NeuroVelo: interpretable learning of cellular dynamics from single-cell transcriptomic data

Idris Kouadri Boudjelthia[1,2], Salvatore Milite[3], Nour El Kazwini[1], Javier Fernandez-Mateos[4], Nicola Valeri[4,5], Yuanhua Huang[6,7], Andrea Sottoriva[3,4], Guido Sanguinetti[1,8*]

[1*]Data Science, SISSA, Trieste, Italy.
[2]Abdus Salam International Centre for Theoretical Physics, Trieste, Italy.
[3]Computational Biology Research Centre, Human Technopole, Milan, Italy.
[4]Institute of Cancer Research, London, UK.
[5]Department of Surgery and Cancer, Imperial College, London, UK.
[6]School of Biomedical Sciences, University of Hong Kong, Hong Kong, China.
[7]Department of Statistics and Actuarial Science, University of Hong Kong, Hong Kong, China.
[8]School of Informatics, University of Edinburgh, Edinburgh, UK.

*Corresponding author(s). E-mail(s): gsanguin@sissa.it;
Contributing authors: ikouadri@sissa.it; salvatore.milite@fht.org; nelkazwi@sissa.it; fernandezmjavier.uo@uniovi.es; nicola.valeri@icr.ac.uk; yuanhua@hku.hk; andrea.sottoriva@fht.org;

Abstract

Reconstructing temporal cellular dynamics from static single-cell transcriptomics remains a major challenge. Methods based on RNA velocity, often in combination with non-linear dimensionality reduction, have been proposed. However, interpreting their results in the light of the underlying biology remains difficult, and their predictive power is limited. Here we propose NeuroVelo, a method that couples learning of an optimal linear projection with a non-linear low-dimensional dynamical system. Using dynamical systems theory, NeuroVelo can then identify genes and biological processes driving temporal cellular dynamics. We benchmark NeuroVelo against several current methods using single-cell multi-omic data, demonstrating that NeuroVelo is superior to competing methods in terms of identifying biological pathways and reconstructing evolutionary dynamics.

Keywords: single-cell transcriptomics, RNA velocity, neural networks, dynamical systems

Main text

Single-cell transcriptomic (scRNA-seq) technologies have transformed our understanding of cellular diversity and heterogeneity[1], yet their destructive nature poses fundamental limits to their ability to capture temporal biological dynamics. Inferring dynamic information from the static snapshots of scRNA-seq has been a major focus of computational research in the last decade. While most early efforts used advanced machine learning techniques to order cells along a pseudo-time trajectory [2–5], more recently the concept of RNA-velocity [6] offered a more mechanistic, biophysically grounded approach to solve the problem.

RNA-velocity leverages the (surprisingly abundant) pre-mRNA reads present in many scRNA-seq data sets, and uses a simple linear ordinary differential equation (ODE) model to deduce whether a gene is transcriptionally activated or repressed. This idea enables researchers to capture a shadow of cellular dynamics even in static data sets. Recent years have witnessed a flourishing of both applications and extensions of the RNA velocity framework [7–12]; nevertheless, both the biochemical foundations and the biological interpretation of RNA-velocity have been questioned [13, 14].

We propose NeuroVelo, a new approach which provides a more readily interpretable and highly effective way to infer cellular dynamics from scRNA-seq data. NeuroVelo combines ideas from Neural Ordinary Differential Equations (ODE) [15, 16] and RNA velocity in a physics-informed neural network architecture. Figure 1(a) illustrates the basic concept of NeuroVelo. scRNA-seq data consisting of both spliced and unspliced reads are encoded to a low dimensional space through two dimensionality reduction channels: one is a non-linear 1D encoder learning a pseudo-time coordinate associated with each cell, while the second is a linear projection to an *effective phase space* for the system. Cellular dynamics is defined through an autonomous system of differential equations parametrised by a neural network (a neural ODE). The resulting dynamics can be extremely rich; to further constrain it, we impose the RNA velocity principle as a further penalty in the loss function in the spirit of physics-informed neural networks. Notice that this constraint is applied locally to each cell, thus removing limiting assumptions on global transcriptional dynamics which are often a major problem for RNA velocity approaches. Because the effective phase space is defined via a linear projection, the RNA velocity constraint can be applied *exactly* in the reduced dimensional space. A second major benefit of the linearity of the phase space is that standard techniques for the analysis of low-dimensional non-linear dynamical systems can be applied to the trained model. The resulting insights, thanks to the linearity of the model, can readily be translated in terms of genes; we propose a novel rank-based

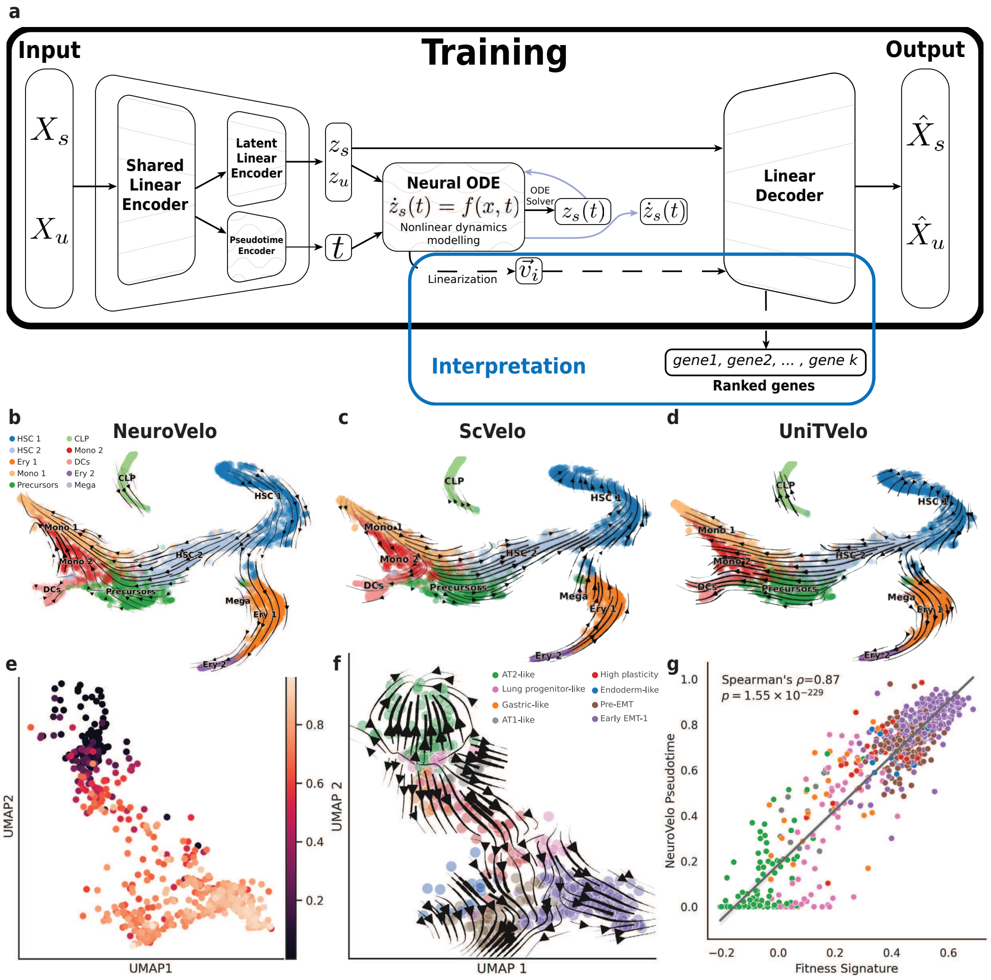

Fig. 1 (a) Schematic representation of the neuroVelo model: two autoencoders, a linear one to define the phase space, and a non-linear 1-D one to define pseudotime, are coupled in latent space through a non-linear neural ODE. RNA-velocity constraints are added as an additional penalty in the loss function. NeuroVelo (b), UniTVelo (c) and scVelo (d) results on the bone marrow data set. Inferred pseudotime (e) and velocity field (f) on the mouse cancer data set, as well as scatterplot of the neuroVelo pseudotime against the fitness signature (inferred from barcoding information) (f). The Spearman correlation coefficient and associated p-value are inset in panel (f).

statistic to provide a robust way to identify genes associated with dynamical changes in cellular state (see *Methods* section).

To test NeuroVelo, we utilize two recent scRNA-seq data sets: a human bone marrow hematopoiesis data set [5] and a CRISPR-based lineage barcoded mouse cancer data set [17], which was used to develop the specialised trajectory inference PhyloVelo.

Figure 1(b-d) shows a visualisation of the inferred velocity fields for NeuroVelo and two competing methods, UniTVelo[9] and scVelo [7], on the human bone barrow hematopoiesis dataset we considered[5]. While all methods provide a good separation of the cell types and treatments, it is worth noticing that only NeuroVelo and UniTVelo could capture the correct dynamics of early differentiation in the bone marrow data set. As a more stringent test, we show the results on the mouse cancer data set in the bottom row of Figure 1; this is modelled on Supplementary Figure 20 of the original PhyloVelo paper [17], which shows comparisons of PhyloVelo with five other methods [7–9, 11, 12], including scVelo and UniTVelo. Figure 1 (e-f) shows visualisations of the cells coloured by the inferred pseudo-time (e) and by cell-type together with the velocity fields (f). NeuroVelo broadly obtains an accurate reconstruction of both pseudo-time and velocity field. The velocity reconstruction in the early times is retrieved incorrectly (and in the same direction as most other RNA-velocity methods); this suggests that at the very early stages of this dynamical process lineage barcoding information (which is used by PhyloVelo) is indeed essential. Nevertheless, NeuroVelo achieves the highest Spearman correlation between inferred pseudo-time and fitness signature (Figure 1 (g), *cf* Supplementary Figure 20 in [17]), substantially higher than the other RNA-velocity methods and even marginally higher than PhyloVelo itself. A similar visualisation for another cancer data set from [17] is given in Suppl. Fig. 1, once again showing a very high consistency with the fitness signature and an excellent reconstruction of the velocity field.

To test NeuroVelo's ability to provide interpretable results, we turned to a single-cell Multiome dataset (nuclear transcriptome + chromatin accessibility - ATAC) of patient-derived colorectal cancer organoids treated with different drugs[18]. The organoid line was generated from a colorectal cancer clinical trial[19]. The parental organoid line was exposed to three different drug regimens: an AKT inhibitor (capivaserib), a MEK inhibitor (trametinib) and finally a sequence of first the AKT inhibitor followed by the MEK inhibitor. Because in this analysis we focus on interpretability, we restrict our comparisons to scVelo [7] and UniTVelo [9], which provide lists of *high velocity* genes. Other deep-learning based methods are less straightforward to interpret as they combine non-linear projections/ embeddings with non-linear dynamics. NeuroVelo, instead, relies only on a low-dimensional non-linear dynamical system, which can be analysed by standard spectral methods. The resulting eigenvectors can then be embedded in the original gene space using the linear embedding (see Methods), providing a list of genes that can be interpreted via standard methods.

Figure 2 (a) shows a visualisation of the neuroVelo latent space. The treatments are clearly separated, and the inferred velocity fields correctly identify an evolutionary process which diverges from a (subset of) the parental cells towards the resistant cells. Similar visualisations for scVelo and UniTVelo are shown in Supplementary Figure 2; notice however that neuroVelo obtains a clearer evolutionary trajectory than the competing methods. We then interrogated the gene lists obtained by the various methods. The genes identified to drive the cellular dynamics in resistant organoid lines by neuroVelo (Figure 2(b) and Supplementary Figure 3) are indeed significantly enriched for meaningful pathways in this case involved in carcinogenesis. By contrast, both scVelo and UniTVelo return very short lists of high velocity genes, which do not identify any significantly enriched pathway, and the top scoring pathways contain fewer cancer related pathways (Figure 2(b); indeed, UniTVelo's list of high velocity genes is too short to provide any meaningfully enriched pathways). As an example, we look in detail at the FoxO pathway, which is identified both by neuroVelo and scVelo. NeuroVelo identifies a large gene set with a strong enrichment score (Fig. 2(c) top), while the high velocity genes identified by scVelo in the FoxO pathway do not exhibit a particularly coordinated behaviour, obtaining an overall modest enrichment score (Fig. 2(c) bottom).

We then focussed on the 10 most important genes found by the different methods. Notice that the three methods identify almost entirely non-overlapping sets of genes (only one gene is shared between NeuroVelo and UniTVelo and between neuroVelo and scVelo). Figure 2 (d) shows that the most important genes identified by NeuroVelo have higher expression levels than those identified with scVelo or UniTVelo. In particular, scVelo identifies genes that have very low expression values, and hence of unclear biological significance. To seek more biological validation of these gene sets, we correlated the expression of the genes identified by NeuroVelo with the chromatin accessibility of their corresponding genomic regions. This was substantially higher than the other methods (Figure 2(e)). Figure 2 (f-g) shows a visual representation of expression (f) and open region probability ((g), see Methods for how we compute this quantity) for the BNIP3 gene, showing a clear correlation between the two measurements.

Taken together, these data provide strong evidence that NeuroVelo can capture biological dynamics from complex single-cell transcriptomic data, in a manner that is both accurate and interpretable in terms of underlying genes driving the dynamics. We believe NeuroVelo will therefore become an important tool in characterising biological dynamics in single-cell experiments.

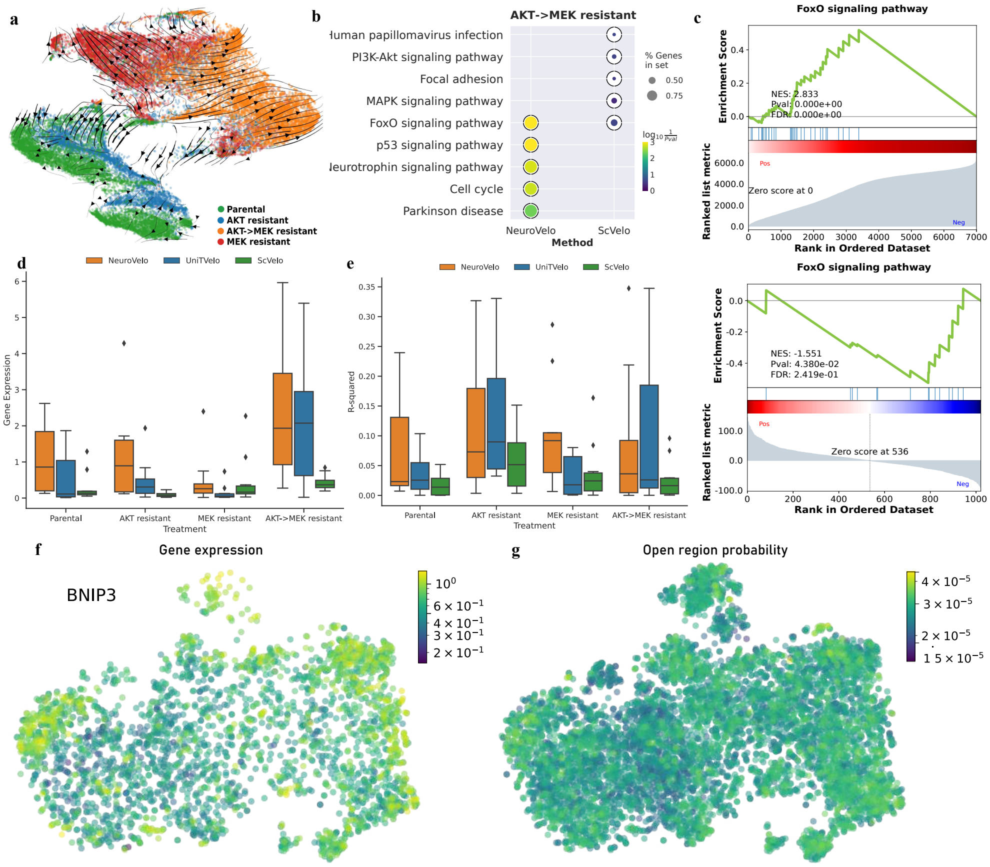

Fig. 2 (a) UMAP visualisation of the latent space inferred by neuroVelo, colored by cell type and with overlaid velocity field. (b) Gene set enrichment analysis of neuroVelo and scVelo genes. (c) NeuroVelo and scVelo respectively gene set enrichment analysis of common pathways of the two methods. (d) Expression of top 10 genes from the three methods in different treatments. (e) Smoothed R-squared between gene expression and open probabilities for associated promoters (cisTopic) for the top 10 genes from the three methods in different treatments. TSNE of expression of the BNIP3 gene (f) and open probability (g) for the associated promoters

References

[1] Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., *et al.*: The human cell atlas. elife **6**, 27041 (2017)

[2] Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., Rinn, J.L.: The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature biotechnology **32**(4), 381–386 (2014)

[3] Haghverdi, L., Büttner, M., Wolf, F.A., Buettner, F., Theis, F.J.: Diffusion pseudotime robustly reconstructs lineage branching. Nature methods **13**(10), 845–848 (2016)

[4] Campbell, K.R., Yau, C.: Uncovering pseudotemporal trajectories with covariates from single cell and bulk expression data. Nature communications **9**(1), 2442 (2018)

[5] Setty, M., Kiseliovas, V., Levine, J., Gayoso, A., Mazutis, L., Pe'Er, D.: Characterization of cell fate probabilities in single-cell data with palantir. Nature biotechnology **37**(4), 451–460 (2019)

[6] La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., *et al.*: Rna velocity of single cells. Nature **560**(7719), 494–498 (2018)

[7] Bergen, V., Lange, M., Peidli, S., Wolf, F.A., Theis, F.J.: Generalizing rna velocity to transient cell states through dynamical modeling. Nature biotechnology **38**(12), 1408–1414 (2020)

[8] Chen, Z., King, W.C., Hwang, A., Gerstein, M., Zhang, J.: Deepvelo: Single-cell transcriptomic deep velocity field learning with neural ordinary differential equations. Science Advances **8**(48), 3745 (2022)

[9] Gao, M., Qiao, C., Huang, Y.: Unitvelo: temporally unified rna velocity reinforces single-cell trajectory inference. Nature Communications **13**(1), 6586 (2022)

[10] Gayoso, A., Weiler, P., Lotfollahi, M., Klein, D., Hong, J., Streets, A., Theis, F.J., Yosef, N.: Deep generative modeling of transcriptional dynamics for rna velocity analysis in single cells. Nature Methods, 1–10 (2023)

[11] Li, S., Zhang, P., Chen, W., Ye, L., Brannan, K.W., Le, N.-T., Abe, J.-i., Cooke, J.P., Wang, G.: A relay velocity model infers cell-dependent rna velocity. Nature biotechnology, 1–10 (2023)

[12] Qiao, C., Huang, Y.: Representation learning of rna velocity reveals robust cell transitions. Proceedings of the National Academy of Sciences **118**(49), 2105859118 (2021)

[13] Barile, M., Imaz-Rosshandler, I., Inzani, I., Ghazanfar, S., Nichols, J., Marioni, J.C., Guibentif, C., Göttgens, B.: Coordinated changes in gene expression kinetics underlie both mouse and human erythroid maturation. Genome biology **22**(1), 1–22 (2021)

[14] Gorin, G., Fang, M., Chari, T., Pachter, L.: Rna velocity unraveled. PLOS Computational Biology **18**(9), 1010492 (2022)

[15] Chen, R.T., Rubanova, Y., Bettencourt, J., Duvenaud, D.K.: Neural ordinary differential equations. Advances in neural information processing systems **31** (2018)

[16] Li, Q.: sctour: a deep learning architecture for robust inference and accurate prediction of cellular dynamics. Genome Biology **24**(1), 1–33 (2023)

[17] Wang, K., Hou, L., Wang, X., Zhai, X., Lu, Z., Zi, Z., Zhai, W., He, X., Curtis, C., Zhou, D., et al.: Phylovelo enhances transcriptomic velocity field mapping using monotonically expressed genes. Nature Biotechnology, 1–12 (2023)

[18] Fernandez-Mateos, J., Milite, S., Oliveira, E., Vlachogiannis, G., Chen, B., Yara, E., Cresswell, G.D., James, C., Patruno, L., Ascolani, G., Acar, A., Heide, T., Spiteri, I., Graudenzi, A., Caravagna, G., Graham, T., Magnani, L., Valeri, N., Sottoriva, A.: Epigenetic heritability of cell plasticity drives cancer drug resistance through one-to-many genotype to phenotype mapping (2023)

[19] Vlachogiannis, G., Hedayat, S., Vatsiou, A., Jamin, Y., Fernández-Mateos, J., Khan, K., Lampis, A., Eason, K., Huntingford, I., Burke, R., Rata, M., Koh, D.-M., Tunariu, N., Collins, D., Hulkki-Wilson, S., Ragulan, C., Spiteri, I., Moorcraft, S.Y., Chau, I., Rao, S., Watkins, D., Fotiadis, N., Bali, M., Darvish-Damavandi, M., Lote, H., Eltahir, Z., Smyth, E.C., Begum, R., Clarke, P.A., Hahne, J.C., Dowsett, M., Bono, J., Workman, P., Sadanandam, A., Fassan, M., Sansom, O.J., Eccles, S., Starling, N., Braconi, C., Sottoriva, A., Robinson, S.P., Cunningham, D., Valeri, N.: Patient-derived organoids model treatment response of metastatic gastrointestinal cancers. Science **359**(6378), 920–926 (2018) https://doi.org/10.1126/science.aao2774 https://www.science.org/doi/pdf/10.1126/science.aao2774

[20] Satpathy, A.T., Granja, J.M., Yost, K.E., Qi, Y., Meschi, F., McDermott, G.P., Olsen, B.N., Mumbach, M.R., Pierce, S.E., Corces, M.R., *et al.*: Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral t cell exhaustion. Nature biotechnology **37**(8), 925–936 (2019)

[21] Bravo González-Blas, C., Minnoye, L., Papasokrati, D., Aibar, S., Hulselmans, G., Christiaens, V., Davie, K., Wouters, J., Aerts, S.: cistopic: cis-regulatory topic modeling on single-cell atac-seq data. Nature methods **16**(5), 397–400 (2019)

[22] El Kazwini, N., Sanguinetti, G.: Share-topic: Bayesian inerpretable modelling of single-cell multi-omic data. bioRxiv, 2023–02 (2023)

[23] Peltzer, A., De Almeida, F.M., Botvinnik, O., Sturm, G., Menden, K., Nf-Core Bot, Sangram K Sahu, Gabernet, G., Alvarez, P., Syme, R., Harshil Patel, Kelly, T., Heylf, Talbot, A., PETER BAILEY, Garcia, M.U., Thiery, A., Reynolds, R.H., Azedine Zoufir, Medova, H., Khajidu, Ribeiro-Dantas, M., Chen, W.-A., Leipzig, J., Ewels, P., Werbrouck, S., Ameynert, Marteau, V.: nf-core/scrnaseq: nf-core/scrnaseq v2.3.2 "Sepia Samarium Salmon". Zenodo (2023). https://doi.

org/10.5281/ZENODO.8015242 . https://zenodo.org/record/8015242

Methods

Model architecture

NeuroVelo is a neural network architecture that consists of a linear auto-encoder initialized with PCA and neural ODEs architecture in the latent representation of the data. The main idea of NeuroVelo is to construct drug-specific cell dynamics from scRNA-seq data in the latent space.

We used spliced and unspliced reads of scRNA-seq data and embedded them using the same encoder function into a latent space $\boldsymbol{z} = (\boldsymbol{z}_s, \boldsymbol{z}_u)$ of dimension l. The encoder is initialized using the principal components of the spliced reads. This step is mainly for dimensionality reduction and feature extraction of the highly dimensional gene space to a much smaller latent space (50 by default).

Part of the encoder is also used to estimate the pseudotime of the cells. Both pseudotime and latent representation of the cell are used to train nonlinear neural ODEs.

Once we project the cells into a lower-dimensional space, we use a set of neural ODEs to describe the cellular dynamics in this latent space

$$\frac{d\boldsymbol{z}_s^{(d)}(t)}{dt} = f^{(d)}(\boldsymbol{z}_s^{(d)}(t), t), \tag{1}$$

where $f^{(d)} : \mathbb{R}^l \to \mathbb{R}^l$ is a two layers neural network with nonlinear activation function, this neural ODE describes the dynamics of cells treated with a particular drug d. Using the nonlinear ODEs in the latent space helps to capture the essential and complex dynamics of drugs in this space.

Notice that, since we are learning both the pseudotime and the generator function of the dynamics (the f function), one could reverse simultaneously the pseudotime and the sign of f without altering the trajectories. This ambiguity may be resolved either by prior knowledge or by inverting both signs *a posteriori*.

The encoder latent representation is decoded with a linear decoder network (initialized with the same principal components as well) to create a reconstruction of the original spliced and unspliced inputs ($\hat{X}_s$, $\hat{X}_u$), learning another function that maps from latent space to gene space.

Model training

The tasks learnt by neural networks are based on what loss function the network is optimizing. To ensure that the cellular dynamics learned by the network have biological meaning, we propose a loss function that puts biophysical constraints on the data and does not require any parameter fine-tuning.

The first part of the loss is the usual mean squared error (MSE) of autoencoders network, we want our spliced and unspliced inputs to match their reconstructed counterparts from the decoder and so the first loss part is a sum of MSE of spliced and unspliced reads.

The second part of the loss is what puts a biophysical constraint on the neural ODEs in the same fashion of physics-informed neural networks. The idea is to make the derivative of latent spliced reads $\dot{\boldsymbol{z}}_s^d(t)$ align with the splicing dynamics from the RNA velocity but in latent representation $(e^{\beta}\boldsymbol{z}_u - e^{\gamma}\boldsymbol{z}_s)$, There is an exponential of β (and γ) because they are analogy of splicing and degradation rate and they must be positive.

The final loss is written as follows:

$$L = \mathrm{MSE}(X_s, \hat{X}_s) + \mathrm{MSE}(X_u, \hat{X}_u) + \sum_{d=1}^{D} \mathrm{MSE}\left(f^{(d)}(\boldsymbol{z}_s^{(d)}(t), t), e^{\beta}\boldsymbol{z}_u^{(d)} - e^{\gamma}\boldsymbol{z}_s^{(d)}\right) \quad (2)$$

Where the sum over d is the sum over specific drugs/treatments. $\dot{z}(t)_s^d$ is the function that describes the splicing dynamics of drug d in the latent space. z_s^d and z_u^d are the projected spliced and unspliced reads of treatment d. Notice that the final term in the loss enforces the RNA-velocity constraint, so that the role of the transcription rate is modelled flexibly and learnt directly from the data.

In this way, we are able to investigate complex and specific cellular dynamics by learning splicing dynamics constrained nonlinear ODEs in a reduced space and without the need to fine tune parameters on the level of the loss function as well.

Interpretation

Any dynamical process is generally governed by a nonlinear ODE, we can understand the behavior of a system by linearizing the ODE around a given point which is in this case an average cell or group of cells, the linearized system is:

$$\dot{\boldsymbol{z}}_s(t) = \mathbf{J}\boldsymbol{z}_s(t), \quad \mathbf{J} = \begin{bmatrix} \frac{\partial f_1}{\partial z_1} & \cdots & \frac{\partial f_1}{\partial z_l} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_l}{\partial z_1} & \cdots & \frac{\partial f_l}{\partial z_l} \end{bmatrix} \quad (3)$$

We perform *eigenanalysis* on the Jacobin matrix. The eigenvalues give the important directions of the correspondent eigenvectors in the latent space, then we use the decoder function to map these important directions in the latent space to the gene space. The output of this decoder is a list of ranked genes based on their importance in a particular direction.

The main issues related to using neural networks are interpretability and robustness, the former is not a problem because NeuroVelo uses both linear embedding and projection, thus the interpretation of the latent space into the gene space is straightforward. The robustness part is more of a concern because a different initialization leads to a different local minima which makes biological interpretation uncertain. Here we propose a geometrical and statistical approach that gives a solid gene list by using many trained models.

The first step is to find the set of eigenvectors from other models that align the most with an eigenvector of a current model $\vec{v}_i^{(m)}$, the alignment is based on the absolute value of the cosine similarity. This set of the most aligned eigenvectors is given by the following equation:

$$U\left(\vec{v}_i^{(m)}\right) = \left\{\max_j \left| S_c\left(\vec{v}_i^{(m)}, \vec{v}_j^{(m')}\right)\right|, \text{for } m' \neq m\right\}$$

where S_c is the cosine similarity function and m' is the index of the trained model.

We take the ranked genes of the aligned eigenvectors, and we average the ranks gene-wise. The output is a list of genes ranked based on the average rank across the aligned eigenvectors from different trained models. This indicates that the genes that are at the top of the list hold significant importance for the dynamics learned by various models and vice versa. A set of different analyses can be performed on these average-ranked genes for validation.

0.1 Smoothed R-squared

The smoothed coefficient of determination R^2 is computed between the gene expression and the probability of open chromatin in the associated region. It is smoothed R^2 because we take the average gene expression around the cell neighborhood (20 neighbors)

Implementation

NeuroVelo is implemented in Python machine learning framework `PyTorch 2.0` and using `torchdiffeq` package for ODE solvers in PyTorch.

Velocity field visualizations are done with `ScVelo`

Gene set enrichment analysis is done with *"GSEApy: Gene Set Enrichment Analysis in Python"* `gseapy`. Passing a dataframe of genes and their average rank as a ranking metric to `prerank` function. The enrichment analysis plots are all generated by the same package.

Datasets

Patient-derived colorectal cancer organoids

10X multiomics sequencing data has been processed with cellranger-arc count version 2.0.2 [20]. In order to get spliced and unspliced read quantification we run velocyto[6] on the in mode run10x on the output of cellranger.

After aligning and generating the spliced and unspliced count of this dataset. We picked the top 7000 highly variables gene (as the number of cells was greater than $30,000$) we used normalized counts for both training and validation.

Chromatin open probabilities for pre-selected genomic regions were obtained by using the cisTopic latent Dirichlet allocation model [21], using the implementation in [22].

Mouse lung cancer

This dataset is CRISPR/Cas9-based lineage tracing dataset from a genetically-engineered mouse model (GEMM) of lung adenocarcinoma. We used it to compare

NeuroVelo to PhyloVelo, mainly using two tumors (3726_NT_T1 and 3435_NT_T1). The published processed count contained only spliced reads, thus we downloaded the fastq files and used `nf-core/scrnaseq` pipeline [23] to align (using STAR) and generate RNA velocity count matrix. The number of cells used in PhyloVelo was in order of 700 thus we only picked the top 1000 highly variable genes to run NeuroVelo. Normalized spliced and unspliced counts were used for training and post-training pseudotime inference.

Human bone marrow hematopoiesis

The human bone marrow dataset is downloaded from ScVelo package. The bone marrow is the primary site of new blood cell production or haematopoiesis. It is composed of hematopoietic cells, marrow adipose tissue, and supportive stromal cells. The dataset is already processed and contains spliced and unspliced counts. We picked the top 2000 highly variables gene and we used moments instead of counts for training.