biology

MDPI

Article

SigPrimedNet: A Signaling-Informed Neural Network for scRNA-seq Annotation of Known and Unknown Cell Types

Pelin Gundogdu [1,2], Inmaculada Alamo [1,2], Isabel A. Nepomuceno-Chamorro [3], Joaquin Dopazo [1,2,4,5,*] and Carlos Loucera [1,2,*]

[1] Computational Medicine Platform, Andalusian Public Foundation Progress and Health-FPS, 41013 Sevilla, Spain
[2] Computational Systems Medicine, Institute of Biomedicine of Seville (IBIS), Hospital Virgen del Rocio, 41013 Sevilla, Spain
[3] Dpto. de Lenguajes y Sistemas Informáticos, Universidad de Sevilla, 41013 Seville, Spain
[4] Bioinformatics in Rare Diseases (BiER), Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), FPS, Hospital Virgen del Rocío, 41013 Sevilla, Spain
[5] FPS/ELIXIR-es, Hospital Virgen del Rocío, 42013 Sevilla, Spain
[*] Correspondence: joaquin.dopazo@juntadeandalucia.es (J.D.); carlos.loucera@juntadeandalucia.es (C.L.)

Simple Summary: Single-cell data has enabled the study of cell dynamics at an unprecedented resolution. Cell type and functional annotation are crucial to address during any analysis involving transcriptomic data at the cell level since both annotations provide the basis to understand the complex biological processes behind the communication machinery. We propose SigPrimedNet, a data-driven solution to identify cells while learning a functional summarization of signaling measurements by incorporating the knowledge stored in pathway databases. To do so, we decompose each signaling pathway into canonical effector circuits, which act as a minimal functional unit. These circuits inform the design of a cell-type classification neural network model, which allows us to extract meaningful features that act as a proxy of the signaling activity of any given cell. Furthermore, we train an unsupervised anomaly detection algorithm on the inferred activities, which enables the model to identify unknown cells when working with previously unseen cells. To illustrate the performance of the proposed model we conduct a series of experiments over publicly available data with promising results across every task: cell-type annotation, unknown cell-type identification, and clustering. Finally, we showcase the biological richness of the signaling activity learned by the model.

Abstract: Single-cell RNA sequencing is increasing our understanding of the behavior of complex tissues or organs, by providing unprecedented details on the complex cell type landscape at the level of individual cells. Cell type definition and functional annotation are key steps to understanding the molecular processes behind the underlying cellular communication machinery. However, the exponential growth of scRNA-seq data has made the task of manually annotating cells unfeasible, due not only to an unparalleled resolution of the technology but to an ever-increasing heterogeneity of the data. Many supervised and unsupervised methods have been proposed to automatically annotate cells. Supervised approaches for cell-type annotation outperform unsupervised methods except when new (unknown) cell types are present. Here, we introduce SigPrimedNet an artificial neural network approach that leverages (i) efficient training by means of a sparsity-inducing signaling circuits-informed layer, (ii) feature representation learning through supervised training, and (iii) unknown cell-type identification by fitting an anomaly detection method on the learned representation. We show that SigPrimedNet can efficiently annotate known cell types while keeping a low false-positive rate for unseen cells across a set of publicly available datasets. In addition, the learned representation acts as a proxy for signaling circuit activity measurements, which provide useful estimations of the cell functionalities.

check for updates

Citation: 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* **2023**, *12*, 579. https://doi.org/10.3390/biology12040579

Academic Editors: W. Brad Barbazuk and Georgios K. Georgakilas

Received: 27 December 2022
Revised: 4 March 2023
Accepted: 8 April 2023
Published: 10 April 2023

Keywords: scRNA-seq; deep learning; explainable artificial intelligence; cell signaling; cell-type identification

1. Introduction

Recent high-throughput technology developments are transforming our view of complex biological systems by providing a detailed picture of their individual components. Single-cell RNA sequencing (scRNA-seq) has enabled RNA activity to be profiled in individual single cells by obtaining profiles of thousands of cells in heterogeneous environments [1]. scRNA-seq increases our understanding of the cell as a functional unit revealing new populations of cells with gene expression profiles previously unnoticed in conventional analyses of bulk cell populations [2].

Facing the huge amount of data provided by scRNA-seq technology, one of the major challenges is cell-type identification within a diverse population of sequenced cells. This challenge, also known as cell retrieval or cell-type annotation, consists of inferring the type of a given cell by querying a reference database of annotated scRNA-seq data. Unsupervised methods, such as clustering analysis, find the closest cell to a sample given a population of cells. However, single-cell data contains high levels of noise from heterogeneous sources, and to mitigate such problems, dimensionality reduction is usually performed before clustering. Scmap projection algorithm [3] explores different strategies for feature selection as highly variable genes (HVGs) [4] and genes with a higher number of dropouts (zero expression) than expected determined using M3Drop [5]. The most popular methods for dimensionality reduction are based on Principal Component Analysis (PCA) [6], dropout modeling (ZIFA) [7], t-distributed stochastic neighbor embedding (TSNE) [8] or uniform manifold approximation and projection (UMAP) [9]. Single reference mapping methods are growing in popularity as Seurat's supervised principal component analysis [10], single-cell architecture surgery (scArches) [11], or an extension of Harmony [12] to map query datasets by minimal modification of the reference atlas [13]. However, the implicitly used latent dimensions for joint data representation are not directly interpretable, and it is a major drawback of these methods [14]. Currently, the development of interpretable models by the addition of statistical assumptions or prior biological information is a trend, but the former approaches have not yielded sufficiently useful latent spaces in the context of scRNA-seq analysis [14].

Supervised methods use a labeled reference to learn a function that maps transcriptomic profiles to cell types. Thereafter, new cells are annotated using the learned mapping. Model training (learning the map) is usually a time-consuming process due to the large size of the reference databases [3], while inference (applying the learned function) is faster and less laborious than the two-step process associated with unsupervised methods [15]. Furthermore, supervised training for cell-type annotation usually performs better than unsupervised methods in most datasets, although this is not the case when unknown cell types arise [16]. One of the more promising methods to overcome such limitations is SciBet, which uses a combination of statistical learning to find informative genes, a multinomial approximation for cell-type annotation, and building a synthetic reference cell to estimate out-of-distribution transcriptomic profiles. Scibet outperforms other state-of-the-art methods like Seurat v3 and scMap across several experiments, achieving a high prediction accuracy while keeping a low false-positive rate when annotating unseen cell types.

In this work, we present SigPrimedNet a domain-informed Artificial Neural Network (ANN) that overcomes the limitations associated with supervised learning methods by combining a signaling circuits-informed sparse architecture with an anomaly detection procedure that uses the latent structure learned by the ANN to elucidate if any given cell is of unknown origin. Sparse domain-informed neural networks are used to solve complex biological problems by incorporating domain-specific constraints on the underlying architectures to develop more interpretable models that avoid overfitting through regularization [17]. For example, P-NET [18] incorporates different biological entities to aid in decision-making when dealing with prostate cancer patients, whereas Dcell models gene interactions on cell growth in yeast [19]. In the context of cell annotation [20] uses algorithmically crafted clusters of protein-protein and protein-DNA interactions to provide the sparse structure, but cannot classify unknown cells.

Our previous work on cell-type identification [21] used broader, all-encompassing, pathways and lacked any form of out-of-distribution learning, which hampered its usefulness when the query dataset representation showed more heterogeneity in cell populations. Contrarily, SigPrimedNet offers a more fine-grained functional characterization of the cell populations due to the use of more specific effector-based signaling proxies based on recent developments in mechanistic models of cell signaling, which ultimately triggers cell functionality and dictates cell behavior and fate [22]. Our method outperforms Scibet, Seurat v3, and ScMap when dealing with unknown cell types while providing a comparable performance on tasks where no cells should be labeled as unknown (using the experiments proposed in [15]). To the best of our knowledge, SigPrimedNet is the first supervised Domain-informed Sparse Neural Network to incorporate unknown cell-type identification.

2. Materials and Methods

2.1. Datasets

In this manuscript, we use three publicly available datasets, which we have called PBMC, Immune, and Melanoma dataset to facilitate their reference throughout the manuscript. All of them are publicly available on two platforms, Gene Expression Omnibus (GEO [23]) and 10× Genomics [24], moreover, they are human sequencing data. The datasets used in this work have been obtained from [15] (PBMC and Melanoma) and [25] (Immune). See Table 1 for cell type details.

Table 1. Cell type distribution for each dataset.

PBMC		**Immune**		**Melanoma**	
Cell Type	**# of Samples**	**Cell Type**	**# of Samples**	**Cell Type**	**# of Samples**
CD14+	2500	B cells	1465	B.cell	818
CD19+	2500	Erythrocytes	1747	Macrophage	420
CD34+	2500	HSPCs	3742	NK	92
CD56+	2500	Monocytes	954	T.CD4+	856
CD8+ Cytotoxic	2500	Neutrophils	485	T.CD8+	1759
CD4+/CD45RO+ Memory	2500	NK	546	Negative cells	2228
Treg	2500	T cells	517		

2.1.1. PBMC Dataset

The full version of the fresh peripheral blood mononuclear cells (PBMCs) datasets is publicly available in 10x Genomics [24]. In this work, we use the preprocessed version proposed in [15], which consists of 2500 cells randomly sampled for each cell type: CD14+, CD19+, CD34+, CD56+, CD8+ Cytotoxic, CD4+/CD45RO+ Memory, and Treg cells. In addition, to test the reliability of the model with unbalanced datasets, we have randomly undersampled each cell type (using a proportion of 0.2, 0.4, and 0.6 of the original population) to produce a total of 21 synthetic datasets derived from the PBMC dataset.

2.1.2. Immune Dataset

This dataset profiles the transcriptomes of bone marrow and peripheral blood-derived hematopoietic cells, which are publicly available from GEO database [23] with identifiers GSE137864 and GSE149938. The dataset profiles 7 cell types for 9456 samples (see Table 1) using a unique molecular identifier (UMI) counting [26]. To be more precise, CD34+ HSPCs, B cells, NK cells, T cells, monocytes, neutrophils, and erythrocytes for bone marrow, and together with regulatory B, naive B, memory B, cytotoxic NK, cytokine NK and T cells for peripheral blood-derived differentiated cells.

2.1.3. Melanoma Dataset

This human melanoma scRNA-seq dataset has malignant cells, CD8+ and CD4+ T cells, B cells, natural killer (NK) cells, macrophages, cancer-associated fibroblasts (CAFs), and endothelial cell types. The cell types of CAF, malignant, and endothelial cells are combined in one group called *negative cell*. In [15] they propose a filtered version of the dataset, which profiles 6 cell types for 6173 samples (see Table 1). The dataset is split into two subsets called reference and query with 70–30% sampling size, where the negative cells only appear in the query set. Note that, contrary to Scibet, SigPrimedNet does not rely on an external synthetic reference cell constructed from the aggregation of several single-cell datasets, so we do not make use of the massive reference set described in [15].

2.2. Analysis Workflow

In this work, we propose an analysis workflow that tries to show how our proposed model (SigPrimedNet) can correctly identify previously unseen cell types without losing the advantages of supervised learning (fast and accurate known cell-type assignment) while providing a biologically useful latent space. The workflow (Figure 1) can be summarized in three steps: (A) data processing and architecture design, (B) knowledge extraction from learned representations (*interpretation*), and (C) cell-type inference.

In broad terms, the model works as follows: (i) the weights of the first hidden layer of a dense network are constrained by a binary matrix that encodes the biological information extracted from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [27], (ii) any given training dataset is decomposed into two sets (learning and validation) stratified by cell-type, (iii) the model is fitted to the learning set while using the validation for early stopping the training, (iv) computes the learned representation (encoding) of the learning and validation sets by evaluating the activations of the last hidden layer, (v) fits an anomaly detection algorithm using the encoding of the learning set as the features, and, (vi) establishes a threshold for detecting anomalies (unknown cell types) using the validation encodings. When a new cell is evaluated, the model computes the corresponding encoding, decides if the cell is of an unknown cell type by applying the anomaly detection algorithm along with the learned threshold, and, finally if the cell is not an anomaly the cell-type mapping learned by the ANN is applied.

To check the performance of the model (when all cell types are known) we have followed [15] using the resampled PBMC dataset to conduct a 50 times repeated cell-type stratified cross-validation. Whereas, to test the capability to identify unknown cell types, we have followed the negative cell melanoma experiment as proposed by [15]. Finally, the functional interpretability of our model has been tested using the Immune dataset, where we have also checked the performance by means of 30 times repeated 10-fold cross-validation strategy (all cell types are known).

2.3. Model Design

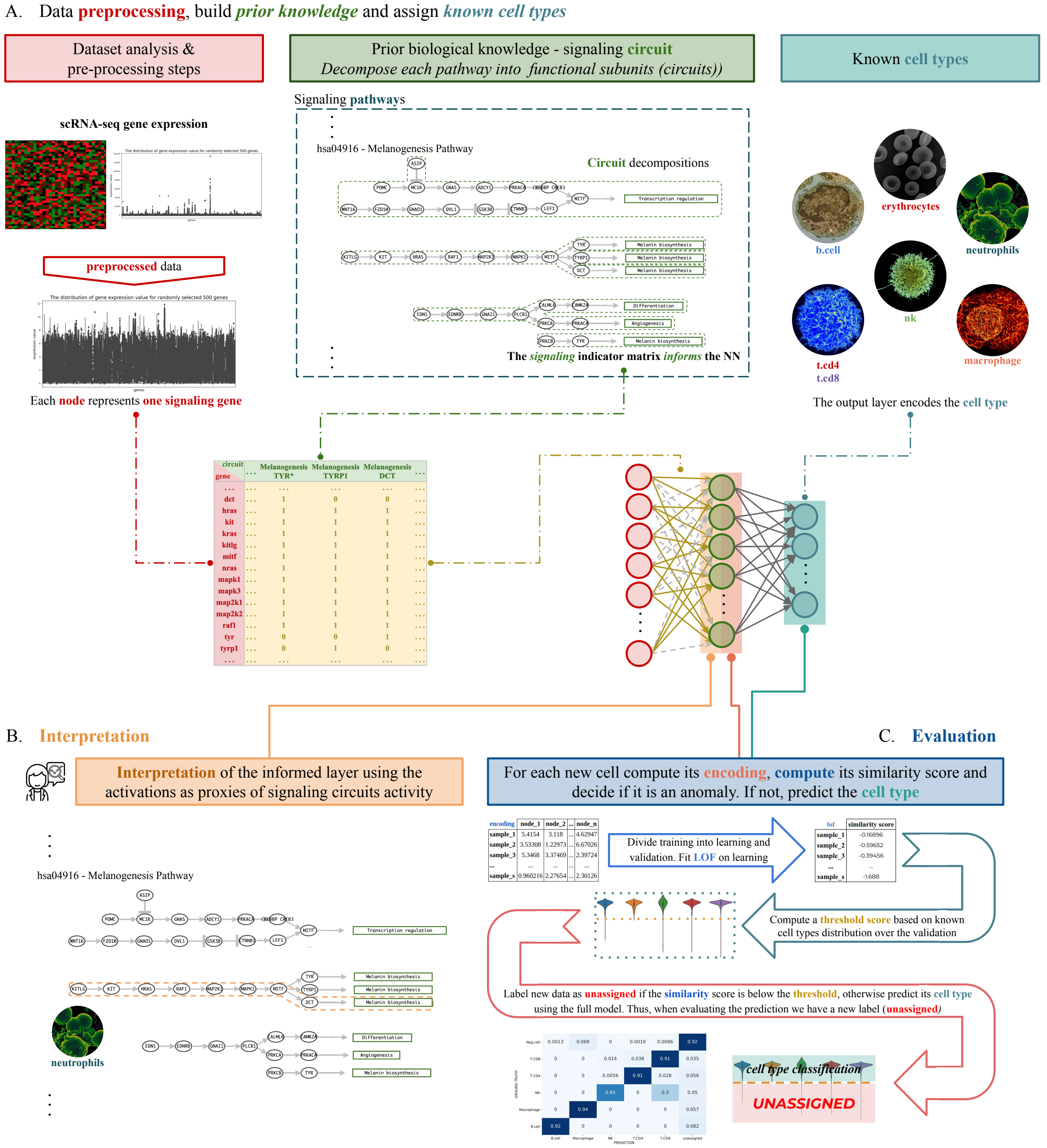
The architecture of the SigPrimedNet is defined as a dense network (all nodes in any given layer are connected to all the nodes of the adjacent layers), where the input layer (one node for each gene) is connected to a signaling-informed layer (the first hidden layer), which is wired to a new dense layer (the encoding layer). Finally, a softmax layer (2) connects the network to the output (the cell types). The model uses Rectified Linear Units (ReLU) [28] activation functions (1) except for the output layer. To train the network, we use the categorical cross-entropy loss function (3), where each known cell type represents a category.

$$\mathrm{relu}(\mathbf{z}) = \max(0, \mathbf{z}). \tag{1}$$

$$\mathrm{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{n_c} e^{z_j}}. \tag{2}$$

$$-\sum_{j=1}^{n_c} y_{i,j} \log(p_{i,j}) \tag{3}$$

where $\mathbf{z}$ refers to real-valued data, n_c to the number of cell types, $y_{i,j}$ is 1 if cell type j is the correct classification for observation i, 0 otherwise. Finally, $p_{i,j}$ is the probability that the observation i belongs to cell type j.


A. Data preprocessing, build prior knowledge and assign known cell types
Dataset analysis & pre-processing steps
Prior biological knowledge - signaling circuit
Decompose each pathway into functional subunits (circuits))
Known cell types
scRNA-seq gene expression
The distribution of gene expression value for randomly selected 500 genes
preprocessed data
The distribution of gene expression value for randomly selected 500 genes
expression value
genes
Each node represents one signaling gene
Signaling pathways
hsa04916 - Melanogenesis Pathway
Circuit decompositions
ASIP
POMC
MC1R
GNAS
ADCY1
PRKACA
CREBBP
CREB3
MITF
WNT16
FZD10
GNAO1
DVL1
GSK3B
CTNNB1
LEF1
Transcription regulation
KITLG
KIT
HRAS
RAF1
MAP2K1
MAPK1
TYR
TYRP1
DCT
Melanin biosynthesis
EDN1
EDNRB
GNAI1
PLCB1
CALML6
CAMK2A
PRKCA
PRKACA
PRKCB
Differentiation
Angiogenesis
The signaling indicator matrix informs the NN
erythrocytes
b.cell
neutrophils
nk
t.cd4
t.cd8
macrophage
The output layer encodes the cell type
B. Interpretation
Interpretation of the informed layer using the activations as proxies of signaling circuits activity
hsa04916 - Melanogenesis Pathway
neutrophils
C. Evaluation
For each new cell compute its encoding, compute its similarity score and decide if it is an anomaly. If not, predict the cell type
Divide training into learning and validation. Fit LOF on learning
Compute a threshold score based on known cell types distribution over the validation
Label new data as unassigned if the similarity score is below the threshold, otherwise predict its cell type using the full model. Thus, when evaluating the prediction we have a new label (unassigned)
cell type classification
UNASSIGNED
GROUND TRUTH
PREDICTION
Neg.cell
T.CD8
T.CD4
NK
Macrophage
B.cell
unassigned
0.0013 0.069 0 0.0019 0.0096 0.92
0 0 0.014 0.038 0.91 0.035
0 0 0.0056 0.91 0.028 0.056
0 0 0.65 0 0.3 0.05
0 0.94 0 0 0 0.057
0.92 0 0 0 0 0.082

circuit / gene	...	Melanogenesis TYR*	Melanogenesis TYRP1	Melanogenesis DCT	...
...	...	...	...	...	...
dct	...	1	0	0	...
hras	...	1	1	1	...
kit	...	1	1	1	...
kras	...	1	1	1	...
kitlg	...	1	1	1	...
mitf	...	1	1	1	...
nras	...	1	1	1	...
mapk1	...	1	1	1	...
mapk3	...	1	1	1	...
map2k1	...	1	1	1	...
map2k2	...	1	1	1	...
raf1	...	1	1	1	...
tyr	...	0	0	1	...
tyrp1	...	0	1	0	...
...	...	...	...	...	...

encoding	node_1	node_2	...	node_n
sample_1	5.4154	3.118	...	4.62947
sample_2	3.53308	1.22973	...	6.67026
sample_3	5.3468	3.37469	...	2.39724
...	...	...	...	...
sample_s	0.960216	2.27654	...	2.30126

lof	similarity score
sample_1	-0.16896
sample_2	-0.59652
sample_3	-0.39456
...	...
sample_s	-1.688

Figure 1. The first step (**A**) consists of preprocessing the data, building the signaling-informed layer S, and designing the architecture of the network based on the constraints imposed by an indicator matrix $\mathbf{I}_S$ (see Methods). The second step (**B**) deals with the interpretation using the functional characterization of each cell cluster by aggregating the activations of the informed layer with respect to each observed or predicted cell type. The final step, (**C**) consists in making new predictions by (i) dividing the *training* set into *learning* and *validation*, (ii) fitting an anomaly detection algorithm to the encodings of the *learning* set, compute a threshold with *validation*, and (iii) label a new cell as unassigned if the threshold is not met, otherwise use the cell type prediction of the full NN.

2.4. Data Preprocessing

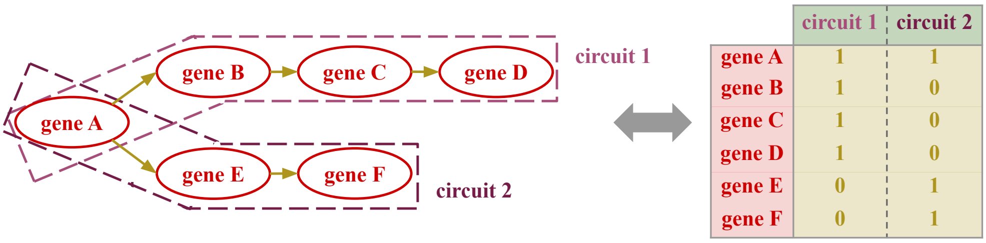
Count data is preprocessed using the Transcripts per Million (TPM) normalization method [29]. To preprocess unique molecular identifier (UMI) data we use Seurat v37 with default parameters (each cell UMI count is normalized using size-factor 10,000). In either case, we end with a gene-wise rescaling to $[-1, 1]$ after a logarithmic transformation of the preprocessed data.

2.5. Signaling-Primed Sparsity-Inducing Layers

SigPrimedNet is an ANN informed by a set of signaling circuits extracted from KEGG. Each pathway is decomposed into multiple effector circuits, so-called because they are the subpathways that end in effector proteins, which are responsible for triggering the associated function. Each effector node (a node with no descendants) defines an effector circuit along with the nodes that lead to it. To parse KEGG and decompose the resulting pathways into effector circuits we have used the HiPathia R package (v 2.11.4) [22]: the resulting (human) pathway list has been curated to remove those related to specific diseases, which totals 92 pathways that give rise to 1210 circuits (see Table A1). Note that our implementation can be extended to other pathway databases as long as each signaling pathway can be decomposed into functional subpathways.

Therefore, given a signaling pathway $\mathcal{P}$, its associated directed graph, and $\{g_0, \ldots, g_n\}$ the set of genes that belong to $\mathcal{P}$, we build the indicator matrix for $\mathcal{P}$ as follows: (i) detect the pathway effector (nodes with no descendants, $\{e_0, \ldots, e_m\}$, and receptor (nodes with no ascendants) nodes, (ii) for each effector node e, define an effector circuit $\mathcal{C}_e$ as the subgraph that contains all the receptor nodes, $\{r_0^e, \ldots, r_k^e\}$, that is connected to e, (iii) construct an indicator vector $\vec{c}_e$ where $\vec{c}_e(i) = 1$ if $g_i \in \mathcal{C}_e$, and $\vec{c}_e(i) = 0$ otherwise. Then, the indicator matrix for pathway $\mathcal{P}$ is defined as $\mathbf{I}(\mathcal{P}) = [\vec{c}_l]_{l=1}^m$. See Figure 2 for a simplified visual representation of how to build an indicator matrix.

To compose the signaling-informed layer each pathway is decomposed into its corresponding indicator matrix, which is used to build the indicator matrix $\mathbf{I}_S$ that informs the signaling layer S by performing the outer join of the previous matrices. Trivially, $\mathbf{I}_S$ is an indicator matrix with $\mathbf{I}_S(i, j) = 1$ if gene i belongs to circuit j, and $\mathbf{I}_S(i, j) = 0$ otherwise, where i and j traverse the set of all the signaling genes and circuits, respectively. This matrix informs the first hidden layer of the model: (i) the layer has as many nodes as effector circuits, (ii) the layer is initialized using Glorot uniform [30], and (iii) a weight that connects an input gene i to a node j is set to 0 if the corresponding entry in the indicator matrix is 0 (i.e. gene i does not belong to circuit j).



	circuit 1	circuit 2
gene A	1	1
gene B	1	0
gene C	1	0
gene D	1	0
gene E	0	1
gene F	0	1

Figure 2. Simplified version on how to decompose a pathway into effector circuits and build the corresponding indicator matrix. On the left side, we see a simplified pathway that gives rise to two effector sub-pathways (referenced as *effector circuits* in this work), which lead to the indicator matrix depicted on the right side.

Therefore, the kernel $\mathbf{W}_S$ of an informed layer S can be written as (Equation (4)):

$$\mathbf{W}_S = \mathbf{W} \odot \mathbf{I}_S \tag{4}$$

where $\mathbf{W}$ is a $\left(n_{genes}, n_{circuits}\right)$ real valued tensor, $\mathbf{I}_S$ is the indicator matrix of dimension $\left(n_{genes}, n_{circuits}\right)$, and $\odot$ refers to element wise (Hadamard) product.

The integration of a signaling-informed layer into the ANN has two aims: on the one hand, the sparsity induced by the informed layer has a regularization effect that prevents overfitting [17,18,20], and on the other hand, the learned representation using effector circuits provides a useful representation of the data through the associated functions, which helps to mitigate the problems associated to uninformative latent spaces [14].

2.6. Network Training and Inference

To add another source of regularization as well as to provide the model with the ability to identify unknown cell types, we split each training set into learning and validation subsets. The model is fitted (using the ADAM optimizer [31]) in a fully supervised way (the cell types are the response) to the learning set using the validation for early stopping of the training phase. Thereafter, we encode both subsets using the resulting network. A Local Outlier Factor (LOF) [32] model is fitted using the learning encodings as the features, whereas the validation set is used for setting a threshold on the similarity score. With these artificial splits, we avoid the overconfidence associated with ANN when evaluating the data where it has been fitted [33], resulting in a more realistic threshold. The threshold is set to the mean of the w measure of the similarity score distributions across the cell types, where w represents the maximum allowed deviation from the distribution set as (Equation (5)):

$$w^c = q_1^c - 1.5(q_3^c - q_1^c) \tag{5}$$

where c represents a cell type, and q_i its i-th quartile.

To predict the cell type of a new sample, the model first encodes its preprocessed transcriptomic profile, then computes the similarity score associated, decides if it is of an unknown cell type (labeling as unassigned) based on the learned threshold and, if this is not the case the model annotates the cell using the mapping function learned during the supervised training.

Therefore, we exploit the richness of the representations learned by SigPrimedNet using an unsupervised anomaly detection algorithm (LOF), which locates unusual data points by evaluating each point's local deviation from its neighbors. The LOF algorithm is based on the local density concept, in which locality is determined by K-nearest neighbors (KNN), whose distances are used for density-based scores. Finally, a point is considered an outlier if and only if the LOF score is greater than one. However, we compute a more realistic threshold by using the secondary (validation) set. See Figure 3 for a visual representation of the SigPrimedNet's prediction mechanism.

See Tables 2 and 3 for a summary of the design choices and the training and inference times, respectively.

Table 2. Hyperparameter values.

Dataset	Hyperparameter	Hyperparameter Value
PBMC	epochs	100
	batch_size	10
Immune	kernel_initializer	glorot_uniform + sig-informed
	bias_initializer	zeros
Melanoma	activation	relu (hidden layers)/softmax (last layer)
	optimizer	Adam

Table 3. Execution times.

		DESIGN	
Dataset	**Experiment**	**1-Layer**	**2-Layer**
PBMC	RepeatedStratifiedKFold (10 k-fold with 50 iterations)	mean, 3.20 min std, 0.77 min total execution time is 13.28 h	mean, 3.26 min std, 0.87 min total execution time is 13.52 h
Immune	RepeatedStratifiedKFold (10 k-fold with 30 iterations)	mean, 2.82 min std, 0.69 min total execution time is 14.07 h	mean, 4.31 min std, 1.30 min total execution time is 21.48 h
	train_test_split (50% test size with 100 iterations)	mean, 1.94 min std, 0.92 min total execution time is 3.2 h	mean, 1.79 min std, 0.51 min total execution time is 2.95 h
Melanoma	training with reference dataset (one iteration)	total execution time is 1.96 min	total execution time is 3.7 min

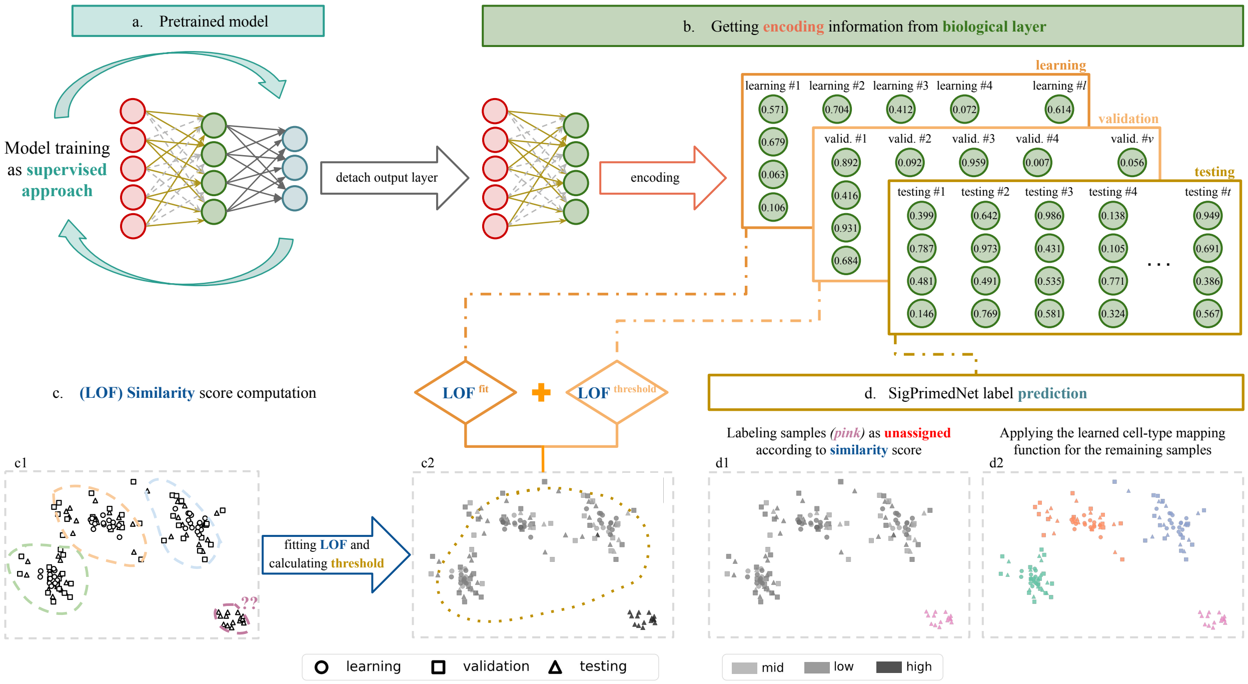

a. Pretrained model
Model training as supervised approach
detach output layer
encoding
b. Getting encoding information from biological layer
learning
learning #1 learning #2 learning #3 learning #4 learning #l
0.571 0.679 0.063 0.106 0.704 0.412 0.072 0.614
validation
valid. #1 valid. #2 valid. #3 valid. #4 valid. #v
0.892 0.416 0.931 0.684 0.092 0.959 0.007 0.056
testing
testing #1 testing #2 testing #3 testing #4 testing #t
0.399 0.787 0.481 0.146
0.642 0.973 0.491 0.769
0.986 0.431 0.535 0.581
0.138 0.105 0.771 0.324
0.949 0.691 0.386 0.567
c. (LOF) Similarity score computation
LOF fit
LOF threshold
d. SigPrimedNet label prediction
Labeling samples (pink) as unassigned according to similarity score
Applying the learned cell-type mapping function for the remaining samples
c1
??
fitting LOF and calculating threshold
c2
d1
d2
○ learning □ validation △ testing
mid low high

Figure 3. SigPrimedNet's prediction mechanism (in two dimensions to simplify). Given the fitted NN (**a**), we detach the output layer and compute the encodings of the learning, validation and test sets (**b**), Then, in (**c1**) we fit a LOF to model the training encodings, (**c2**) predict the similarity scores of the learning and validation encodings to compute a more realistic threshold. Finally, (**d1**) a sample is labeled as unassigned if its similarity score is bellow the threshold, if not (**d2**) we apply the cell-type mapping learned by the NN.

2.7. Functional Proxies and Representation Learning

Once the model has been fitted to a collection of annotated cells, we extract the features learned by the ANN, also known as representation learning [34], by detaching the last layer and computing the activations of the encoding layer. As the model has learned to map the gene profiles to the cell types, the encoding layer captures a lower-dimensional representation of the data necessary for the mapping. Note that, if the activations are computed for the signaling-informed layer we obtain a functional representation of the data as the nodes act as a proxy for the effector circuits. For visualization purposes, we can map the encodings to a 2D space by using TSNE (See Supplementary material).

3. Results and Discussion

We provide here a series of validation procedures to test the performance of SigPrimeNet under different scenarios: a synthetically balanced data set based on PBMC where all cell types are known, a synthetic collection of unbalanced data sets made by undersampling each of the cell types that appear in PBMC, a real-world unbalanced data set (Immune) where the cell types are known and a data set (Melanoma) built for benchmark unknown cell-type identification methods. In the Supplementary material, we also provide results for a two-layer version of SigPrimedNet (adding a second dense hidden layer), and a set of experiments designed to showcase the supervised performance of SigPrimedNet with the aim of making it easier to compare it to other methods that lack the ability to identify unknown cell types.

3.1. Model Performance When All Cell Types Are Known

3.1.1. Synthetically Balanced PBMC

We tested the performance of our method employing 50 times repeated stratified by cell type 10-fold cross-validation schema using the balanced PBMC dataset (see Materials). The confusion matrix aggregated across the test folds shows that SigPrimedNet has a high ability to distinguish between cell types, as can be seen in Figure 4. In general, SigPrimedNet exhibits excellent performance across all the cell types with a slight decrease when dealing with those that are very closely related, such as Memory and Regulatory T cells. It should be noted that these results are similar to those obtained by Scibet and better than those obtained by Seurat and Scmap in a similar experiment shown in [15] since our approach reduces the misclassification of cytotoxic T cells.

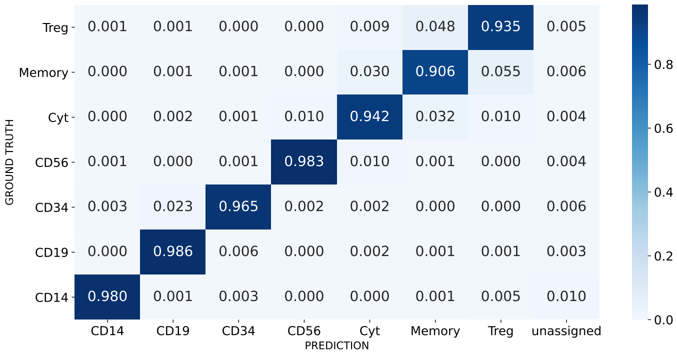


Figure 4. PBMC aggregated cross-validation confusion matrix. The unassigned label refers to cells that the model could not assign a known cell type.

3.1.2. Synthetically Unbalanced PBMC

To check the model performance in unbalanced scenarios, while still holding some control over the cell populations, we have randomly undersampled one cell type at a time in the PBMC dataset for different undersampling ratios (0.2, 0.4, 0.6). Then, we evaluated the performance of SigPrimedNet using a 10-fold cross-validation schema for each of the simulated datasets.

Figure 5 shows the aggregated cross-validation matrix for each synthetic dataset. As expected, underpopulated cell types lead to a decrease in the predictive power with respect to the minority class for those cell types that were hard to classify originally (Treg, Memory), while the performance of cytotoxic T cells is severely hampered when undersampling their population, similar to the experiments conducted with the balanced PBMC dataset in [15]. However, the performance of the model over the other known types remains at levels equivalent to those obtained when evaluating SigPrimedNet in the balanced scenario. In addition, the rate of cells incorrectly labeled as unassigned remains in the same range as in the balanced simulation.

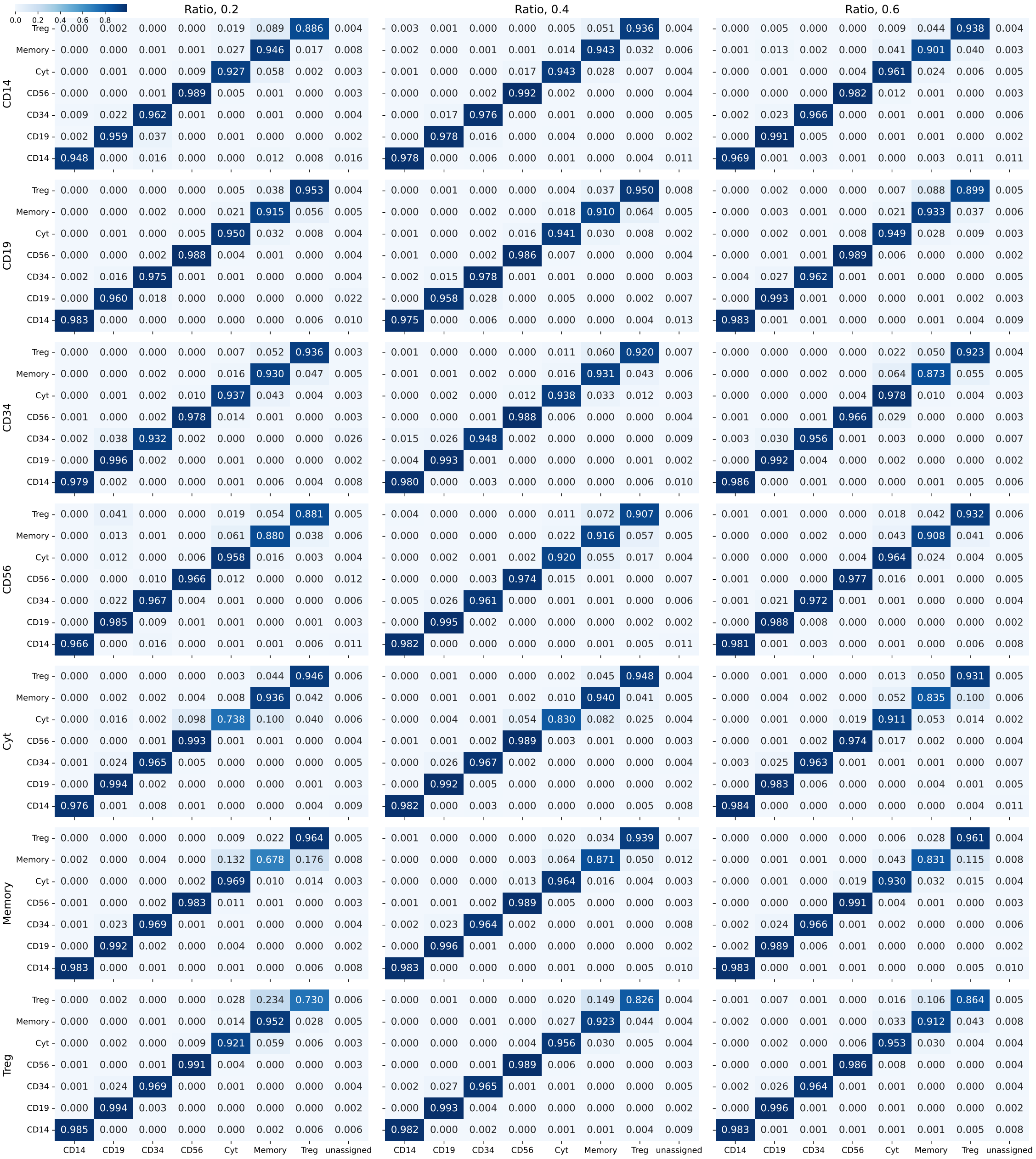


Figure 5. Summarized confusion matrices for the different unbalanced simulations using the PBMC dataset. The rows represent the cell type being undersampled, while the columns are the fraction of cells kept for each cell type.

3.1.3. Real-World Unbalanced Scenario

To check the performance of our model in a class-imbalanced scenario we performed 30 times cell-type stratified repeated 10-fold cross-validation using the Inmune dataset. Despite the added difficulty due to the disproportion between the classes, a ratio of 7.72 between the highest (HSPCs) and lowest (T cells) populated cell types, our method could still provide a high discriminating power as can be observed in the aggregated confusion matrix depicted in Figure 6. Most misclassifications are cells incorrectly labeled as HSPCs, which could be explained in machine learning terms, as a bias towards the majority class (HSPCs), or in biological terms, since HSPCs are very heterogeneous with transcriptomic profiles that match other cell types patterns [35,36]. Furthermore, the proportion of incorrectly labeled cells remains low as shown in the PBMC experiments.

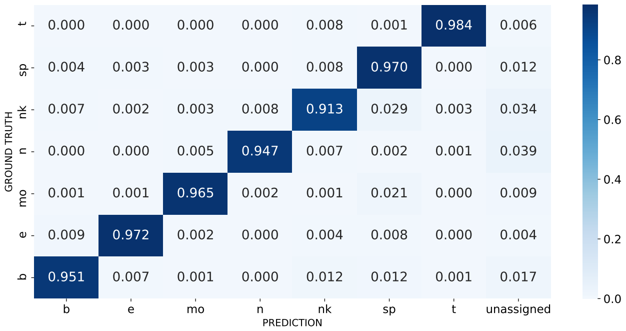


Figure 6. The aggregated cross-validation confusion matrix of the Immune dataset. The cell types b, e, mo, n, nk, sp, and t refer to B cells, erythrocytes, monocytes, neutrophils, NK cells, CD34+ HSPCs, and T cells, respectively. The unassigned label refers to cells that the model that could not correctly assign a known cell type.

3.1.4. Design Comparison

The expressive power of the informed layer is evident when comparing the results of the two designs tested in this paper: the model's performance is not noticeably improved by adding more capacity to the network by including a dense layer. Thus, the signaling-informed layer is capable of constructing, by itself, the necessary meta-features to differentiate cell types from the point of view of cell signaling. This can be deducted by inspecting Figure 7: the recall and the proportion of cells with an assigned label are higher in the one-layer design, while the precision is similar for both designs. Note that we have used the weighted version of precision and recall to account for the label imbalance. See Supplementary Material for the complete set of results for the two-layer design.

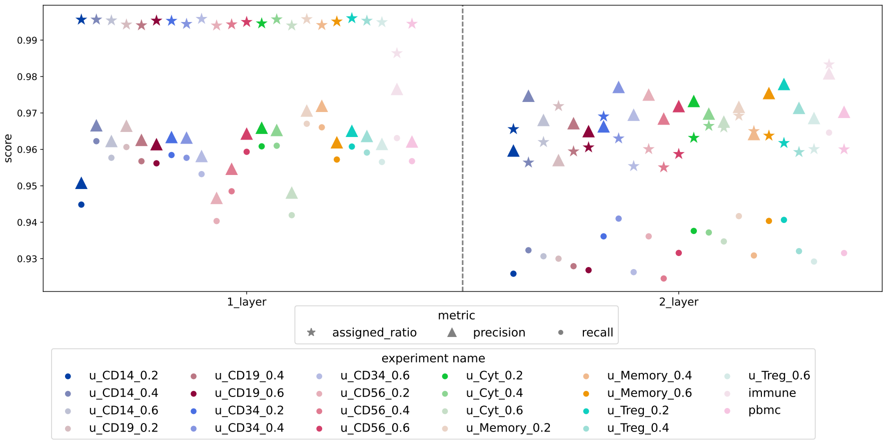


Figure 7. The colors represent the datasets used for testing the performance. In the legend "immune" refers to the Immune experiment, "pmbc" to the balanced PBMC experiment, and u_cell-type_ratio refers to the undersampled version of the PBMC where the "ratio" indicates the proportion of cells of sampled cells a given "cell-type". Triangles and dots represent the mean across the test sets of the weighted precision and recall scores, respectively, while the crosses resent the mean of the proportion of cells assigned a cell type (since all cells are known, the higher the better).

3.2. *Unknown Cell-Type Identification*

Novelty Detection in the Melanoma Dataset

Due to the incomplete nature of the reference scRNA-seq data, cell types not present in the reference dataset may be falsely predicted as those used during the model training. To analyze our approach to this issue, we used the Melanoma dataset with immune cells as positive cells and the other cells as negative cells. Figure 8 depicts the confusion matrix for the case study of false-positive control, with normalization for each row (origin label): the task consists in annotating the negative cells as unassigned while assigning the corresponding label to the other cells. Note that negative cells including malignant cells, CAF cells, and endothelial cells were removed from the training set. Query cells identified as anomalies by SigPrimedNet were labeled as unassigned. The results show that our method consistently outperforms Scibet for all the known labels (except NK) while maintaining a similar false-positive ratio. As mentioned in the Datasets section, this experiment was designed for this specific task in [15], and we have been able to reproduce it with both models: Scibet and SigPrimedNet. This NK deficit could be easily understood as it is something shared across all the experiments conducted: very low-populated cell types are harder to classify.

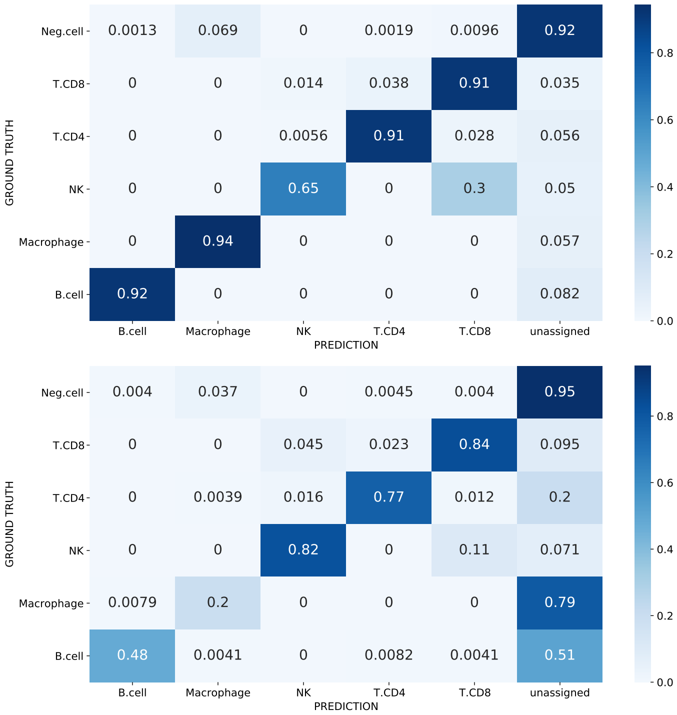


Figure 8. The confusion matrix of the melanoma dataset for the unknown cell-type identification task. SigPrimedNet (top) and Scibet (bottom).

The Melanoma dataset was used in our previous study with a limited pathway-driven neural network (PDNN) [21], which only works for supervised tasks. The performance is similar between PDNN and SigPrimedNet with balanced accuracy scores of 0.844 and 0.8837 for the test split, respectively (see Table 4 for a more comprehensive comparison). Note that the results are not fully comparable since the dataset was filtered in [21] by removing all the negative cells in order to be able to use the PDNN (which results in a more favorable scenario for supervised models, like PDNN). Note that if the PDNN is used when unknown cells are present, it would label all the unknown cells with one of the known labels (a critical limitation), which can be assessed by looking at the PDNN (*) entry in Table 4 where we have run the PDNN model on the full dataset. This is not the case for SigPrimedNet, where unknown cell types are properly labeled as “unassigned”.

Figure 9 shows the distribution of similarity scores. The graph shows the similarity scores computed by fitting the ANN to the learning set (70% of the reference), then we fit a LOF model using their encodings as features, and finally, a threshold is learned to use the similarity scores of the remaining 30% validation (blue colored). When assigning labels to the test (query unseen cells), the first step is computing the ANN encodings followed by the LOF scores (ocher colored): those scores below the threshold are labeled as unassigned, and the remaining cells are assigned the cell type using the mapping learned by SigPrimedNet. Supplementary material shows an analogous result for the two-layer architecture, although the performance is worse than the one-layer interpretable design used here.

Table 4. Comparison between SigPrimedNet (one-layer (1L) and two-layer (2L) designs) and the best PDNN design on the Melanoma test split.

	MACRO			WEIGHTED				
Design	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**Accuracy**	**Balanced Accuracy**
SigPrimedNet (1L)	0.838	0.823	0.884	0.926	0.945	0.919	0.919	0.884
SigPrimedNet (2L)	0.743	0.785	0.796	0.878	0.927	0.846	0.846	0.796
PDNN	0.861	0.922	0.844	0.933	0.938	0.936	0.936	0.844
PDNN (*)	0.499	0.454	0.753	0.241	0.224	0.326	0.326	0.753

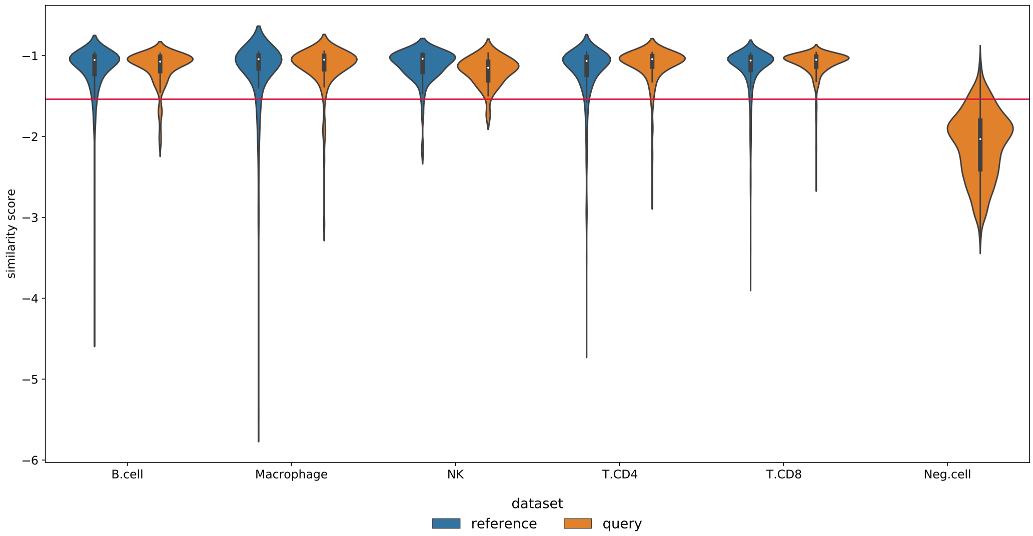


Figure 9. Similarity score distribution for each cell type on the validation and test splits Melanoma dataset). The horizontal line shows the threshold obtained using the reference set inner splits as detailed in Section 2.6.

3.3. SigPrimedNet Provides Biologically Interpretable Results

To illustrate the potential of our approach in producing biologically interpretable results, we have selected, for each cell type, the ten highest-weighted nodes (Table A2) from the signaling-informed layer, each representing a circuit from KEGG (Table A1).

For example, circuit Hedgehog signaling pathway (hsa04340): GLI SUFU is known to be involved in the control of hematopoietic differentiation [37], and it is present in the rank for HSPCs, NKs, erythrocytes, and B cells. The GO annotations for this circuit include cell differentiation and cell proliferation.

Another example is circuit Hippo signaling pathway (hsa04390): SERPINE1, which is present in ranks for B cells, NKs and HSPCs. Ref. [38] describe the role of SERPINE1 in the regulation of immune-related biological processes in glioma, relating high expression of SERPINE1 to gene expression patterns enriched in immune-related signaling pathways such as B cell receptor signaling pathway, Natural Killer cell mediated cytotoxicity, primary immunodeficiency, and T cell receptor signaling pathway, among others. Additionally, ref. [39] describe a mechanism by which PAI-1 (SERPINE1) regulates the localization of HSPCs between the bone marrow or its migration to other tissues. HSPCs also list circuits regulating cell survival and cell adhesion, relevant for their proliferative activity as hematopoietic cell precursors.

We also find that the circuit Calcium signaling pathway (hsa04020): Sphingosine 1-phosphate is active for monocytes and neutrophils. Ref. [40] provide evidence for the need for SphK2 kinase in processes of intracellular catalytic lipid degradation, which should be necessary for the phagocytic activity of monocytes and neutrophils.

Neutrophils also list circuits related to secretion and cellular mobility. Interestingly, among the neutrophils rank we find a circuit from the Melanogenesis pathway (hsa04916): DCT, that is implied in tyrosine metabolism. Neutrophils release several types of amino acids upon adhesion and spreading onto fibronectin, a process especially relevant in tissues undergoing healing and regeneration processes [41]. This connects to the already described link of neutrophilic activity to inflammation-related skin pigmentation [42].

The rank for Erythrocytes' top 10 includes circuits related to the regulation of the location of precursor cells during hematopoiesis, regulation of apoptosis, and transendothelial migration. It also includes circuits that could regulate erythrocyte micromechanical properties and fluidity, which are necessary to adapt the size and viscosity enabling circulation through thin terminal capillaries [43].

The top-ranked circuits for T cells include the circuit Cell cycle: TFDP1 E2F4, which regulates a circuit regulating the entry of cells in the S-phase of the cell cycle. Although T cell selection occurs in the thymus, there is evidence that they undergo further differentiation in peripheral tissues [44].

4. Conclusions

SigPrimedNet is a highly efficient neural network for cell-type annotation in single-cell transcriptomics, one of the main challenges arising from a field with exponential growth, while providing useful biological features. The tool has been successfully tested on three tasks, namely: supervised cell type classification, unknown cell type annotation, and representation learning usefulness. To that effect, different publicly available benchmarks on multiple datasets have been carried out, with an outstanding known cell-type annotation performance while keeping a low false-positive rate for cell types unknown to the model. The ability to successfully identify cells of unknown origin lies in the high expressiveness of the features learned by the neural network, which are successfully used to train an unsupervised secondary model that detects anomalous cell types. These features are proxies for the signaling circuits used to inform the layers of the model, which have a regularization benefit due to weight sparsification and present a meaningful set of biological functions. The model has very low latency when annotating new cells and provides rich and useful interpretable features.

Supplementary Materials: The following are available at https://www.mdpi.com/article/10.3390/biology12040579/s1, Figure S1: 2D TSNE visualization of the features learned by SigPrimedNet for a test split of the Immune dataset. The cell types b, e, mo, n, nk, sp, and t refer to B cells, erythrocytes, monocytes, neutrophils, NK cells, CD34+ HSPCs, and T cells, respectively; Figure S2: SigPrimedNet with two-layer design; Figure S3: The confusion matrix of the Melanoma dataset for the unknown cell-type identification tas; Figure S4: Similarity score distribution for each cell type on the validation and test splits using the two-layer architecture (Melanoma dataset). The horizontal line shows the threshold obtained using the reference set inner splits as detailed in the Methods section of the main manuscript; Figure S5: The confusion matrix of the Immune dataset using SigPrimedNet with 2 layers; Figure S6: The confusion matrix of the PBMC balanced dataset using SigPrimedNet with 2 layers; Figure S7: The confusion matrix of the PBMC unbalanced dataset using SigPrimedNet with 2 layers; Figure S8: PBMC experiment aggregated cross-validation confusion matrix for SigPrimedNet (1-layer design); Figure S9: PBMC experiment aggregated cross-validation confusion matrix for SigPrimedNet (2-layer design); Figure S10: Performance of SigPrimedNet (1 and 2 layer designs) for the PBMC experiment: F1, Precision and Recall score distribution across the test sets of 50 times repeated 10-fold cross-validation; Figure S11: Performance of SigPrimedNet (1 and 2 layer designs) for the PBMC experiment: F1, Precision and Recall score distribution across each cell type of the test set of 50 times repeated 10-fold cross-validation; Figure S12: The aggregated confusion matrix of the Immune dataset (1-layer design for the reduced model); Figure S13: The aggregated confusion matrix of the Immune dataset (2-layer design for the reduced model); Figure S14: SigPrimedNet overall performance for Immune datase; Figure S15: SigPrimedNet performance desegregated for each cell type for Immune dataset; Table S1: Cell type, number of samples detail, and percentage of samples above or below the encoding-based threshold of Melanoma dataset during the testing phase. Note that Neg. cells including malignant cells, CAF cells, and endothelial cells were removed from the training set (see Materials).

Author Contributions: Conceptualization, I.A.N.-C., J.D. and C.L.; methodology, P.G., I.A.N.-C. and C.L.; software, P.G.; formal analysis, P.G., I.A.N.-C. and C.L.; biological interpretation (validation), I.A. and I.A.N; writing—original draft preparation, P.G., I.A.N.-C. and C.L.; writing—review and editing, J.D. and C.L.; supervision, C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work is 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), 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.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The full version of the PBMC dataset can be obtained from 10× Genomics [24], the Melanoma dataset can be obtained from GEO [23], and the Immune dataset can be obtained from [26]. Finally, the filtered versions used in this study can be obtained from [15] (PBMC and Melanoma) and [25] (Immune).

Acknowledgments: The authors thankfully acknowledge all members of SciBet for their helpful comments and guidelines to reproduce their experiments.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

scRNA-seq	Single-cell RNA sequencing
HVGs	Highly variable genes
PCA	Principal Component Analysis
UMAP	Uniform Manifold Approximation and Projection
ANN	Artificial Neural Network
UMI	Unique Molecular Identifier
KEGG	Kyoto Encyclopedia of Genes and Genomes Database
TSNE	t-distributed stochastic neighbor embedding
TPM	Transcripts per Million
LOF	Local Outlier Factor
KNN	K-nearest neighbors
PDNN	Pathway-driven Neural Network
MACRO	Unweighted average across cell types for any given classification metric
WEIGHTED	Support-weighted average across cell types for any given classification metric

Appendix A

In this appendix, we provide several tables that are useful to understand the model's prior knowledge and the most biologically relevant results.

Table A1. List of pathways used as the sources for the circuit decomposition to encode the prior knowledge of SigPrimedNet.

KeggID	Pathway Name	KeggID	Pathway Name	KeggID	Pathway Name
hsa03320	PPAR signaling pathway	hsa04370	VEGF signaling pathway	hsa04727	GABAergic synapse
hsa04010	MAPK signaling pathway	hsa04380	Osteoclast differentiation	hsa04728	Dopaminergic synapse
hsa04012	ErbB signaling pathway	hsa04390	Hippo signaling pathway	hsa04740	Olfactory transduction
hsa04014	Ras signaling pathway	hsa04510	Focal adhesion	hsa04742	Taste transduction
hsa04015	Rap1 signaling pathway	hsa04520	Adherens junction	hsa04750	Inflammatory mediator regulation of TRP channels
hsa04020	Calcium signaling pathway	hsa04530	Tight junction	hsa04810	Regulation of actin cytoskeleton
hsa04022	cGMP-PKG signaling pathway	hsa04540	Gap junction	hsa04910	Insulin signaling pathway
hsa04024	cAMP signaling pathway	hsa04550	Signaling pathways regulating pluripotency of stem cells	hsa04911	Insulin secretion
hsa04062	Chemokine signaling pathway	hsa04610	Complement and coagulation cascades	hsa04912	GnRH signaling pathway
hsa04064	NF-kappa B signaling pathway	hsa04611	Platelet activation	hsa04913	Ovarian steroidogenesis
hsa04066	HIF-1 signaling pathway	hsa04612	Antigen processing and presentation	hsa04914	Progesterone-mediated oocyte maturation
hsa04068	FoxO signaling pathway	hsa04620	Toll-like receptor signaling pathway	hsa04915	Estrogen signaling pathway
hsa04071	Sphingolipid signaling pathway	hsa04621	NOD-like receptor signaling pathway	hsa04916	Melanogenesis
hsa04072	Phospholipase D signaling pathway	hsa04622	RIG-I-like receptor signaling pathway	hsa04917	Prolactin signaling pathway
hsa04110	Cell cycle	hsa04623	Cytosolic DNA-sensing pathway	hsa04918	Thyroid hormone synthesis
hsa04114	Oocyte meiosis	hsa04630	Jak-STAT signaling pathway	hsa04919	Thyroid hormone signaling pathway
hsa04115	p53 signaling pathway	hsa04650	Natural killer cell mediated cytotoxicity	hsa04920	Adipocytokine signaling pathway
hsa04150	mTOR signaling pathway	hsa04660	T cell receptor signaling pathway	hsa04921	Oxytocin signaling pathway
hsa04151	PI3K-Akt signaling pathway	hsa04662	B cell receptor signaling pathway	hsa04922	Glucagon signaling pathway
hsa04152	AMPK signaling pathway	hsa04664	Fc epsilon RI signaling pathway	hsa04923	Regulation of lipolysis in adipocytes
hsa04210	Apoptosis	hsa04666	Fc gamma R-mediated phagocytosis	hsa04924	Renin secretion
hsa04211	Longevity regulating pathway - mammal	hsa04668	TNF signaling pathway	hsa04925	Aldosterone synthesis and secretion
hsa04213	Longevity regulating pathway - multiple species	hsa04670	Leukocyte transendothelial migration	hsa04960	Aldosterone-regulated sodium reabsorption
hsa04218	Cellular senescence	hsa04710	Circadian rhythm	hsa04961	Endocrine and other factor-regulated calcium reabsorption
hsa04261	Adrenergic signaling in cardiomyocytes	hsa04713	Circadian entrainment	hsa04962	Vasopressin-regulated water reabsorption
hsa04270	Vascular smooth muscle contraction	hsa04720	Long-term potentiation	hsa04970	Salivary secretion
hsa04310	Wnt signaling pathway	hsa04722	Neurotrophin signaling pathway	hsa04971	Gastric acid secretion
hsa04330	Notch signaling pathway	hsa04723	Retrograde endocannabinoid signaling	hsa04972	Pancreatic secretion
hsa04340	Hedgehog signaling pathway	hsa04724	Glutamatergic synapse	hsa04973	Carbohydrate digestion and absorption
hsa04350	TGF-beta signaling pathway	hsa04725	Cholinergic synapse	hsa04976	Bile secretion
hsa04360	Axon guidance	hsa04726	Serotonergic synapse	hsa05100	Bacterial invasion of epithelial cells

Table A2. List of the most relevant signaling circuits (defined as pathway:effector protein) for each cell type in the Immune dataset.

CT	KeggID	Circuit Name	CT	KeggID	Circuit Name
B cells	hsa03320	PPAR signaling pathway: DBI	NKs	hsa04115	p53 signaling pathway: TP73
	hsa04670	Leukocyte transendothelial migration: CDH5		hsa04151	PI3K-Akt signaling pathway: EIF4B
	hsa04666	Fc gamma R-mediated phagocytosis: PLA2G4B		hsa04390	Hippo signaling pathway: SERPINE1
	hsa04115	p53 signaling pathway: CD82		hsa04152	AMPK signaling pathway: CCNA2
	hsa04340	Hedgehog signaling pathway: GLI1 SUFU		hsa04151	PI3K-Akt signaling pathway: CDKN1B
	hsa04390	Hippo signaling pathway: SERPINE1		hsa04520	Adherens junction: LEF1 CTNNB1
	hsa04064	NF-kappa B signaling pathway: PLCG2		hsa04210	Apoptosis: BID
	hsa04724	Glutamatergic synapse: ADRBK1		hsa04064	NF-kappa B signaling pathway: PLCG2
	hsa04115	p53 signaling pathway: TP73		hsa04340	Hedgehog signaling pathway: GLI1 SUFU
	hsa04620	Toll-like receptor signaling pathway: CCL5		hsa04620	Toll-like receptor signaling pathway: CCL5
Erythrocytes	hsa04724	Glutamatergic synapse: MAPK1	HSPCs	hsa04340	Hedgehog signaling pathway: GLI1 SUFU
	hsa05100	Bacterial invasion of epithelial cells: ACTB		hsa04390	Hippo signaling pathway: SERPINE1
	hsa04115	p53 signaling pathway: CD82		hsa04724	Glutamatergic synapse: ADRBK1
	hsa04340	Hedgehog signaling pathway: GLI1 SUFU		hsa04151	PI3K-Akt signaling pathway: EIF4B
	hsa03320	PPAR signaling pathway: FADS2		hsa04919	Thyroid hormone signaling pathway: SLC9A1
	hsa04110	Cell cycle: ORC3 ORC5 ORC4 ORC2 ORC1 ORC6 MCM7 MCM6 MCM5 MCM4 MCM3 MCM2		hsa04520	Adherens junction: LEF1 CTNNB1
	hsa04670	Leukocyte transendothelial migration: ACTB CTNNA1 CTNNB1		hsa04210	Apoptosis: BID
	hsa04810	Regulation of actin cytoskeleton: MYL12B MYH9 ACTB		hsa04740	Olfactory transduction: PDE2A
	hsa04014	Ras signaling pathway: PLCE1		hsa04064	NF-kappa B signaling pathway: TNFSF13B
	hsa03320	PPAR signaling pathway: DBI		hsa04014	Ras signaling pathway: PAK4
Monocytes	hsa04110	Cell cycle: CDC6 ORC3 ORC5 ORC4 ORC2 ORC1 ORC6	T cells	hsa04110	Cell cycle: CDC6 ORC3 ORC5 ORC4 ORC2 ORC1 ORC6
	hsa04740	Olfactory transduction: PDE2A		hsa04919	Thyroid hormone signaling pathway: THRA Triiodothyronine
	hsa04022	cGMP-PKG signaling pathway: ITPR1		hsa04724	Glutamatergic synapse: MAPK1
	hsa04670	Leukocyte transendothelial migration: CDH5		hsa04713	Circadian entrainment: PRKCA
	hsa04914	Progesterone-mediated oocyte maturation: CDK1		hsa04666	Fc gamma R-mediated phagocytosis: ARF6
	hsa04210	Apoptosis: BCL2L1		hsa04014	Ras signaling pathway: PLCE1
	hsa05100	Bacterial invasion of epithelial cells: ACTB		hsa04650	Natural killer cell mediated cytotoxicity: TNFRSF10D
	hsa04915	Estrogen signaling pathway: ESR1 Estradiol-17beta		hsa04024	cAMP signaling pathway: LIPE
	hsa04668	TNF signaling pathway: DNM1L		hsa04110	Cell cycle: TFDP1 E2F4
	hsa04020	Calcium signaling pathway: Sphingosine 1-phosphate		hsa04919	Thyroid hormone signaling pathway: SLC9A1
Neutrophyls	hsa04668	TNF signaling pathway: DNM1L		hsa04919	Thyroid hormone signaling pathway: NOTCH1
	hsa04915	Estrogen signaling pathway: ESR1 Estradiol-17beta		hsa04020	Calcium signaling pathway: Sphingosine 1-phosphate
	hsa04916	Melanogenesis: DCT		hsa04660	T cell receptor signaling pathway: CD40LG
	hsa04970	Salivary secretion: KCNN4		hsa04915	Estrogen signaling pathway: CREB3
	hsa04014	Ras signaling pathway: RHOA		hsa04340	Hedgehog signaling pathway: SMO

Appendix B

In this appendix, we provide the software used to carry out the proposed experiments. Note that we show only major libraries, the full conda environment specification can be obtained from the project's repository (v.1.0.0 refers to the version described in this work): https://github.com/babelomics/sigprimednet/releases/tag/v1.0.0 (accessed on 9 April 2023).

- Python 3.8
- scikit-learn (v 0.24.1) [45]
- numpy (v 1.19.2) [46]
- scipy (v 1.6.0) [47]
- tensorflow (v 2.2.0) [48]

References

1. Alavi, A.; Ruffalo, M.; Parvangada, A.; Huang, Z.; Bar-Joseph, Z. A Web Server for Comparative Analysis of Single-Cell RNA-seq Data. *Nat. Commun.* **2018**, *9*, 4768. [CrossRef] [PubMed]
2. AlJanahi, A.A.; Danielsen, M.; Dunbar, C.E. An Introduction to the Analysis of Single-Cell RNA-Sequencing Data. *Mol. Ther. Methods Clin. Dev.* **2018**, *10*, 189–196. [CrossRef] [PubMed]
3. Kiselev, V.Y.; Yiu, A.; Hemberg, M. Scmap: Projection of Single-Cell RNA-seq Data across Data Sets. *Nat. Methods* **2018**, *15*, 359–362. [CrossRef] [PubMed]

4. Brennecke, P.; Anders, S.; Kim, J.K.; Kołodziejczyk, A.A.; Zhang, X.; Proserpio, V.; Baying, B.; Benes, V.; Teichmann, S.A.; Marioni, J.C.; et al. Accounting for Technical Noise in Single-Cell RNA-seq Experiments. *Nat. Methods* **2013**, *10*, 1093–1095. [CrossRef]
5. Andrews, T.S.; Hemberg, M. M3Drop: Dropout-based feature selection for scRNASeq. *Bioinformatics* **2019**, *35*, 2865–2867. [CrossRef]
6. Tsuyuzaki, K.; Sato, H.; Sato, K.; Nikaido, I. Benchmarking Principal Component Analysis for Large-Scale Single-Cell RNA-sequencing. *Genome Biol.* **2020**, *21*, 9. [CrossRef]
7. Pierson, E.; Yau, C. ZIFA: Dimensionality Reduction for Zero-Inflated Single-Cell Gene Expression Analysis. *Genome Biol.* **2015**, *16*, 241. [CrossRef]
8. van der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
9. Becht, E.; McInnes, L.; Healy, J.; Dutertre, C.A.; Kwok, I.W.H.; Ng, L.G.; Ginhoux, F.; Newell, E.W. Dimensionality Reduction for Visualizing Single-Cell Data Using UMAP. *Nat. Biotechnol.* **2019**, *37*, 38–44. [CrossRef]
10. Lopez, R.; Regier, J.; Cole, M.B.; Jordan, M.I.; Yosef, N. Deep Generative Modeling for Single-Cell Transcriptomics. *Nat. Methods* **2018**, *15*, 1053–1058. [CrossRef]
11. Lotfollahi, M.; Naghipourfar, M.; Luecken, M.D.; Khajavi, M.; Büttner, M.; Wagenstetter, M.; Avsec, Ž.; Gayoso, A.; Yosef, N.; Interlandi, M.; et al. Mapping Single-Cell Data to Reference Atlases by Transfer Learning. *Nat. Biotechnol.* **2022**, *40*, 121–130. [CrossRef] [PubMed]
12. Kang, J.B.; Nathan, A.; Weinand, K.; Zhang, F.; Millard, N.; Rumker, L.; Moody, D.B.; Korsunsky, I.; Raychaudhuri, S. Efficient and Precise Single-Cell Reference Atlas Mapping with Symphony. *Nat. Commun.* **2021**, *12*, 5890. [CrossRef] [PubMed]
13. Korsunsky, I.; Millard, N.; Fan, J.; Slowikowski, K.; Zhang, F.; Wei, K.; Baglaenko, Y.; Brenner, M.; Loh, P.R.; Raychaudhuri, S. Fast, Sensitive and Accurate Integration of Single-Cell Data with Harmony. *Nat. Methods* **2019**, *16*, 1289–1296. [CrossRef] [PubMed]
14. Lotfollahi, M.; Rybakov, S.; Hrovatin, K.; Hediyeh-zadeh, S.; Talavera-López, C.; Misharin, A.V.; Theis, F.J. Biologically Informed Deep Learning to Infer Gene Program Activity in Single Cells. *Nat. Cell Biol.* **2023**, *25*, 337–350. [CrossRef]
15. Li, C.; Liu, B.; Kang, B.; Liu, Z.; Liu, Y.; Chen, C.; Ren, X.; Zhang, Z. SciBet as a Portable and Fast Single Cell Type Identifier. *Nat. Commun.* **2020**, *11*, 1818. [CrossRef]
16. Sun, X.; Lin, X.; Li, Z.; Wu, H. A Comprehensive Comparison of Supervised and Unsupervised Methods for Cell Type Identification in Single-Cell RNA-seq. *Brief. Bioinform.* **2022**, *23*, bbab567. [CrossRef]
17. Xu, Q.; Zhang, M.; Gu, Z.; Pan, G. Overfitting Remedy by Sparsifying Regularization on Fully-Connected Layers of CNNs. *Neurocomputing* **2019**, *328*, 69–74. [CrossRef]
18. Elmarakeby, H.A.; Hwang, J.; Arafeh, R.; Crowdis, J.; Gang, S.; Liu, D.; AlDubayan, S.H.; Salari, K.; Kregel, S.; Richter, C.; et al. Biologically Informed Deep Neural Network for Prostate Cancer Discovery. *Nature* **2021**, *598*, 348–352. [CrossRef]
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. *Nat. Methods* **2018**, *15*, 290–298. [CrossRef]
20. Lin, C.; Jain, S.; Kim, H.; Bar-Joseph, Z. Using Neural Networks for Reducing the Dimensions of Single-Cell RNA-Seq Data. *Nucleic Acids Res.* **2017**, *45*, e156. [CrossRef]
21. 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 Min.* **2022**, *15*, 1. [CrossRef] [PubMed]
22. 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* **2016**, *8*, 5160–5178. [CrossRef] [PubMed]
23. Barrett, T.; Wilhite, S.E.; Ledoux, P.; Evangelista, C.; Kim, I.F.; Tomashevsky, M.; Marshall, K.A.; Phillippy, K.H.; Sherman, P.M.; Holko, M.; et al. NCBI GEO: Archive for Functional Genomics Data Sets—Update. *Nucleic Acids Res.* **2013**, *41*, D991–D995. [CrossRef] [PubMed]
24. Zheng, G.X.Y.; Terry, J.M.; Belgrader, P.; Ryvkin, P.; Bent, Z.W.; Wilson, R.; Ziraldo, S.B.; Wheeler, T.D.; McDermott, G.P.; Zhu, J.; et al. Massively Parallel Digital Transcriptional Profiling of Single Cells. *Nat. Commun.* **2017**, *8*, 14049. [CrossRef] [PubMed]
25. Xie, X.; Liu, M.; Zhang, Y.; Wang, B.; Zhu, C.; Wang, C.; Li, Q.; Huo, Y.; Guo, J.; Xu, C.; et al. Single-Cell Transcriptomic Landscape of Human Blood Cells. *Natl. Sci. Rev.* **2021**, *8*, nwaa180. [CrossRef] [PubMed]
26. Kivioja, T.; Vähärautio, A.; Karlsson, K.; Bonke, M.; Enge, M.; Linnarsson, S.; Taipale, J. Counting Absolute Numbers of Molecules Using Unique Molecular Identifiers. *Nat. Methods* **2012**, *9*, 72–74. [CrossRef]
27. Kanehisa, M.; Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30. [CrossRef]
28. Nair, V.; Hinton, G.E. Rectified Linear Units Improve Restricted Boltzmann Machines. In Proceedings of the 27th International Conference on International Conference on Machine Learning, Madison, WI, USA, 21–24 June 2010; pp. 807–814.
29. Wagner, G.P.; Kin, K.; Lynch, V.J. Measurement of mRNA Abundance Using RNA-seq Data: RPKM Measure Is Inconsistent among Samples. *Theory Biosci. = Theor. Den Biowiss.* **2012**, *131*, 281–285. [CrossRef]
30. Glorot, X.; Bengio, Y. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, Sardinia, Italy, 13–15 May 2010; pp. 249–256.
31. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2017**, arXiv:cs/1412.6980.

32. Breunig, M.M.; Kriegel, H.P.; Ng, R.T.; Sander, J. LOF: Identifying Density-Based Local Outliers. *ACM Sigmod Rec.* **2000**, *29*, 93–104. [CrossRef]
33. Hein, M.; Andriushchenko, M.; Bitterwolf, J. Why Relu Networks Yield High-Confidence Predictions Far Away from the Training Data and How to Mitigate the Problem. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 41–50.
34. Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [CrossRef] [PubMed]
35. Stumpf, P.S.; Du, X.; Imanishi, H.; Kunisaki, Y.; Semba, Y.; Noble, T.; Smith, R.C.G.; Rose-Zerili, M.; West, J.J.; Oreffo, R.O.C.; et al. Transfer Learning Efficiently Maps Bone Marrow Cell Types from Mouse to Human Using Single-Cell RNA Sequencing. *Commun. Biol.* **2020**, *3*, 736. [CrossRef] [PubMed]
36. Velten, L.; Haas, S.F.; Raffel, S.; Blaszkiewicz, S.; Islam, S.; Hennig, B.P.; Hirche, C.; Lutz, C.; Buss, E.C.; Nowak, D.; et al. Human Haematopoietic Stem Cell Lineage Commitment Is a Continuous Process. *Nat. Cell Biol.* **2017**, *19*, 271–281. [CrossRef] [PubMed]
37. Detmer, K.; Walker, A.N.; Jenkins, T.M.; Steele, T.A.; Dannawi, H. Erythroid Differentiation in Vitro Is Blocked by Cyclopamine, an Inhibitor of Hedgehog Signaling. *Blood Cells Mol. Dis.* **2000**, *26*, 360–372. [CrossRef] [PubMed]
38. Huang, X.; Zhang, F.; He, D.; Ji, X.; Gao, J.; Liu, W.; Wang, Y.; Liu, Q.; Xin, T. Immune-Related Gene SERPINE1 Is a Novel Biomarker for Diffuse Lower-Grade Gliomas via Large-Scale Analysis. *Front. Oncol.* **2021**, *11*, 646060. [CrossRef]
39. Yahata, T.; Ibrahim, A.A.; Muguruma, Y.; Eren, M.; Shaffer, A.M.; Watanabe, N.; Kaneko, S.; Nakabayashi, T.; Dan, T.; Hirayama, N.; et al. TGF-β–Induced Intracellular PAI-1 Is Responsible for Retaining Hematopoietic Stem Cells in the Niche. *Blood* **2017**, *130*, 2283–2294. [CrossRef] [PubMed]
40. Ishimaru, K.; Yoshioka, K.; Kano, K.; Kurano, M.; Saigusa, D.; Aoki, J.; Yatomi, Y.; Takuwa, N.; Okamoto, Y.; Proia, R.L.; et al. Sphingosine Kinase-2 Prevents Macrophage Cholesterol Accumulation and Atherosclerosis by Stimulating Autophagic Lipid Degradation. *Sci. Rep.* **2019**, *9*, 18329. [CrossRef]
41. Galkina, S.I.; Fedorova, N.V.; Ksenofontov, A.L.; Stadnichuk, V.I.; Baratova, L.A.; Sud'Ina, G.F. Neutrophils as a Source of Branched-Chain, Aromatic and Positively Charged Free Amino Acids. *Cell Adhes. Migr.* **2019**, *13*, 98–105. [CrossRef]
42. Rijken, F.; Bruijnzeel, P.L.B. The Pathogenesis of Photoaging: The Role of Neutrophils and Neutrophil-Derived Enzymes. *J. Investig. Dermatol. Symp. Proc.* **2009**, *14*, 67–72. [CrossRef]
43. Semenov, A.N.; Shirshin, E.A.; Muravyov, A.V.; Priezzhev, A.V. The Effects of Different Signaling Pathways in Adenylyl Cyclase Stimulation on Red Blood Cells Deformability. *Front. Physiol.* **2019**, *10*, 923. [CrossRef]
44. Simonetti, S.; Natalini, A.; Folgori, A.; Capone, S.; Nicosia, A.; Santoni, A.; Di Rosa, F. Antigen-Specific CD8 T Cells in Cell Cycle Circulate in the Blood after Vaccination. *Scand. J. Immunol.* **2019**, *89*, e12735. [CrossRef] [PubMed]
45. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
46. 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.; et al. Array Programming with NumPy. *Nature* **2020**, *585*, 357–362. [CrossRef] [PubMed]
47. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261–272. [CrossRef]
48. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. *arXiv* **2015**, arXiv:1603.04467.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.