Cell-level pathway scoring comparison with a biologically constrained variational autoencoder

Pelin Gundogdu[1,2][0000−0002−1791−9457], Miriam Payá-Milans[1,2][0000−0002−7718−9011], Inmaculada Alamo[1,3][0000−0002−6295−0218], Isabel A. Nepomuceno-Chamorro[4][0000−0002−4255−7160], Joaquin Dopazo[1,2,5,6,✉][0000−0003−3318−120X], and Carlos Loucera[1,2,✉][0000−0001−9598−6965]

[1] Computational Medicine Platform, Andalusian Public Foundation Progress and Health-FPS, Sevilla, Spain
[2] Computational Systems Medicine, Institute of Biomedicine of Seville (IBIS), Hospital Virgen del Rocio, Sevilla, Spain
[3] Department of Immunology, Institute of Biomedicine of Seville (IBIS), Hospital Virgen del Rocio, Sevilla, Spain
[4] Department of Computer Languages and Systems, Universidad de Sevilla, Sevilla, Spain
[5] Bioinformatics in Rare Diseases (BiER), Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), FPS, Hospital Virgen del Rocío, Sevilla, Spain
[6] FPS/ELIXIR-es, Hospital Virgen del Rocío, Sevilla, Spain;
{joaquin.dopazo,cloucera}@juntadeandalucia.es

Abstract. Unsupervised techniques are ubiquitous to study and understand the complex patterns that arise when analyzing genomic data at single-cell resolution. Particularly, unsupervised deep learning models provide state-of-the-art solutions for the most common tasks that arise when dealing with scRNA-seq data. However, the biological usefulness of these complex models is burdened by their black-box nature. To address such limitations several lines of research have emerged, from post hoc approximations to ante hoc modeling. In this work, we study the behavior of two biologically-constrained variational autoencoders (ante hoc modeling). On the one hand, we use a one-layer architecture where the constraints come from the signaling pathways, and, on the other hand, we propose a two-layer architecture following the recent trends in mechanistic models of signal transduction. We use the representations learned by the model as proxies of the signaling activity at the single-cell level. We check the performance of the scoring model using a known scRNA-seq public dataset with a clearly established ground truth. Although both models capture the relevant signals, the most pronounced differences are better captured by the one-layer architecture, while the two-layer design is able to learn more fine-grained features that can expose less prominent aspects of the data.

Keywords: variational autoencoder (VAE) · single-cell RNA sequencing (scRNA-seq) · latent space representation · interpretable neural network (xAI).

1 Introduction

The emergence of single-cell RNA sequencing (scRNA-seq) technologies has enabled the study of the complexity and heterogeneity at the transcriptomic level with an unprecedented resolution [24]. However, as the technologies advance so does the computational needs, in order to produce tools, protocols, and models that can cope with the increasing data dimensionality [16,32]: at the sample (more cells), variable (e.g. more genes), and modality axes (different kinds of measurements).

From a data science point of view, there are innumerable challenges that arise when trying to decipher the complex patterns across the myriad of single-cell datasets released on a yearly basis: from cell-type clustering, annotation, visualization, quality control, dataset integration, to name a few [16]. Machine Learning [23], and in particular Deep Learning-based solutions [29], are especially well suited for scRNA-seq data-driven tasks.

However, the latent representations learned by deep-learning-based methods are not useful for interpreting the underlying biology [30], which is a major drawback from a systems-biology point of view. To address these limitations several ante hoc, which make the model more interpretable a priori, and post hoc, which explain the model a posteriori using subrogate interpretable models, have been proposed. In this work, we are interested in domain-constrained models [5], where the wiring of the neural networks (NNs) is conditioned by one or multiple sources of a priori knowledge of the domain (biology in our case).

Among the plethora of explainable modeling solutions, domain-constrained deep learning models have drawn the attention of the research community. For instance, Visible Neural Networks (VNNs), NNs where the layers are coupled with representations of the biological components of human cells, have been used to better understand eukaryotic cells [19] or to model the human cell structure to predict anti-cancer drug responses [15]. In [18] they propose ExiMap, an interpretable network for data integration and gene program discovery. Although a review of the different explainability and interpretability methods and terminology is beyond the scope of this work, we refer the reader to consult [7] for a general review and [33] for a biology-focused one.

In this work, we aim to analyze the biological utility of the representations learned by a variational autoencoder (VAE) constrained by cell signaling entities. More precisely, we are interested in studying how the source of a priori knowledge conditions the biological results offered by the model. To this aim, we train the informed VAE using two different sources of domain knowledge: i) the Reactome pathway database [6] and ii), the Kyoto Encyclopedia of Genes and Genomes database (KEGG) [22]. While on Reactome we use the gene sets as is, to make it easily comparable with the literature [11], when using the KEGG database we follow the recent trends in mechanistic modeling to decompose the pathways into smaller functional units (the so-called circuits) [12,4]. The models are fit to a dataset of human peripheral blood mononuclear cells (PBMCs) of lupus patients where the ground truth of an interferon-β perturbation experiment is known [13].

We empirically demonstrate how the biological signal learned by a simple known VAE architecture can be refined by using more fine-grained constraints.

2 Materials and Methods

2.1 Dataset

In this work, we evaluate the proposed model's biological usefulness by scoring the pathway and circuit activities in a publicly available dataset that contains untreated and IFN-β stimulated human peripheral blood mononuclear cells (PBMC) from eight patients with Lupus [13], see Table 1 for a brief summary of the cell type distribution. The data was obtained using the `scanpy` library [31], as described in the book "Single-cell best practices" [11].

The dataset allows us to easily measure the performance of unsupervised methods since we have the ground truth for the perturbations, thus we can use an unsupervised method for inferring the signaling activity and compare the results across the two groups: *control* and *stimulated*. Due to the nature of the perturbation, we expect that interferon-based pathways are scored higher in the *stimulated* population than in the *control* (*non-stimulated*) cells. This is a known result [11,18] when studying the dataset with the Reactome pathway database [6].

Table 1. Cell type distribution of the untreated and IFN-stimulated human PBMC cells dataset by Kang et al. [13]

Condition	B cells	CD14+ Monocytes	CD4 T cells	CD8 T cells	Dendritic cells	FCGR3A+ Monocytes	Megakaryocytes	NK
control	1316	2932	5560	811	258	520	63	855
stimulated	1335	2765	5678	810	271	569	69	861

2.2 Sources of biological information

In order to elucidate if the VAE activity scoring is stable under changes of the *a priori* knowledge used to inform the network, we trained two Neural Networks using different sources of biological knowledge. On the one hand, we have trained a VAE informed by the Reactome pathway knowledgebase [6] and, on the other hand, we have used the Kyoto Encyclopedia of Genes and Genomes database (KEGG) [22] to inform the neural network architecture. See Section 2.3 for a concise explanation of how the information is used to inform the variational autoencoder design.

2.3 Model Design

In this work, we proposed a biologically constrained artificial neural network, i.e. a NN whose weights and kernels (the informed layers) are conditioned by a set of *a priori* knowledge. We aim to improve the explainability of the model by learning biologically interpretable scoring functions from each informed layer.

Variational autoencoders To learn a latent representation of the data we adopt a variational autoencoder (VAE) architecture (see Figure 1), which is a type of deep generative model used for unsupervised learning that learns a probability distribution over the latent representation. It aims to reconstruct data points ($\mathbf{x}$), let's call ($\hat{\mathbf{x}}$) the reconstructed approximation, with a minimum error as possible by means of a composite network conformed by an encoder (e) and a decoder (d) block joined in a non-trivial way (i.e. there is a bottleneck layer where the information is necessarily compressed).

The model assumes that there is an underlying data distribution, where the encoder gives us the distribution (Gaussians μ_x, σ_x)) of this latent representation of the data (Equation 1a), while the decoder block samples from the distribution generating new data points (Equation 1b), the approximation ($\hat{\mathbf{x}}$). These latent variables are used to sample a vector ($\mathbf{z}$) (see Equation 2) which is used to feed to the decoder block (see Equation 1b) to reconstruct the input data.

To optimize the model, two terms are needed to compose the final loss function (Equation 3a): the *reconstruction loss* (Equation 3b) and the Kullback–Leibler divergence or *similarity loss*. These terms are derived from the probabilistic model, in our case, given the Gaussian assumptions, the reconstruction loss takes the form of the mean squared error between the input ($\mathbf{x}$) and the approximation ($\hat{\mathbf{x}}$) data points, while the KL loss guides the model to *become* a unit normal distribution (Equation 3c). To compute the expectation value of the *reconstruction* term we use the sampling approximation.

$$encoded\ data(e) = \mathcal{N}(\mu_x, \sigma_x) \tag{1a}$$

$$decoded\ data = d(z) \tag{1b}$$

$$sampling = z \sim \mathcal{N}(\mu_x, \sigma_x) \tag{2}$$

$$loss = reconstruction\ loss + similarity\ loss \tag{3a}$$

$$reconstruction\ loss = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \hat{\mathbf{x}}_i)^2 \tag{3b}$$

$$similarity\ loss = KL[\,\mathcal{N}(\mu_x, \sigma_x), \mathcal{N}(0, 1)\,] \tag{3c}$$

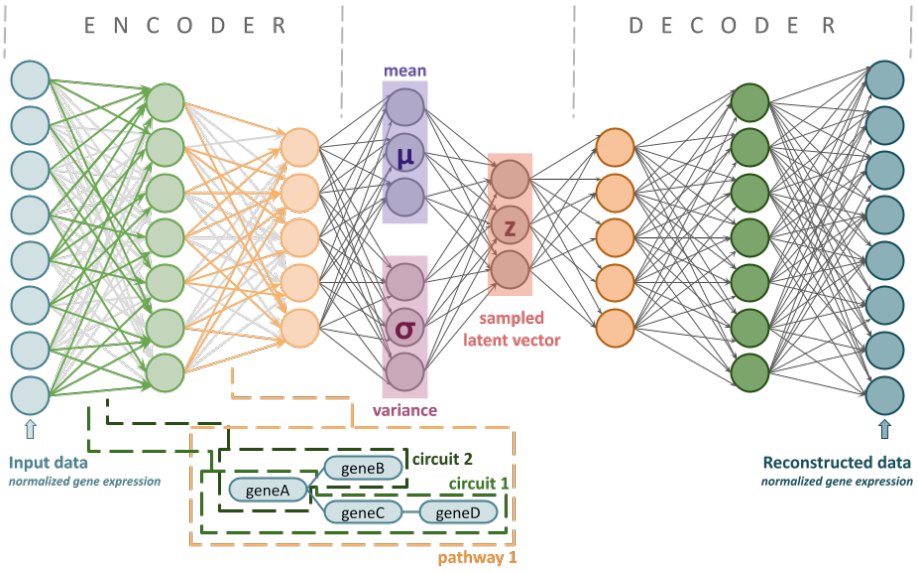

Fig. 1. Proposed Variational autoencoder (VAE) architecture

Signaling-primed layers In this work, we constrain the VAE architecture by using *a priori* biological knowledge that informs the construction and operations of the layers. An informed layer behaves as a linear layer except that the kernel is informed by an indicator matrix ($\mathbf{I}_S$) that informs, using a collection of biological entities $S = \{s\}_{j=1}^{n_S}$, which inputs of the previous layer should be used. Thus, if the previous layer has the biological entities $\mathcal{A} = \{a_1, \ldots, a_k\}$, $\mathbf{I}_S(i,j) = 1$ if the entity a_i belongs to the entity s_j, otherwise $\mathbf{I}_S(i,j) = 0$.

An S-informed layer $\mathbf{H}_S^\star$ is updated using the following formula:

$$\mathbf{H}_S^\star = \text{activation}\left((\mathbf{W} \odot \mathbf{I}_S)\,\mathbf{H} + \mathbf{b}\right) \tag{4}$$

where $\mathbf{W}$ and $\mathbf{b}$ denote the corresponding weight and bias tensors, $\mathbf{I}_S$ the indicator matrix of the signalization gene set S, $\mathbf{H}$ the previous layer outputs, activation in an activation function and, $\odot$ represents the element wise (Hadamard) product.

As has been mentioned before, in this work we define two different architecture designs. On the one hand, we use the Reactome pathway database to inform the first hidden layer of the VAE, thus we only have one informed layer, where $\mathcal{A}$ are the genes, S is the collection of the pathway gene sets and $\mathbf{I}_S$ encodes the information of which gene belongs to every gene set. On the other hand, the architecture for the KEGG database is constructed using the circuit decomposition proposed in [12], where each pathway is decomposed into effector sub-pathways (so-called circuits) that represent the minimal functional units. We inform the first hidden layer of the architecture using the gene sets defined by the

circuits in a similar fashion as with Reactome pathways, where $\mathcal{A}, \mathcal{S}$ represent the genes circuits, respectively. However, given that each pathway is decomposed into as many circuits as effector genes, the resulting neuronal wiring could be too sparse, since each circuit has potentially new genes compared to whole pathways. To overcome such limitations, we create a second hidden layer based on the pathways, a node of the first hidden layer is connected to a node of the second layer if the circuit that it represents belongs to the pathway represented by the node of the second hidden layer. To avoid connecting the layers in blocks, we also connect a pathway ℓ with a circuit j if they share a gene. This KEGG-based design improves upon other works by the authors when conditioning supervised neural networks to identify cell types [9,8].

The circuit decomposition and gene set extraction for the KEGG database has been done using the HiPathia R package (v 2.11.4) [12] for the *Homo sapiens* organism.

Hyperparameter selection We trained the networks for 100 epochs, used Hyperbolic tangent activation functions for all layers except the output layer for the decoder which uses a linear activation function, and the informed layers were regularized using ℓ_2-based activity regularizers, and we used the ADAM optimizer [14]) for the optimization with a learning rate of 1e-5. It was implemented in Python 3.10 using numpy (v 1.23.5) [10], scipy (v 1.10.1) [28] and TensorFlow 2.10 [20]

2.4 Visualization and comparison of the signaling activity

Since the signaling entities, either from circuits or pathways, are first-class entities of our model, we can use standard statistical tools to analyze the inferred signaling activity at the single-cell resolution.

The `Scanpy` library [31] has been used to visualize the signaling activities inferred by the VAE at single-cell resolution following the standard procedure to visualize gene expression: i) we create an `AnnData` object [27,26] where the bottleneck layer substitutes the slot of gene expression data, ii) compute the neighbor graph, iii) cluster the cells using the `leiden` algorithm [25] as proposed in [17], iv) use the UMAP dimensionality reduction technique [21] to produce a two-dimensional space that is easy to visualize, v) add new layers with the inferred activities, and vi) color the cells by the inferred activity of any given circuit or pathway.

In this work, we have used the tools provided by the `Scanpy` library to perform Wilcoxon rank-sum tests to compare the inferred signaling activity across different groups of cells: IFN-β *stimulated* versus *control* (i.e. non-stimulated) cells. In all cases, p values have been corrected for multiple testing with False Discovery Rate (FDR) [3]. We report circuit/pathway names, the test scores (scores), FDR-adjusted p values (pvals_adj), and the log fold change (logFC) between the conditions.

2.5 Code availability

The code required to train the networks and execute all of the analyses reported in this work can be found at https://github.com/babelomics/ivae_scorer.

3 Results

In this section, we present the results of our signaling scoring model trained with KEGG or Reactome as the source of *a priori* knowledge (see Section 2.3). All the models have been fitted using the same hyperparameter schema (see Section 2.3) to the dataset of PBMCs cells described in Section 2.1. The ground truth is labeled as *stimulated* or *control* to indicate if the cells have been treated with interferon-β (IFN-β) or not, respectively.

3.1 Pathway activity at single-cell resolution using KEGG as prior knowledge

Here we present the results on the IFN-β perturbed dataset [13] (see Section 2.1) for the proposed VAE scoring model using the KEGG pathway database as the source of a priori information 2. Once the model has been fitted, we compute the circuit activities using the first hidden layer of the encoder (see Section 2.3) and use them to carry out a Wilcoxon rank-sums test between the *control* and *stimulated* cell groups. The results for the top-10 ranked circuits are summarized in Table 2. The nomenclature for the circuits is "pathway: effector genes".

As expected, the top-ranked circuits detected by our method capture parts of interferon-related signaling pathways. The list is dominated by the "Rig-I-like receptor" pathway since most of its circuits are represented, which could be driven by the fact that RIG-I-like receptors (RLRs) recruit particular intracellular adaptor proteins to activate signaling pathways that result in the creation of type I interferon, among other inflammatory cytokines. Moreover, the circuits that have IFN-α or IFN-β as effectors, namely the "Toll-like receptor signaling pathway: INF-α" and "Toll-like receptor signaling pathway: INF-β" circuits, are among the top-ranked circuits, which should be expected given the interferon-based treatment.

One of the advantages of cell-level gene-set scoring mechanisms is that we can identify cells where specific gene sets are active [1], in our case signaling circuits/pathways. Although a complete analysis of the scoring performance across the different cell groups is beyond the scope of the present work, Figure 2 shows the activity at the single-cell resolution of the major sources of variation across the *stimulated* and *control* groups. As outlined in Section 2.4 each axis represents a component of the UMAP reduction of the latent space learned by the VAE. The top sub-figures represent the cells grouped by either the condition (*control* or *stimulated*) or the cell type. Instead, the rest of the sub-figures show the change in the level of activity of each signaling circuit across the cells.

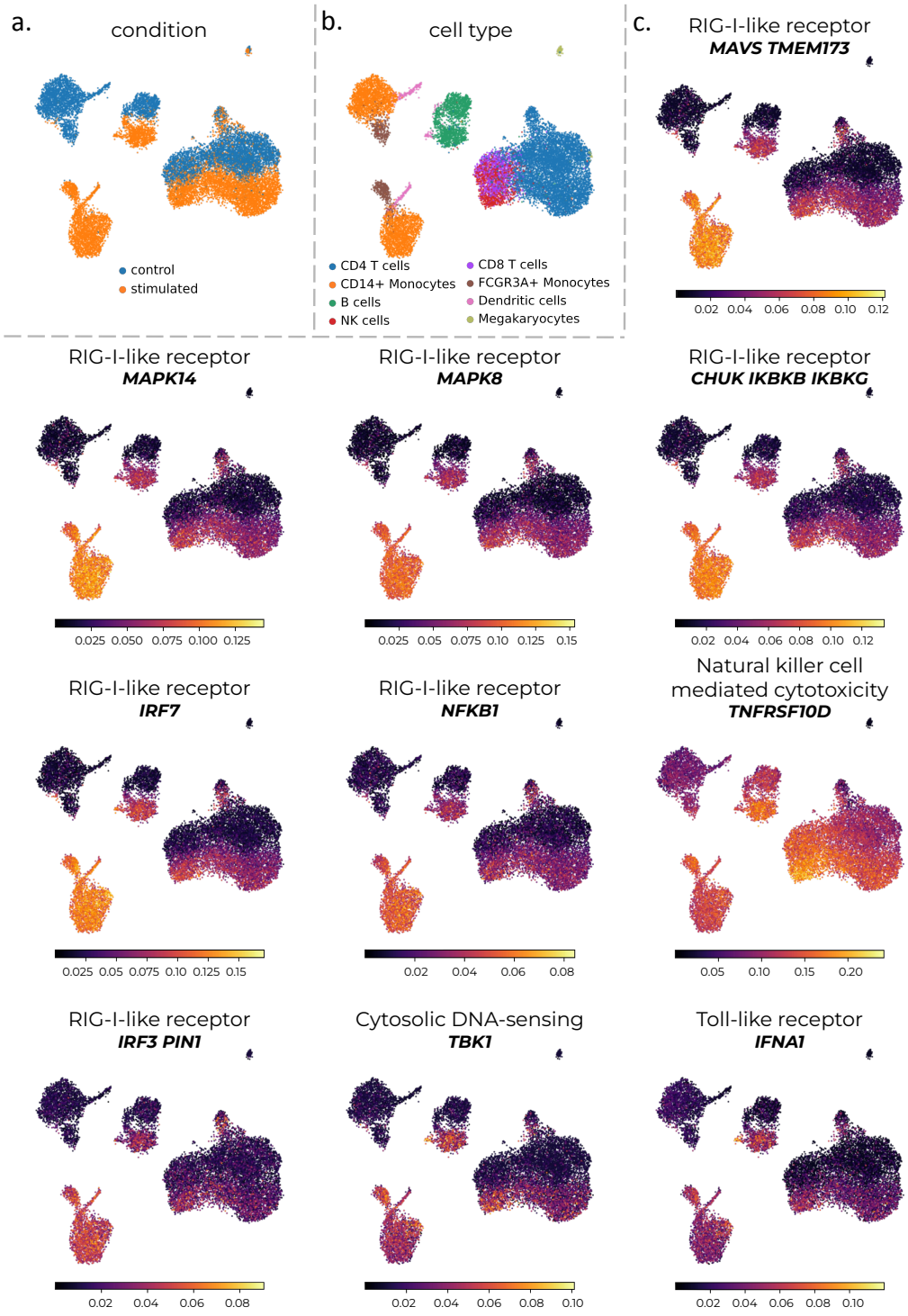

Fig. 2. Circuit activity scoring using the KEGG-informed VAE.

Table 2. Top 10 differentially activated KEGG circuits with respect to the *stimulated* versus *control* comparison

circuits	scores	pvals_adj	logFC
RIG-I-like receptor: MAVS TMEM173	104.57	0.00	2.63
RIG-I-like receptor: MAPK14	103.98	0.00	2.63
RIG-I-like receptor: MAPK8	103.82	0.00	2.62
RIG-I-like receptor: CHUK IKBKB IKBKG	100.95	0.00	2.34
RIG-I-like receptor: IRF7	100.26	0.00	2.46
RIG-I-like receptor: NFKB1	97.65	0.00	2.06
Natural killer cell mediated cytotoxicity: TNFRSF10D	62.89	0.00	0.55
RIG-I-like receptor: IRF3 PIN1	62.15	0.00	1.36
Cytosolic DNA-sensing pathway: TBK1	61.26	0.00	1.36
Toll-like receptor: IFNA1	58.35	0.00	1.37

3.2 Pathway activity at single-cell resolution using Reactome as prior knowledge

Following an analogous procedure to the one presented in the previous section for the KEGG database, we fitted a Reactome-informed VAE to the same IFN-β perturbed dataset [13].

Table 3 shows the top 10 deferentially activated Reactome pathways with respect to the *stimulated* versus *control* comparison following the Wilcoxon procedure (see Section 2.4). Given that a group of cells has been stimulated with IFN-β, it is expected that interferon-related pathways should be ranked higher in the *stimulated* cells with respect to the *control* population [13,11], as it is the case: the top expressed pathways in *stimulated* cells according to the Reactome-informed VAE include the "Interferon α, β Signaling", "Interferon Signaling", "Interferon Gamma Signaling", and "Antiviral Mechanism By IFN Stimulated Genes" pathways, among others. activated Reactome pathways with respect to the *stimulated* versus *control* comparison following the Wilcoxon procedure (see Section 2.4). Given that a group of cells has been stimulated with IFN-β, it is expected that interferon-related pathways should be ranked higher in the *stimulated* cells with respect to the *control* population [13,11], as it is the case: the top expressed pathways in *stimulated* cells according to the Reactome-informed VAE include the "Interferon α, β Signaling", "Interferon Signaling", "Interferon Gamma Signaling", and "Antiviral Mechanism By IFN Stimulated Genes" pathways, among others.

Figure 3 shows the pathway activity scores at the single-cell level for the top-scored interferon-related pathways. The figure is composed as in the KEGG-informed case, the axes represent the UMAP reduction of the cell representation learned by the Reactome-informed VAE, we show the cell type and condition groups, as well as the activity of each top-ranked Reactome pathway across the cells.

The results of the Reactome-informed VAE are on par with those presented in the "Single-cell best practices" manual [11] where they use the same IFN-β

Table 3. Top 10 differentially activated Reactome pathways with respect to the *stimulated* versus *control* comparison

pathways	scores	pvals_adj	logFC
Interferon α, β Signaling	109.32	0.00	3.75
Interferon Signaling	105.27	0.00	2.42
Cytokine Signaling In Immune System	104.96	0.00	2.43
Negative Regulators Of DDX58 IFIH1 Signaling	102.65	0.00	2.39
Antiviral Mechanism By IFN Stimulated Genes	101.54	0.00	2.38
NS1 Mediated Effects On Host Pathways	95.99	0.00	2.15
Interferon Gamma Signaling	88.25	0.00	0.97
OAS Antiviral Response	87.39	0.00	2.44
Post Translational Modification...	87.00	0.00	2.19
DNA Damage Bypass	84.06	0.00	1.83

perturbed dataset to assert the performance of the `decoupleR` tool [2] when inferring the activity of the pathways at the single-cell resolution by means of the `AUCell` method [1].

Furthermore, the difference at the signaling level between the *stimulated* and *control* condition is clearer and easier to interpret when using Reactome to inform the VAE since it includes a specific entity for the interferon signaling pathway.

3.3 Gene-set precision

In this section, we use the previously fitted models (KEGG and Reactome informed VAE) to study the Influenza pathway, whose enrichment has been already described for the dataset under study in [13].

When using the KEGG-informed VAE we observe a significant alteration of the "Influenza A: IRF7" circuit (part of the "Influenza A" KEGG pathway) between the *control* and *stimulated* condition (FDR-adjusted p-value < 0.05, logFC > 1). Interestingly, the "Influenza A: IRF7" circuit leads to the interferon-based T cell activation and antibody response. Figure 4 shows the "Influenza A: IRF7" circuit scores across the cells using the same representation schema of the previous section.

However, when using the broader Reactome pathways to inform the VAE, the model losses the ability to capture the Influenza-based differences that require more precise gene sets. Instead, the model captures differences at the cell-type level, as can be seen in Figure 5.

Note that, these differences in performance could not be related to differences in the intrinsic quality of the databases. Further research is needed since it is likely that the Influenza signal could be captured with the Reactome database by finding a way to decompose each pathway into functional subunits akin to what we have done with the KEGG database.

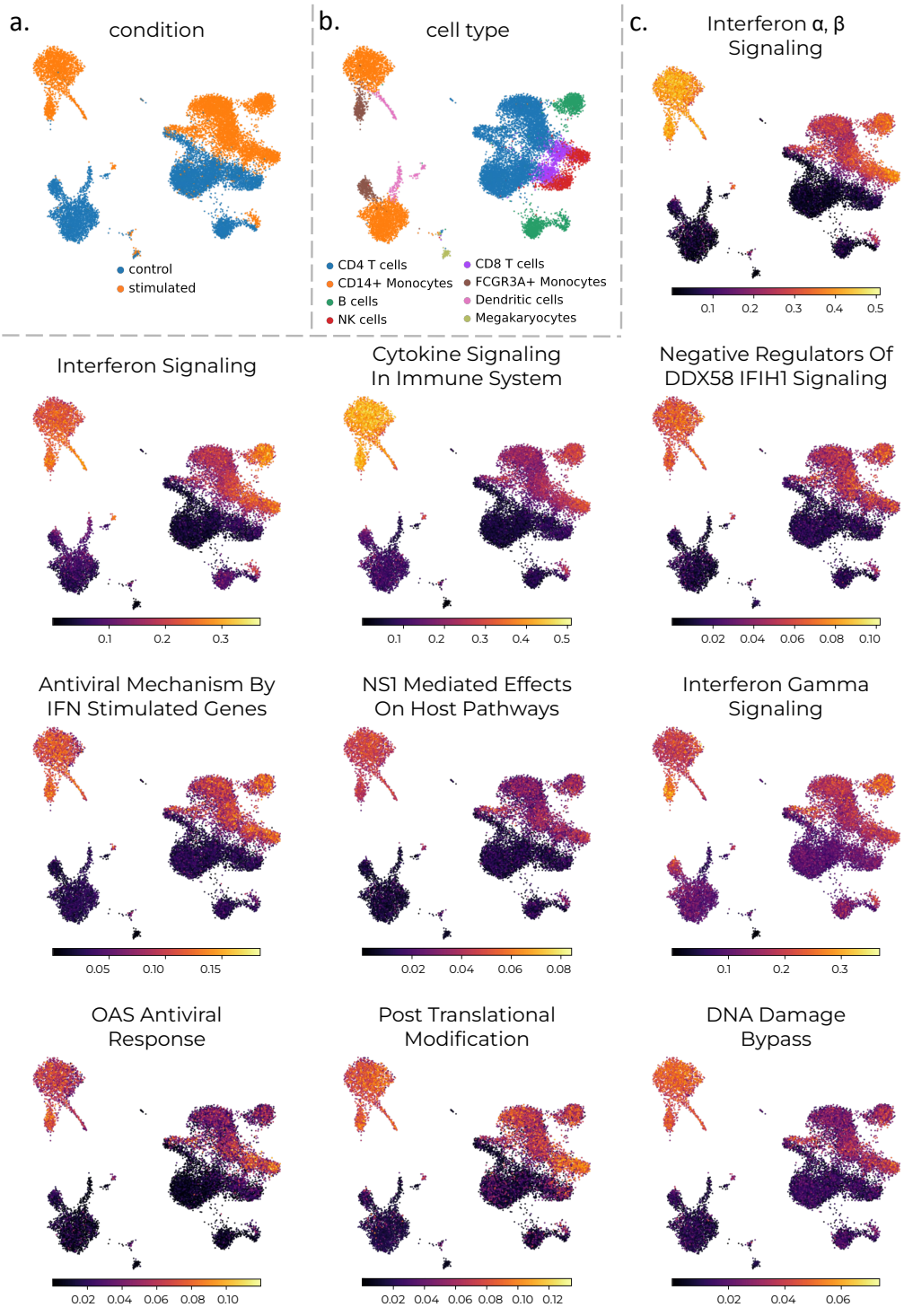

Fig. 3. Pathway activity scoring using the Reactome-informed VAE.

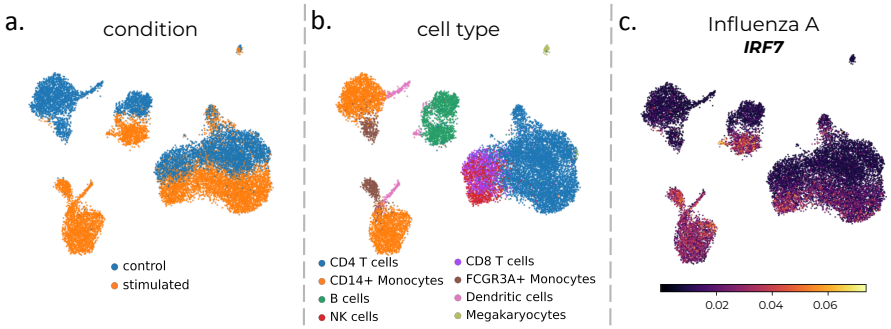

Fig. 4. Influenza circuit scoring using the KEGG-informed VAE.

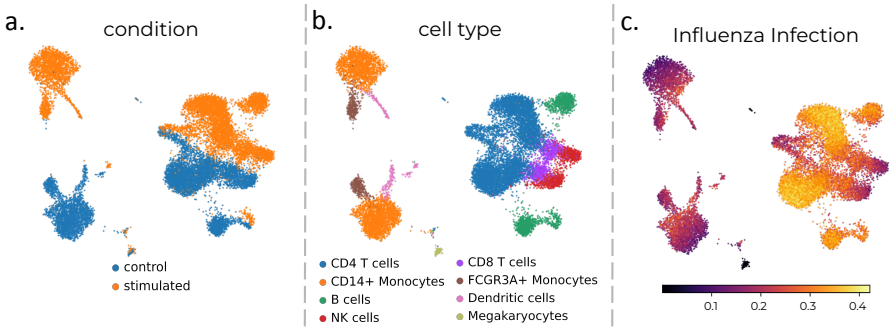

Fig. 5. Influenza pathway activity scoring using the Reactome-informed VAE.

4 Conclusions

This work provides an empirical evaluation of how unsupervised neural networks can be conditioned with a priori biological knowledge to infer cell-level pathway activities. Our proposed signaling-informed variational autoencoder provides results comparable to those found in the literature while being easily extensible to include other biological entities. However, as expected, our model works better if the knowledge base used to inform the wiring of the neural network includes entities that are better aligned with the underlying biological conditions.

Acknowledgements. This work has been partially supported by grants PID2020-117979RB-I00 and PID2020-117954RB-C22 from the Spanish Ministry of Science and Innovation, IMP/0019 from the Instituto de Salud Carlos III (ISCIII), PIP-0087-2021 from Junta de Andalucía, co-funded with European Regional Development Funds (ERDF); grant H2020 Programme of the European Union grants Marie Curie Innovative Training Network "Machine Learning Frontiers in Precision Medicine" (MLFPM) (GA 813533). The authors also acknowledge Junta de Andalucía for the postdoctoral contract of Carlos Loucera (PAIDI2020-DOC_00350) co-funded by the European Social Fund (FSE) 2014–2020.

References

1. Aibar, S., González-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.C., Geurts, P., Aerts, J., van den Oord, J., Atak, Z.K., Wouters, J., Aerts, S.: SCENIC: Single-cell regulatory network inference and clustering. Nature Methods **14**(11), 1083–1086 (Nov 2017). https://doi.org/10.1038/nmeth.4463
2. Badia-i-Mompel, P., Vélez Santiago, J., Braunger, J., Geiss, C., Dimitrov, D., Müller-Dott, S., Taus, P., Dugourd, A., Holland, C.H., Ramirez Flores, R.O., Saez-Rodriguez, J.: decoupleR: Ensemble of computational methods to infer biological activities from omics data. Bioinformatics Advances **2**(1), vbac016 (Jan 2022). https://doi.org/10.1093/bioadv/vbac016
3. Benjamini, Y., Hochberg, Y.: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) **57**(1), 289–300 (1995)
4. Çubuk, C., Loucera, C., Peña-Chilet, M., Dopazo, J.: Crosstalk between Metabolite Production and Signaling Activity in Breast Cancer. International Journal of Molecular Sciences **24**(8), 7450 (Jan 2023). https://doi.org/10.3390/ijms24087450
5. Dash, T., Chitlangia, S., Ahuja, A., Srinivasan, A.: A review of some techniques for inclusion of domain-knowledge into deep neural networks. Scientific Reports **12**(1), 1040 (Jan 2022). https://doi.org/10.1038/s41598-021-04590-0
6. Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senff-Ribeiro, A., Griss, J., Sevilla, C., Matthews, L., Gong, C., Deng, C., Varusai, T., Ragueneau, E., Haider, Y., May, B., Shamovsky, V., Weiser, J., Brunson, T., Sanati, N., Beckman, L., Shao, X., Fabregat, A., Sidiropoulos, K., Murillo, J., Viteri, G., Cook, J., Shorser, S., Bader, G., Demir, E., Sander, C., Haw, R., Wu, G., Stein, L., Hermjakob, H., D'Eustachio, P.: The reactome pathway knowledgebase 2022. Nucleic Acids Research **50**(D1), D687–D692 (Jan 2022). https://doi.org/10.1093/nar/gkab1028
7. Graziani, M., Dutkiewicz, L., Calvaresi, D., Amorim, J.P., Yordanova, K., Vered, M., Nair, R., Abreu, P.H., Blanke, T., Pulignano, V.: A global taxonomy of interpretable AI: Unifying the terminology for the technical and social sciences. Artificial intelligence review pp. 1–32 (2022)
8. Gundogdu, P., Alamo, I., Nepomuceno-Chamorro, I.A., Dopazo, J., Loucera, C.: SigPrimedNet: A Signaling-Informed Neural Network for scRNA-seq Annotation of Known and Unknown Cell Types. Biology **12**(4), 579 (Apr 2023). https://doi.org/10.3390/biology12040579
9. Gundogdu, P., Loucera, C., Alamo-Alvarez, I., Dopazo, J., Nepomuceno, I.: Integrating pathway knowledge with deep neural networks to reduce the dimensionality in single-cell RNA-seq data. BioData Mining **15**(1), 1 (Jan 2022). https://doi.org/10.1186/s13040-021-00285-4
10. Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E.: Array programming with NumPy. Nature **585**(7825), 357–362 (Sep 2020). https://doi.org/10.1038/s41586-020-2649-2
11. Heumos, L., Schaar, A.C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., Lücken, M.D., Strobl, D.C., Henao, J., Curion, F., Single-cell Best Practices Consortium, Aliee, H., Ansari, M., Badia-i-Mompel, P., Büttner, M., Dann, E., Dimitrov, D., Dony, L., Frishberg, A., He, D., Hediyeh-zadeh, S., Hetzel, L., Ibarra, I.L.,

Jones, M.G., Lotfollahi, M., Martens, L.D., Müller, C.L., Nitzan, M., Ostner, J., Palla, G., Patro, R., Piran, Z., Ramírez-Suástegui, C., Saez-Rodriguez, J., Sarkar, H., Schubert, B., Sikkema, L., Srivastava, A., Tanevski, J., Virshup, I., Weiler, P., Schiller, H.B., Theis, F.J.: Best practices for single-cell analysis across modalities. Nature Reviews Genetics (Mar 2023). https://doi.org/10.1038/s41576-023-00586-w

12. Hidalgo, M.R., Cubuk, C., Amadoz, A., Salavert, F., Carbonell-Caballero, J., Dopazo, J.: High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes. Oncotarget **8**(3), 5160–5178 (Dec 2016). https://doi.org/10.18632/oncotarget.14107
13. Kang, H.M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., Wan, E., Wong, S., Byrnes, L., Lanata, C.M., Gate, R.E., Mostafavi, S., Marson, A., Zaitlen, N., Criswell, L.A., Ye, C.J.: Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nature Biotechnology **36**(1), 89–94 (Jan 2018). https://doi.org/10.1038/nbt.4042
14. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization (Jan 2017). https://doi.org/10.48550/arXiv.1412.6980
15. Kuenzi, B.M., Park, J., Fong, S.H., Sanchez, K.S., Lee, J., Kreisberg, J.F., Ma, J., Ideker, T.: Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells. Cancer Cell **38**(5), 672–684.e6 (Nov 2020). https://doi.org/10.1016/j.ccell.2020.09.014
16. Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D.J., Hicks, S.C., Robinson, M.D., Vallejos, C.A., Campbell, K.R., Beerenwinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamatakis, A., Attolini, C.S.O., Aparicio, S., Baaijens, J., Balvert, M., de Barbanson, B., Cappuccio, A., Corleone, G., Dutilh, B.E., Florescu, M., Guryev, V., Holmer, R., Jahn, K., Lobo, T.J., Keizer, E.M., Khatri, I., Kielbasa, S.M., Korbel, J.O., Kozlov, A.M., Kuo, T.H., Lelieveldt, B.P., Mandoiu, I.I., Marioni, J.C., Marschall, T., Mölder, F., Niknejad, A., Raczkowski, L., Reinders, M., de Ridder, J., Saliba, A.E., Somarakis, A., Stegle, O., Theis, F.J., Yang, H., Zelikovsky, A., McHardy, A.C., Raphael, B.J., Shah, S.P., Schönhuth, A.: Eleven grand challenges in single-cell data science. Genome Biology **21**(1), 31 (Feb 2020). https://doi.org/10.1186/s13059-020-1926-6
17. Levine, J.H., Simonds, E.F., Bendall, S.C., Davis, K.L., Amir, E.a.D., Tadmor, M.D., Litvin, O., Fienberg, H.G., Jager, A., Zunder, E.R., Finck, R., Gedman, A.L., Radtke, I., Downing, J.R., Pe'er, D., Nolan, G.P.: Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell **162**(1), 184–197 (Jul 2015). https://doi.org/10.1016/j.cell.2015.05.047
18. Lotfollahi, M., Rybakov, S., Hrovatin, K., Hediyeh-zadeh, S., Talavera-López, C., Misharin, A.V., Theis, F.J.: Biologically informed deep learning to query gene programs in single-cell atlases. Nature Cell Biology **25**(2), 337–350 (Feb 2023). https://doi.org/10.1038/s41556-022-01072-x
19. Ma, J., Yu, M.K., Fong, S., Ono, K., Sage, E., Demchak, B., Sharan, R., Ideker, T.: Using deep learning to model the hierarchical structure and function of a cell. Nature Methods **15**(4), 290–298 (Apr 2018). https://doi.org/10.1038/nmeth.4627
20. Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Jia, Y., Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vin-

cent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, Xiaoqiang Zheng: TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems (2015)

21. McInnes, L., Healy, J., Melville, J.: UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction (Sep 2020). https://doi.org/10.48550/arXiv.1802.03426
22. Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., Kanehisa, M.: KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research **27**(1), 29–34 (Jan 1999). https://doi.org/10.1093/nar/27.1.29
23. Petegrosso, R., Li, Z., Kuang, R.: Machine learning and statistical methods for clustering single-cell RNA-sequencing data. Briefings in Bioinformatics **21**(4), 1209–1223 (Jul 2020). https://doi.org/10.1093/bib/bbz063
24. Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Göttgens, B., Hacohen, N., Haniffa, M., Hemberg, M., Kim, S., Klenerman, P., Kriegstein, A., Lein, E., Linnarsson, S., Lundberg, E., Lundeberg, J., Majumder, P., Marioni, J.C., Merad, M., Mhlanga, M., Nawijn, M., Netea, M., Nolan, G., Pe'er, D., Phillipakis, A., Ponting, C.P., Quake, S., Reik, W., Rozenblatt-Rosen, O., Sanes, J., Satija, R., Schumacher, T.N., Shalek, A., Shapiro, E., Sharma, P., Shin, J.W., Stegle, O., Stratton, M., Stubbington, M.J.T., Theis, F.J., Uhlen, M., van Oudenaarden, A., Wagner, A., Watt, F., Weissman, J., Wold, B., Xavier, R., Yosef, N., Human Cell Atlas Meeting Participants: The Human Cell Atlas. eLife **6**, e27041 (Dec 2017). https://doi.org/10.7554/eLife.27041
25. Traag, V., Waltman, L., van Eck, N.J.: From Louvain to Leiden: Guaranteeing well-connected communities. Scientific Reports **9**(1), 5233 (Mar 2019). https://doi.org/10.1038/s41598-019-41695-z
26. Virshup, I., Bredikhin, D., Heumos, L., Palla, G., Sturm, G., Gayoso, A., Kats, I., Koutrouli, M., Berger, B., Pe'er, D., Regev, A., Teichmann, S.A., Finotello, F., Wolf, F.A., Yosef, N., Stegle, O., Theis, F.J.: The scverse project provides a computational ecosystem for single-cell omics data analysis. Nature Biotechnology pp. 1–3 (Apr 2023). https://doi.org/10.1038/s41587-023-01733-8
27. Virshup, I., Rybakov, S., Theis, F.J., Angerer, P., Wolf, F.A.: Anndata: Annotated data (Dec 2021). https://doi.org/10.1101/2021.12.16.473007
28. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P.: SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nature Methods **17**(3), 261–272 (Mar 2020). https://doi.org/10.1038/s41592-019-0686-2
29. Wang, J., Zou, Q., Lin, C.: A comparison of deep learning-based pre-processing and clustering approaches for single-cell RNA sequencing data. Briefings in Bioinformatics **23**(1), bbab345 (Jan 2022). https://doi.org/10.1093/bib/bbab345
30. Way, G.P., Greene, C.S.: Discovering Pathway and Cell Type Signatures in Transcriptomic Compendia with Machine Learning. Annual Review of Biomedical Data Science **2**(1), 1–17 (2019). https://doi.org/10.1146/annurev-biodatasci-072018-021348

31. Wolf, F.A., Angerer, P., Theis, F.J.: SCANPY: Large-scale single-cell gene expression data analysis. Genome Biology **19**(1), 15 (Feb 2018). https://doi.org/10.1186/s13059-017-1382-0
32. Zappia, L., Theis, F.J.: Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape. Genome Biology **22**(1), 301 (Oct 2021). https://doi.org/10.1186/s13059-021-02519-4
33. Zhao, Y., Shao, J., Asmann, Y.W.: Assessment and Optimization of Explainable Machine Learning Models Applied to Transcriptomic Data. Genomics, Proteomics & Bioinformatics **20**(5), 899–911 (Oct 2022). https://doi.org/10.1016/j.gpb.2022.07.003