Machine learning methods for single-cell transcriptomic data

Guido Sanguinetti

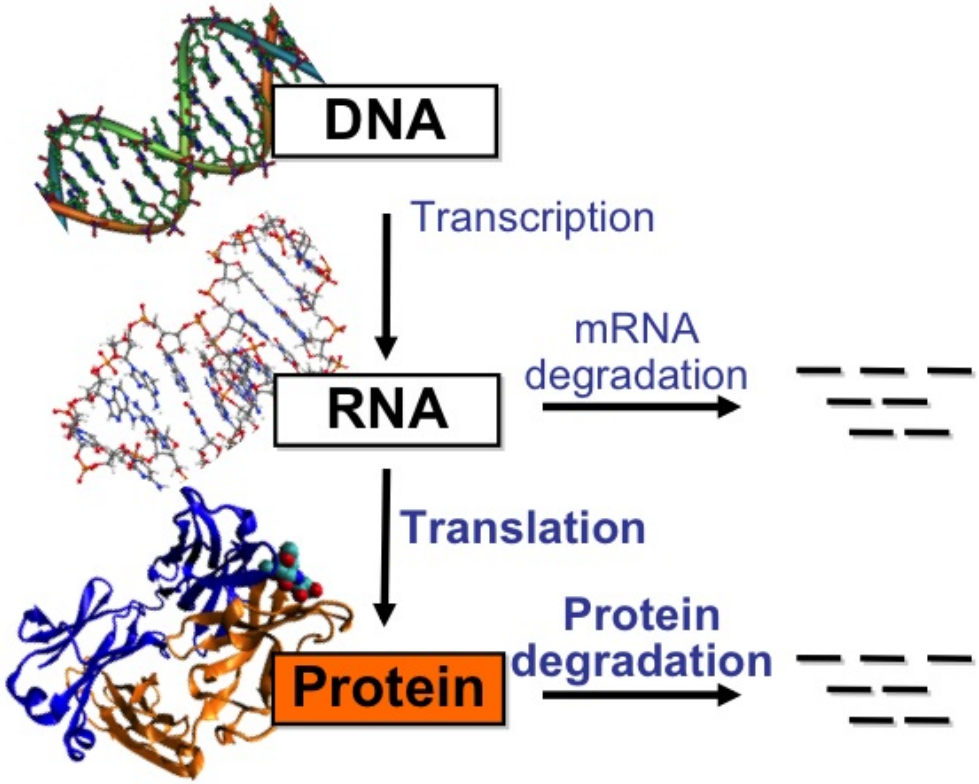
Data Science @ SISSA, Trieste

Bologna, 8 July 2024

Roadmap of today

Roadmap of today

The central dogma

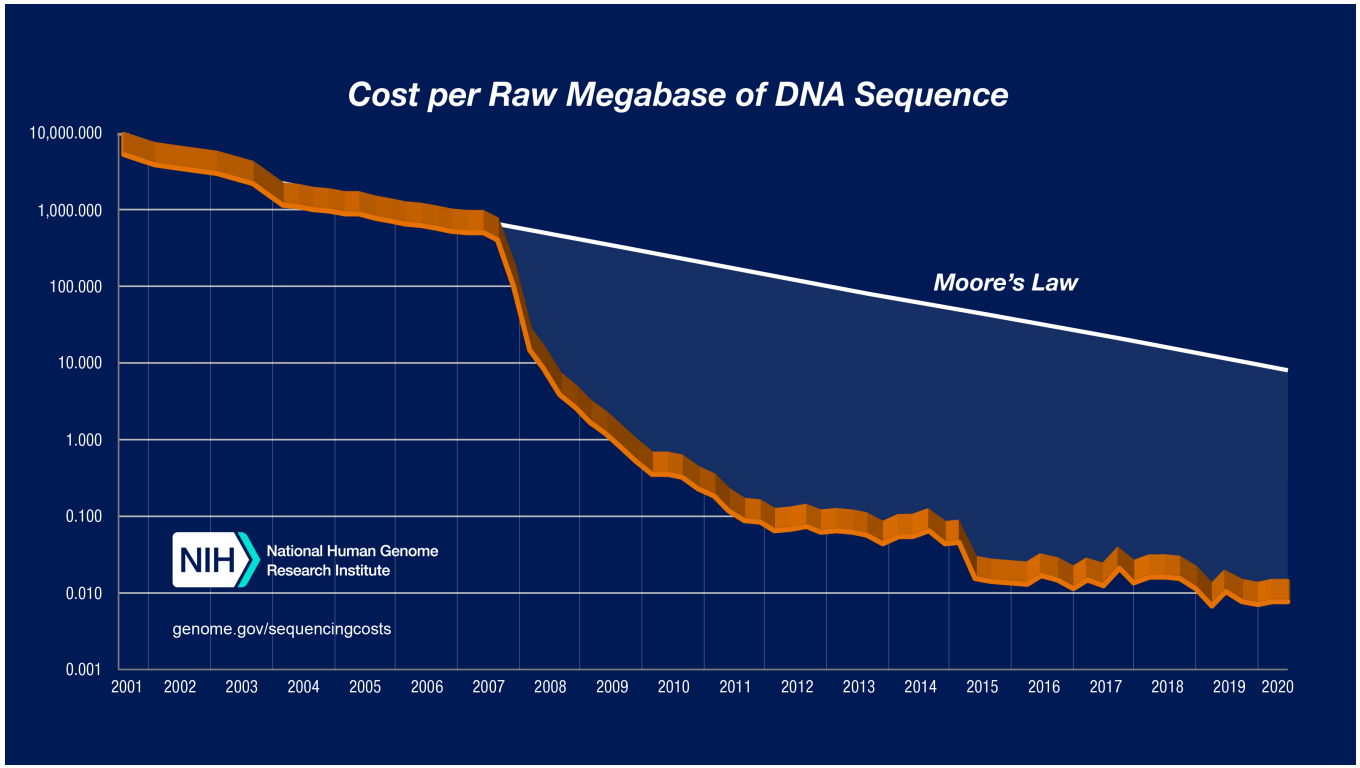


Where does variability come to play?

What is sequencing?

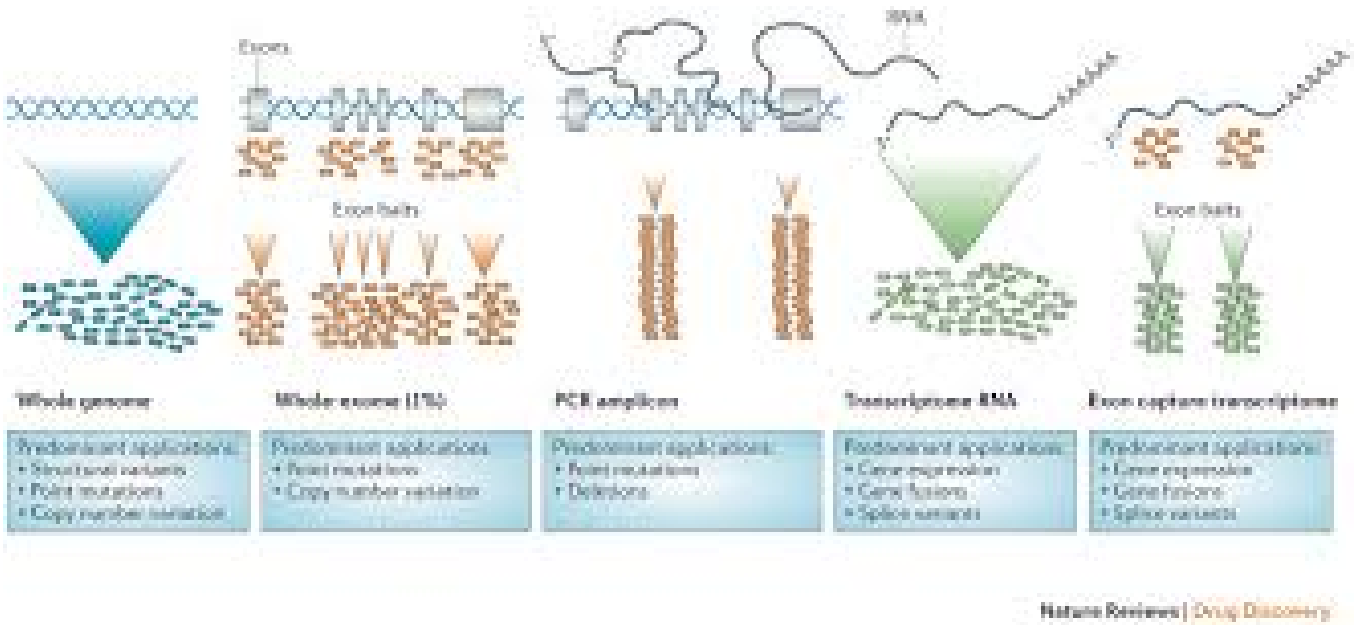
- Sequencing: (chemical or otherwise) procedure for reading the sequence of constituents of a stretch of DNA (or a protein)
- IMPORTANT 1: all DNA sequencing happens in short chunks (from a few tens to a few tens of thousands bp)
- IMPORTANT 2: mistakes are possible (depending on the technology) but quality controls available (Phred scores)
- IMPORTANT 3: some sequences are easier to sequence (biases in the data)

More than Moore



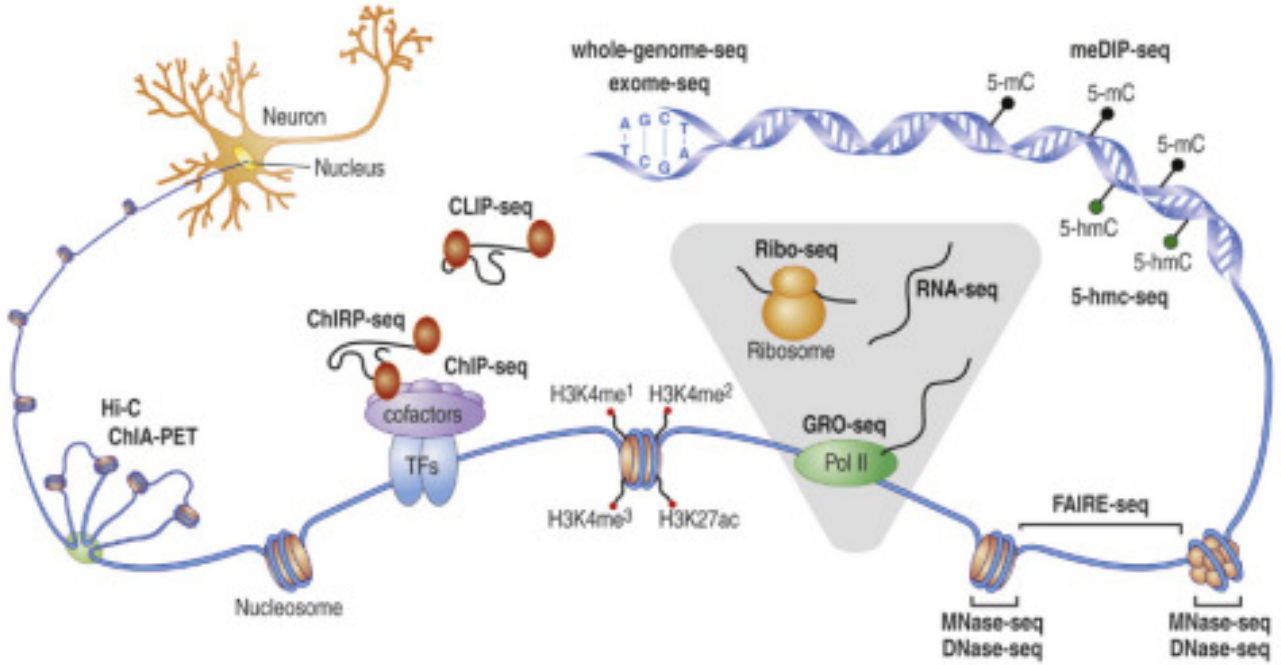
Source: National Human Genome Research Institute

Now sequencing is easier?



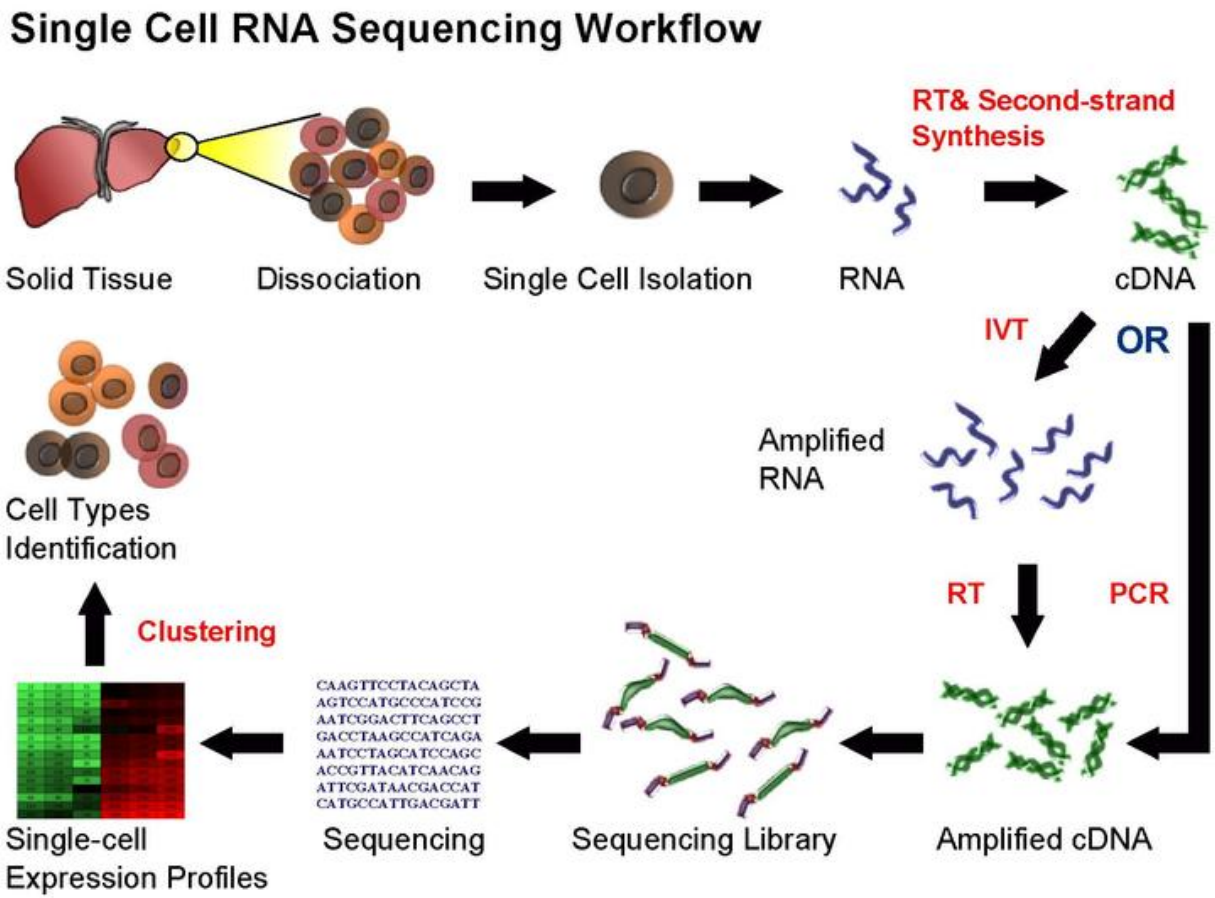
Major technical advances + massive parallelisation, technology started coming online around 2008. Throughput (most recent versions) in the region 10^5Mb/hr.

Sequencing everywhere



- NGS can be "prefaced" by any biochemical treatment
- IMPORTANT: when doing that biases are often introduced/ becomes unclear how to compare samples

scRNA-seq



Can do 100K cells in single experiment. High dropout rate, huge variability in coverage. Dominant technology now.

Roadmap of today

What single-cell 'omics look like

- For each cell, we normally obtain $\sim$ 10K RNA fragments mapped to the transcriptome $\rightarrow$ *most genes* are missed in every single cell
- We apply some pre-filtering criterion, e.g. discard genes not measured in at least 50% of cells, cells with fewer than 100 non-zero genes
- We end up with a gene expression matrix typically $\sim$ 6K rows (genes) and a few thousands columns (cells)

What single-cell 'omics look like

- For each cell, we normally obtain $\sim$ 10K RNA fragments mapped to the transcriptome $\rightarrow$ *most genes* are missed in every single cell
- We apply some pre-filtering criterion, e.g. discard genes not measured in at least 50% of cells, cells with fewer than 100 non-zero genes
- We end up with a gene expression matrix typically $\sim$ 6K rows (genes) and a few thousands columns (cells)
- A large fraction of the entries are zero, either genuine or dropout

The general problem of Data Science

- We have noisy, high dimensional data
- Several factors contribute to the variance we observe in the data: experimental noise, intrinsic stochasticity, physiological processes, disease processes

The general problem of Data Science

- We have noisy, high dimensional data
- Several factors contribute to the variance we observe in the data: experimental noise, intrinsic stochasticity, physiological processes, disease processes
- **ALL** of Data Science/ AI consists in partitioning this variance and using this (implicit or explicit) decomposition for predictions

The general problem of Data Science

- We have noisy, high dimensional data
- Several factors contribute to the variance we observe in the data: experimental noise, intrinsic stochasticity, physiological processes, disease processes
- **ALL** of Data Science/ AI consists in partitioning this variance and using this (implicit or explicit) decomposition for predictions
- The difference lies in the assumptions about what is an important direction of variation

Problem 1: Visualisation/ dimensionality reduction

- Assumption: Only few degrees of freedom exist in the data

Problem 1: Visualisation/ dimensionality reduction

- Assumption: Only few degrees of freedom exist in the data
- Dominant tool: UMAP (Uniform Manifold Approximation (McInnes et al 2018)
- Creates a nearest neighbour graph in gene space, then tries to find points in low D such that the graph distances are preserved
- Usually prefaced by a linear dimensionality reduction step (PCA)
- Not easy to understand what UMAP directions mean

Sub-problem 2: Pseudo-time

- Assumption: the major direction of variation is along a developmental direction
- E.g., cells are collectively following a dynamical process (development, differentiation, drug response) but individually they are at slightly different stages of the process

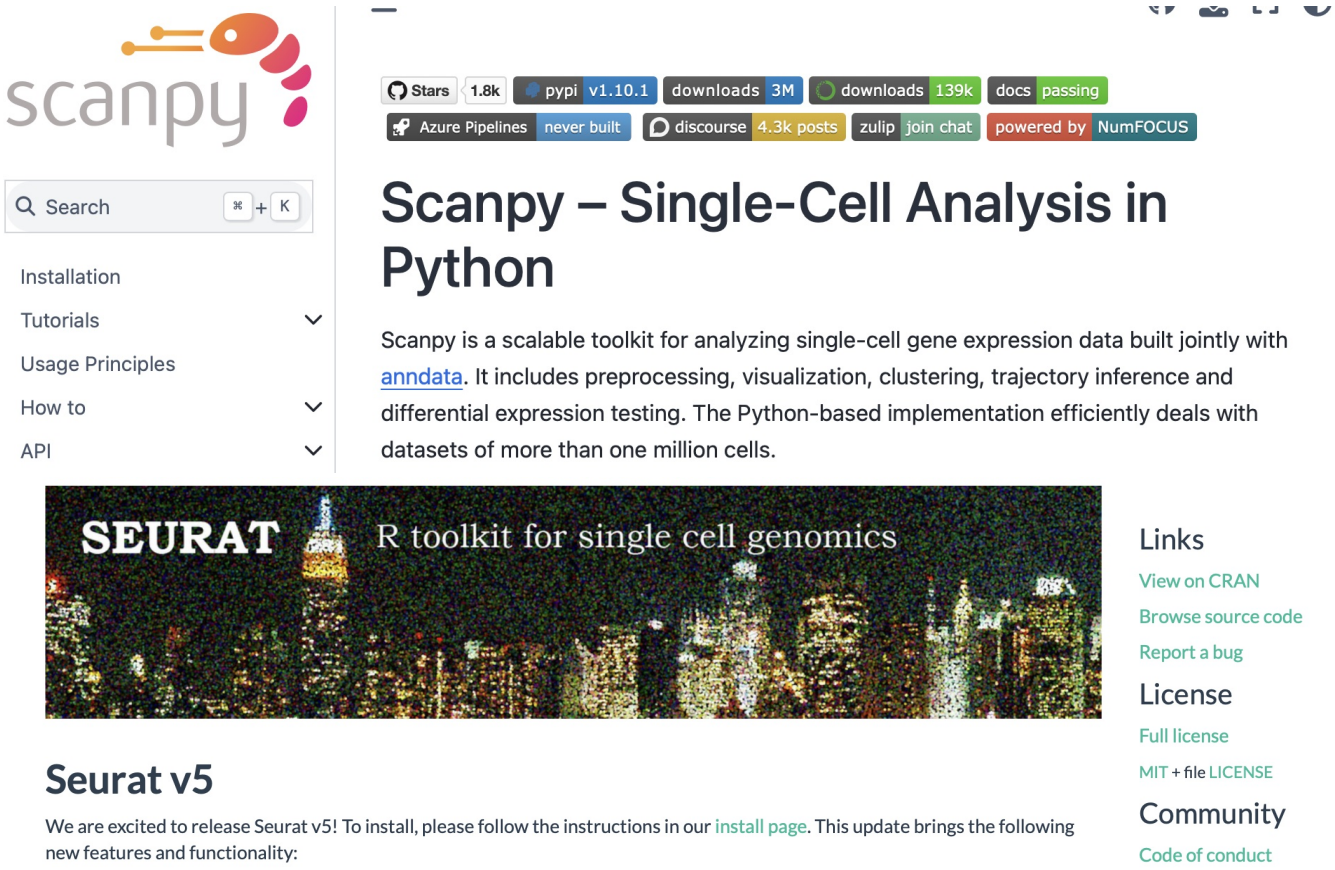
Sub-problem 2: Pseudo-time

- Assumption: the major direction of variation is along a developmental direction
- E.g., cells are collectively following a dynamical process (development, differentiation, drug response) but individually they are at slightly different stages of the process
- Dimensionality reduction to 1D, either VAE or graph-based (e.g. Diffusion Maps Haghverdi 2016)
- Output: (partial) ordering of cells, identification of branching events

Problem 3: clustering

- Assumption: the major variation is caused by the existence of distinct groups of cells which are transcriptomically homogeneous
- Solution: clustering algorithms of various flavours
- Often prefaced by dimensionality reduction
- Generally multiple algorithms to reduce noise (e.g. SC3, Kiselev et al 2017)

Software packages



Both R and Python well used

Roadmap of today

Uncovering dynamics: RNA velocity

- scRNA-seq is destructive $\rightarrow$ static snapshots from a dynamic process

Uncovering dynamics: RNA velocity

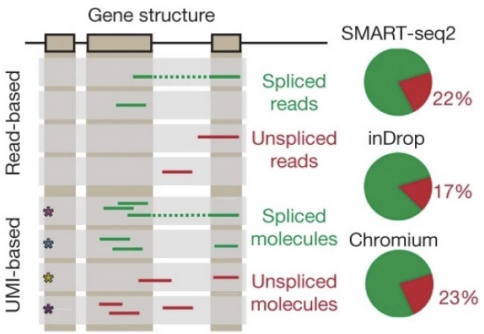
- scRNA-seq is destructive $\rightarrow$ static snapshots from a dynamic process
- **IDEA** (La Manno et al, 2018): use spliced/ unspliced reads to derive *rate of change* of RNA levels

$$\frac{dx_u}{dt} = \alpha - \beta x_u \qquad \frac{dx_s}{dt} = \beta x_u - \gamma x_s$$

Uncovering dynamics: RNA velocity

- scRNA-seq is destructive $\rightarrow$ static snapshots from a dynamic process
- **IDEA** (La Manno et al, 2018): use spliced/ unspliced reads to derive *rate of change* of RNA levels

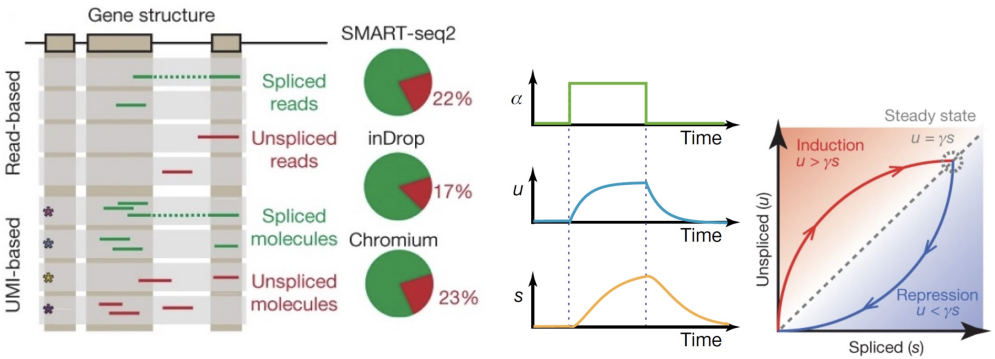
$$\frac{dx_u}{dt} = \alpha - \beta x_u \qquad \frac{dx_s}{dt} = \beta x_u - \gamma x_s$$



Uncovering dynamics: RNA velocity

- scRNA-seq is destructive $\rightarrow$ static snapshots from a dynamic process
- **IDEA** (La Manno et al, 2018): use spliced/ unspliced reads to derive *rate of change* of RNA levels

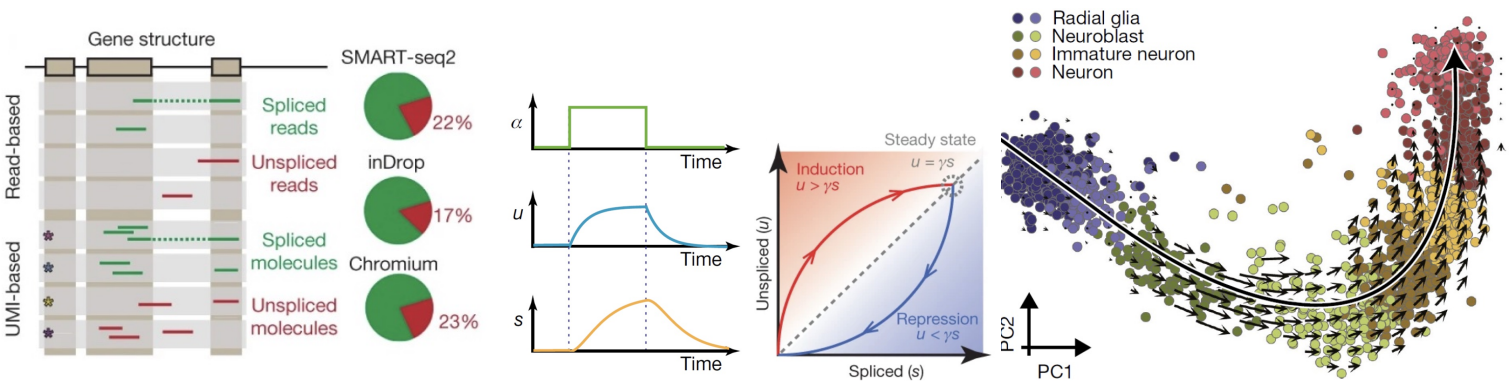
$$\frac{dx_u}{dt} = \alpha - \beta x_u \qquad \frac{dx_s}{dt} = \beta x_u - \gamma x_s$$



Uncovering dynamics: RNA velocity

- scRNA-seq is destructive $\rightarrow$ static snapshots from a dynamic process
- **IDEA** (La Manno et al, 2018): use spliced/ unspliced reads to derive *rate of change* of RNA levels

$$\frac{dx_u}{dt} = \alpha - \beta x_u \qquad \frac{dx_s}{dt} = \beta x_u - \gamma x_s$$



Problems and solutions

- Splicing signal is **very noisy** in single cells
- No reason why timescale of splicing should be the relevant one

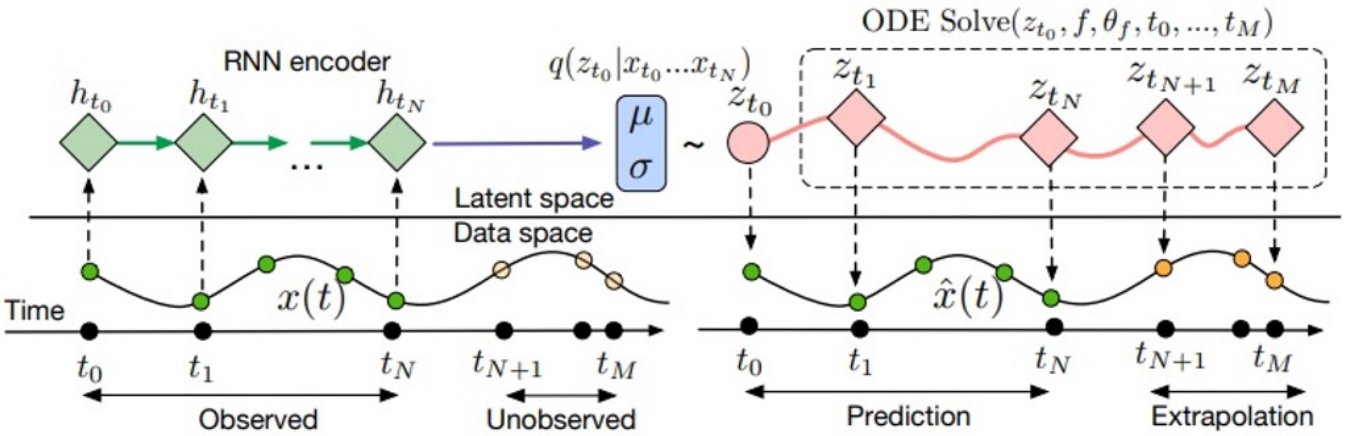
Problems and solutions

- Splicing signal is **very noisy** in single cells
- No reason why timescale of splicing should be the relevant one
- **IDEA**: Underlying (low dimensional) nonlinear dynamical system should govern long-term evolution of cells' transcriptomes
- Spliced/ unspliced ratio gives a noisy measurement of *instantaneous* rate of change

Problems and solutions

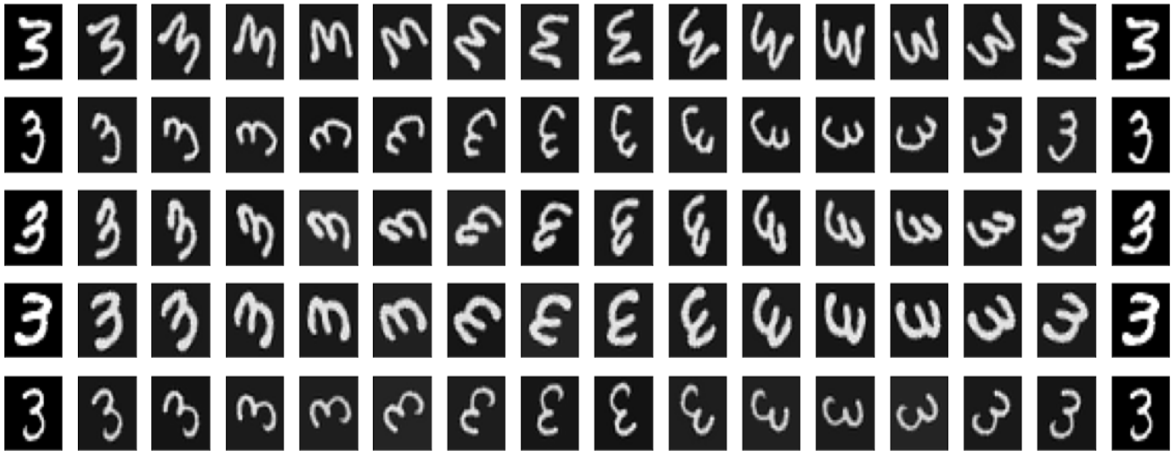
- Splicing signal is **very noisy** in single cells
- No reason why timescale of splicing should be the relevant one
- **IDEA**: Underlying (low dimensional) nonlinear dynamical system should govern long-term evolution of cells' transcriptomes
- Spliced/ unspliced ratio gives a noisy measurement of *instantaneous* rate of change
- Couple the two components in the spirit of *physics informed machine learning*

Neural ODEs (Chen et al 2018)

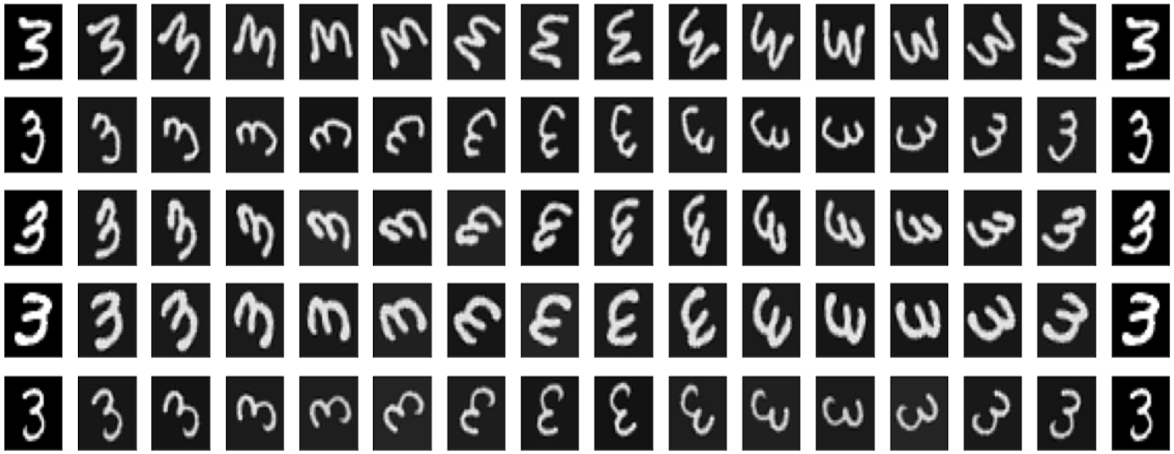


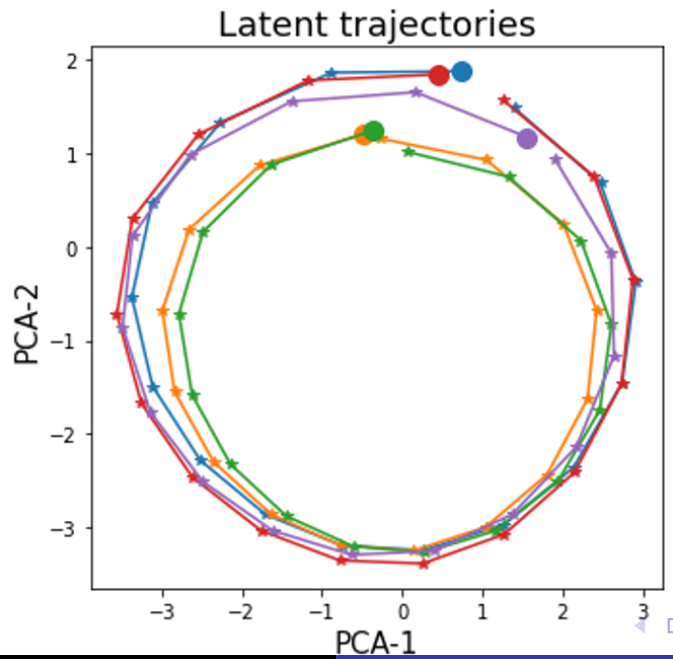
Autoencoding structure in time. ODE in latent space with drift parametrised by a NN. Efficient evaluation of gradients by Pontryagin adjoint.

A (somewhat contrived) example

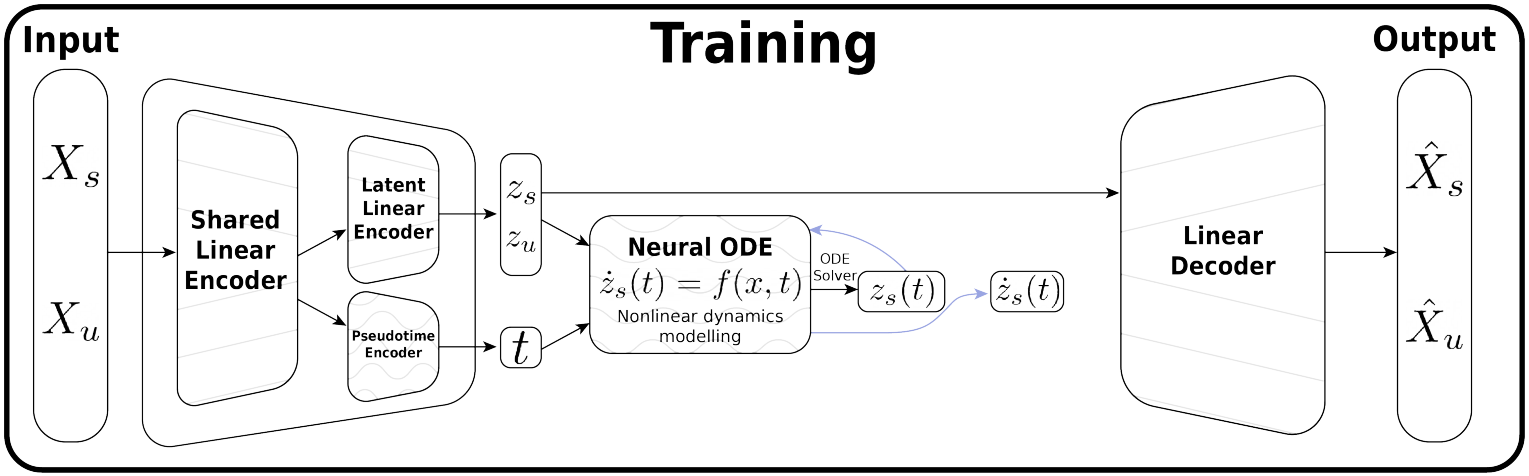


A (somewhat contrived) example





NeuroVelo (Idris Kouadri Boudjelthia)



$$\mathcal{L} = \mathrm{MSE}(X, \hat{X}) + \mathrm{MSE}(\dot{z}_s, \beta z_u - \gamma z_s)$$

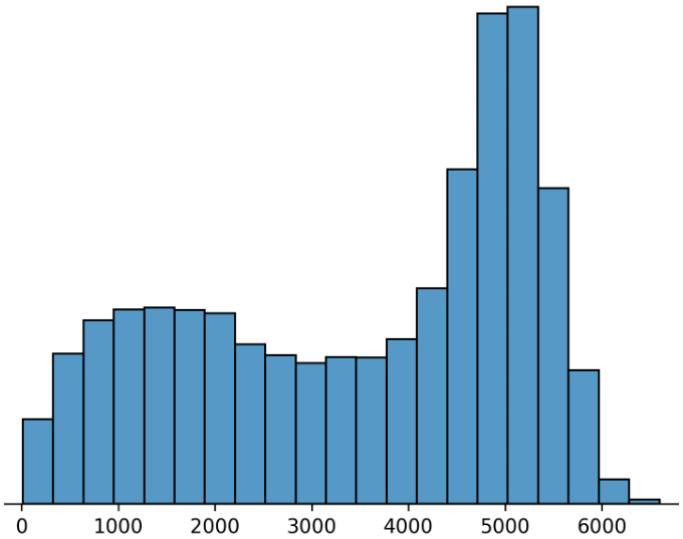
Because the encoding/ decoding is linear, the RNA velocity equations apply also in latent space. Notice we need no assumptions on the transcription rate function.

Interpreting Neurovelo

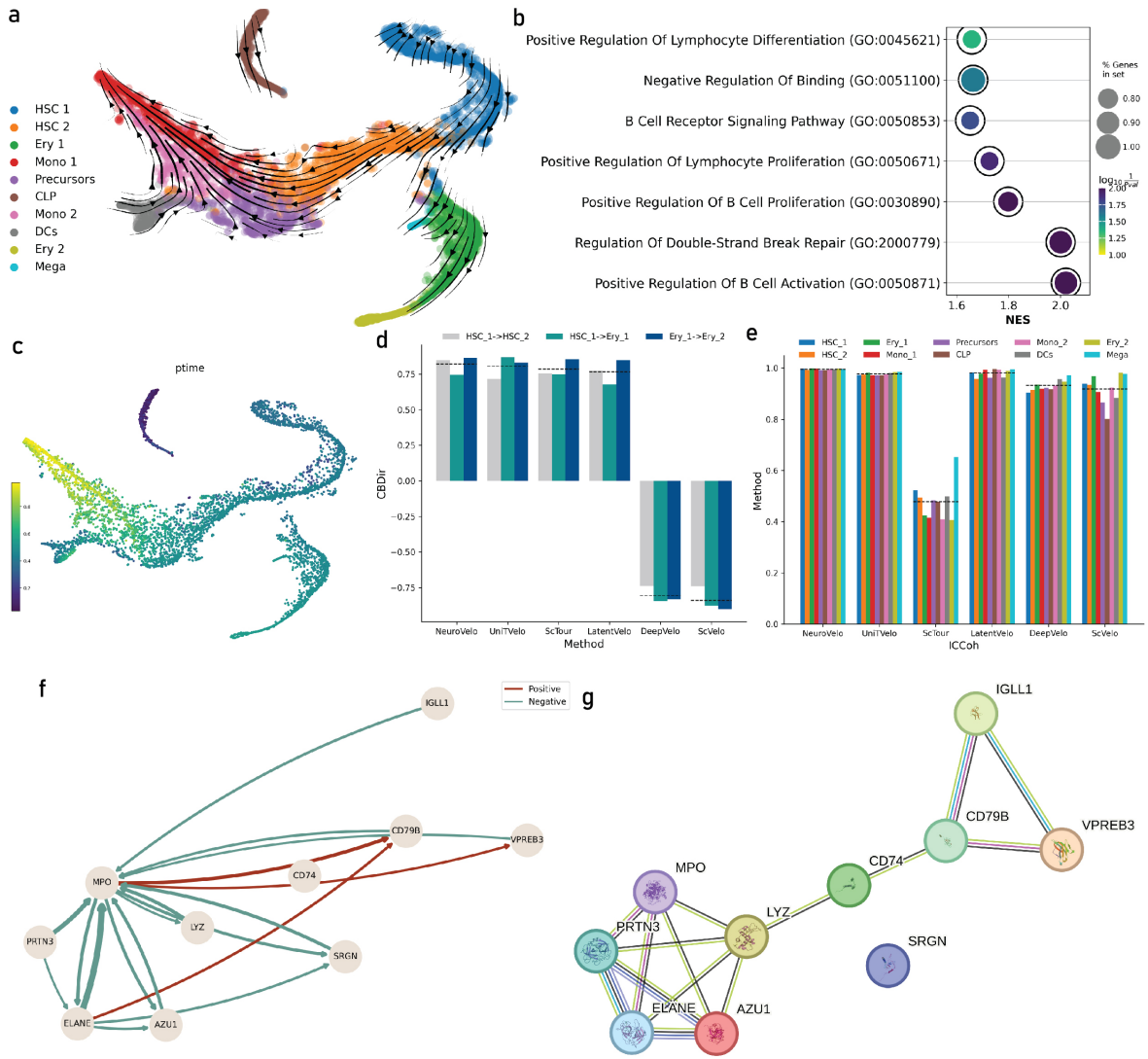
- NeuroVelo learns a low-dimensional nonlinear dynamical system
- Principal dynamics are given (locally) by the *eigenvectors* of the Jacobian matrix
- These eigenvectors can be decoded linearly to give a ranked list of genes
- The decoded Jacobian matrix gives itself a description of the network of interactions between genes
- Robustness is ensured by computing a stability index w.r.t. multiple initializations

Interpreting Neurovelo cont'd

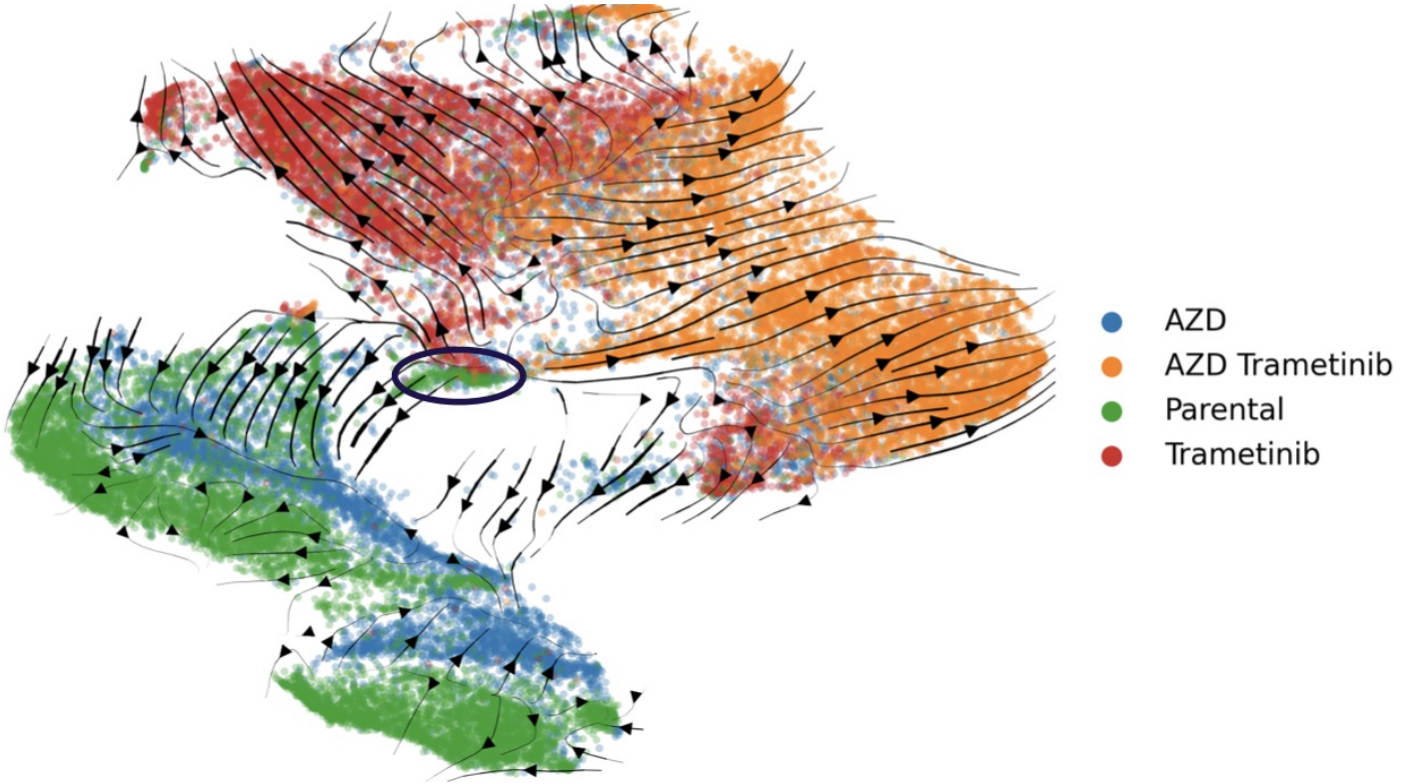
- Noise genes should have ranks uniformly distributed ← Gaussian average
- Relevant genes should have consistently high ranks
- Expect bimodal distribution



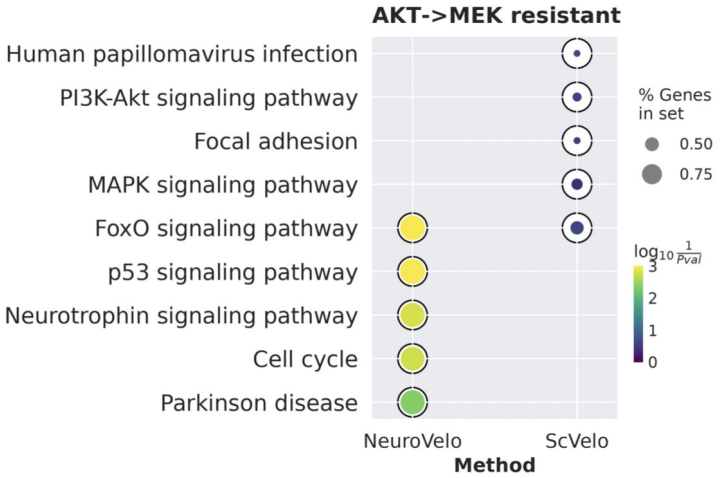
NeuroVelo on HBM

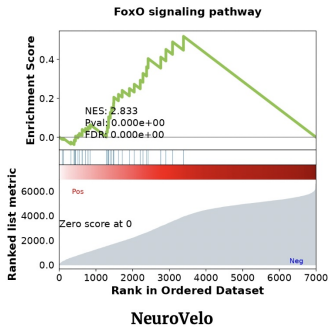


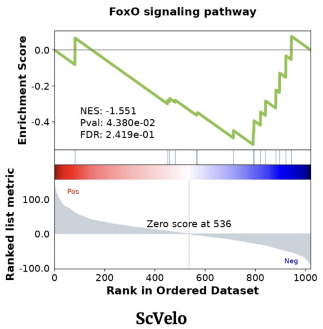
NeuroVelo on CRC



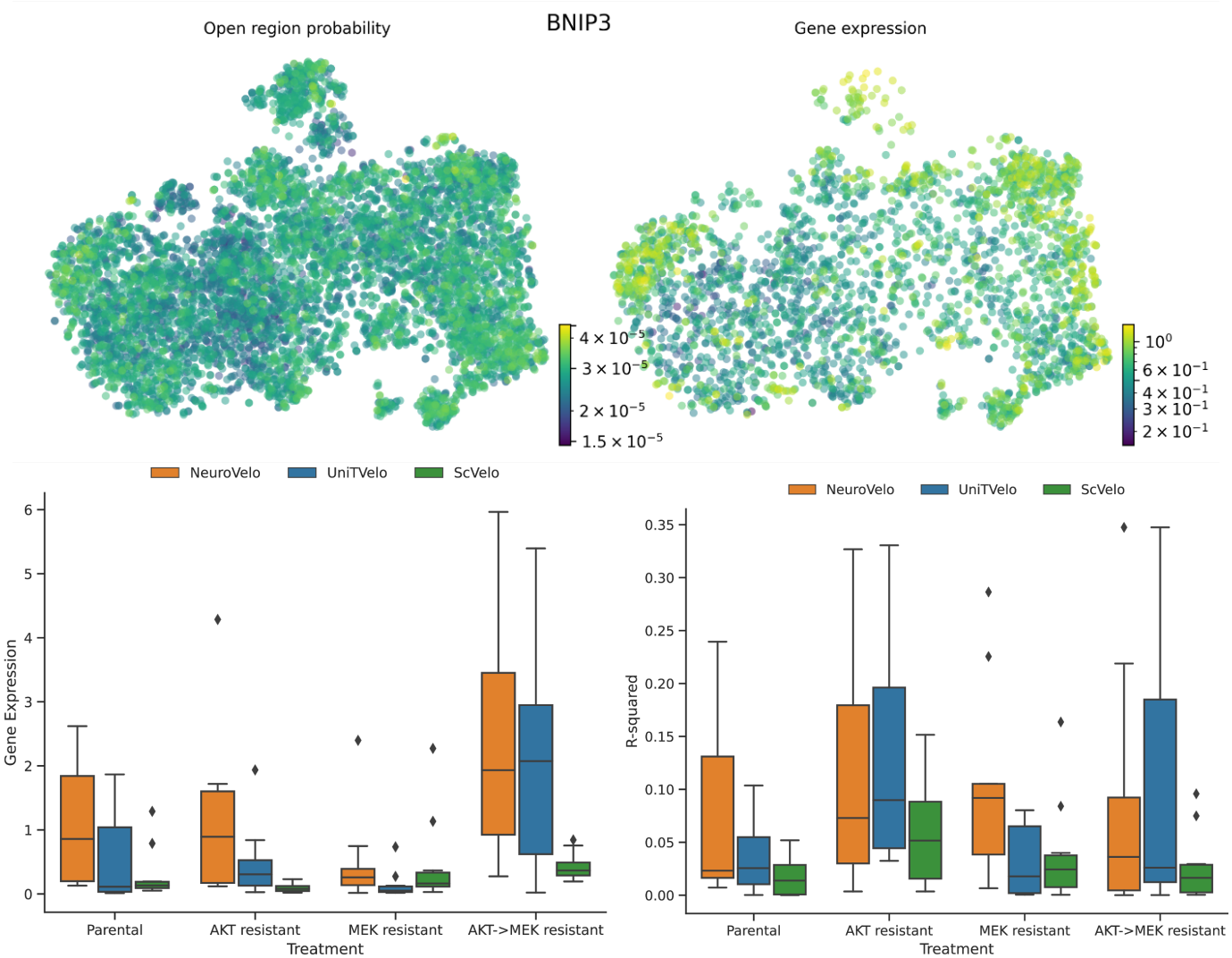
Validating NeuroVelo: enrichment







Validating NeuroVelo: multiome



Roadmap of today

Conclusions

- Single-cell 'omics provide a potential goldmine, but you need the right pick-axe
- Must go beyond simply plotting cells in latent space
- Combining interpretability and nonlinearity is still a major challenge
- Interpretability is key to progress to the clinic!

Thanks!

Collaborators and lab members/ alumni

SISSA

Riccardo Margiotta
Rongrong Xie
Viplove Arora
Nour el Kazwini
Alex Zhang
Idris Kouadri Boudjelthia
Federico Caretti
Katsiaryna Davydzenka

University of Edinburgh

Kashyap Chhatbar
Kaan Ocal
Christos Maniatis

Andreas Kapourani
Yuanhua Huang
Catalina Vallejos

Human Technopole

Andrea Sottoriva
Salvatore Milite

Funding: ERC, AIRC, SISSA/ MUR

Some references

- El Kazwini, Nour and GS, *Genom Biology* 25(55) (2024), https://link.springer.com/article/10.1186/s13059-024-03180-3
- Kouadri-Boudjelthia, Idris, et al https://www.biorxiv.org/content/10.1101/2023.11.17.567500v1
- La Manno, Gioele et al, *Nature* 460 (494-498) (2018), https://www.nature.com/articles/s41586-018-0414-6
- McInnes, L. et al, *J. Open Source Softw.* 3(29) (2018) https://joss.theoj.org/papers/10.21105/joss.00861
- Haghverdi, L. et al, *Nature Meth.* 13 (845-848) (2016), https://www.nature.com/articles/nmeth.3971
- Kiselev, V. et al, *Nature Meth.* 14 (483-86) (2017), https://www.nature.com/articles/nmeth.4236