Machine learning methodologies for single-cell 'omic data

Guido Sanguinetti

Theoretical and Scientific Data Science
SISSA, Trieste

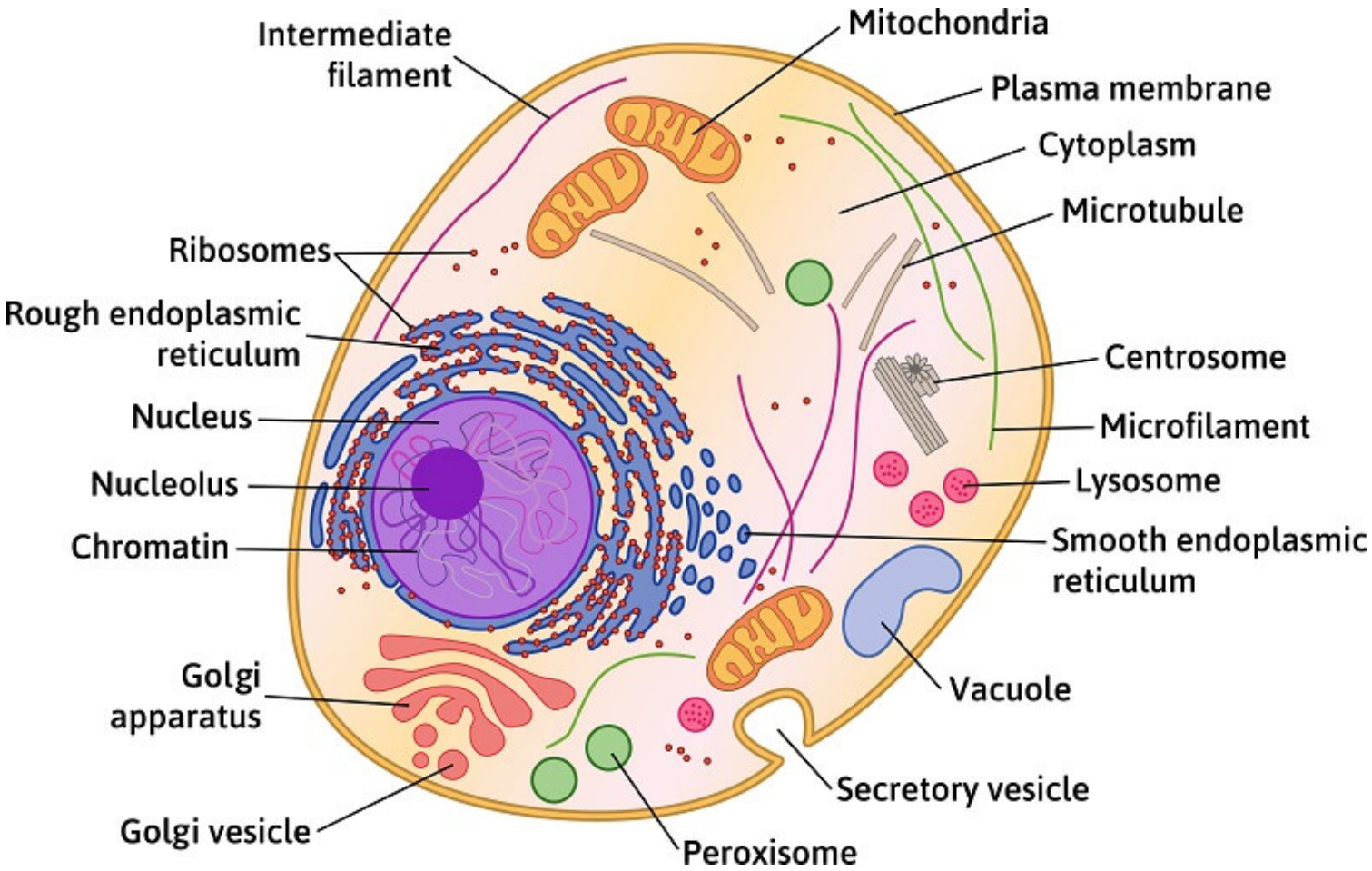
ALPACA 2023, Trento

Roadmap of today

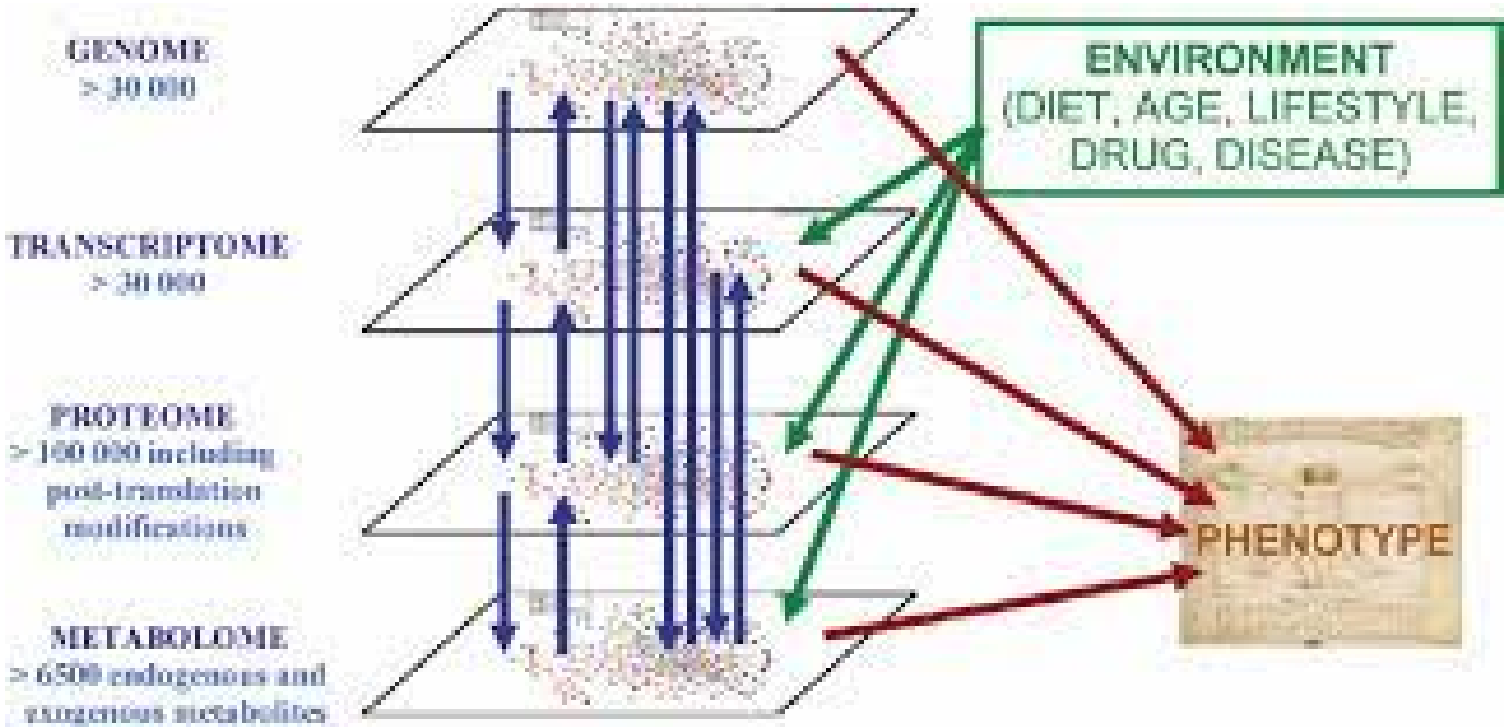
1. Biology from an ML perspective
2. Some core concepts
3. Some earlier work
4. Topic modelling for single-cell multi-omics
5. Dynamics from static data

Roadmap of today

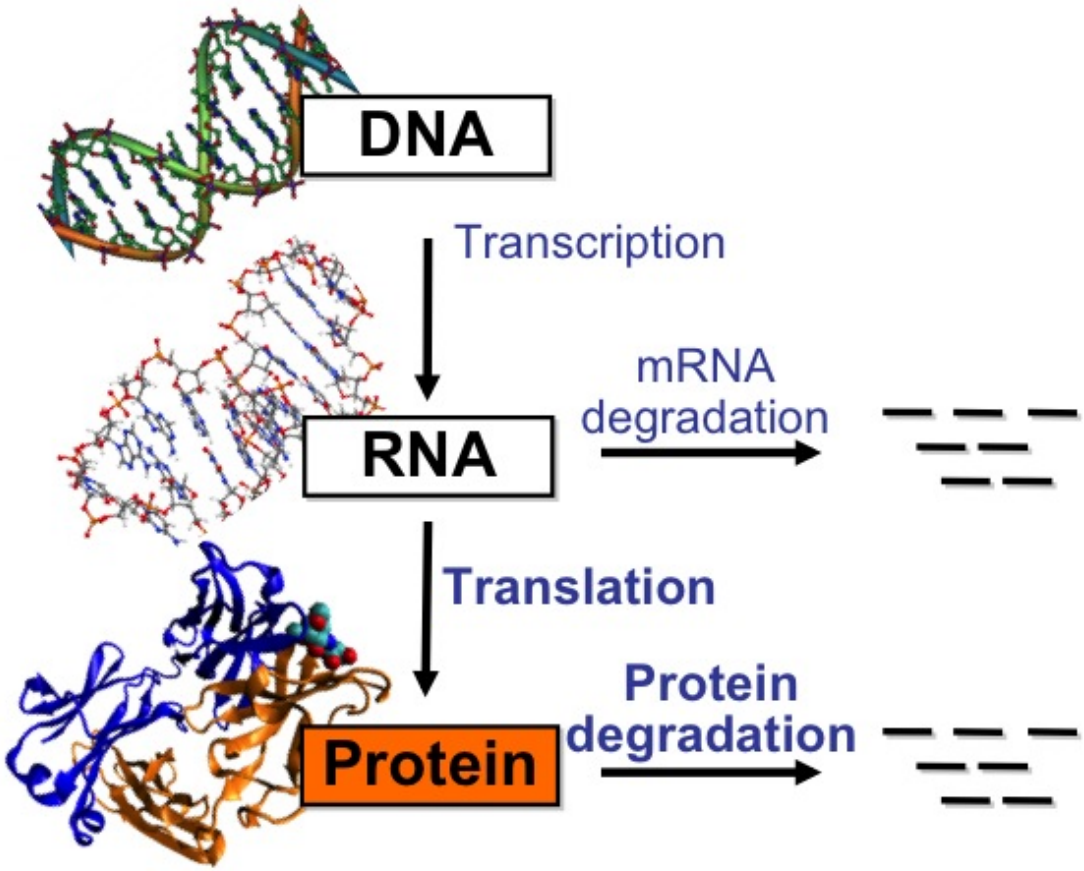
Too much detail!



A more abstract view

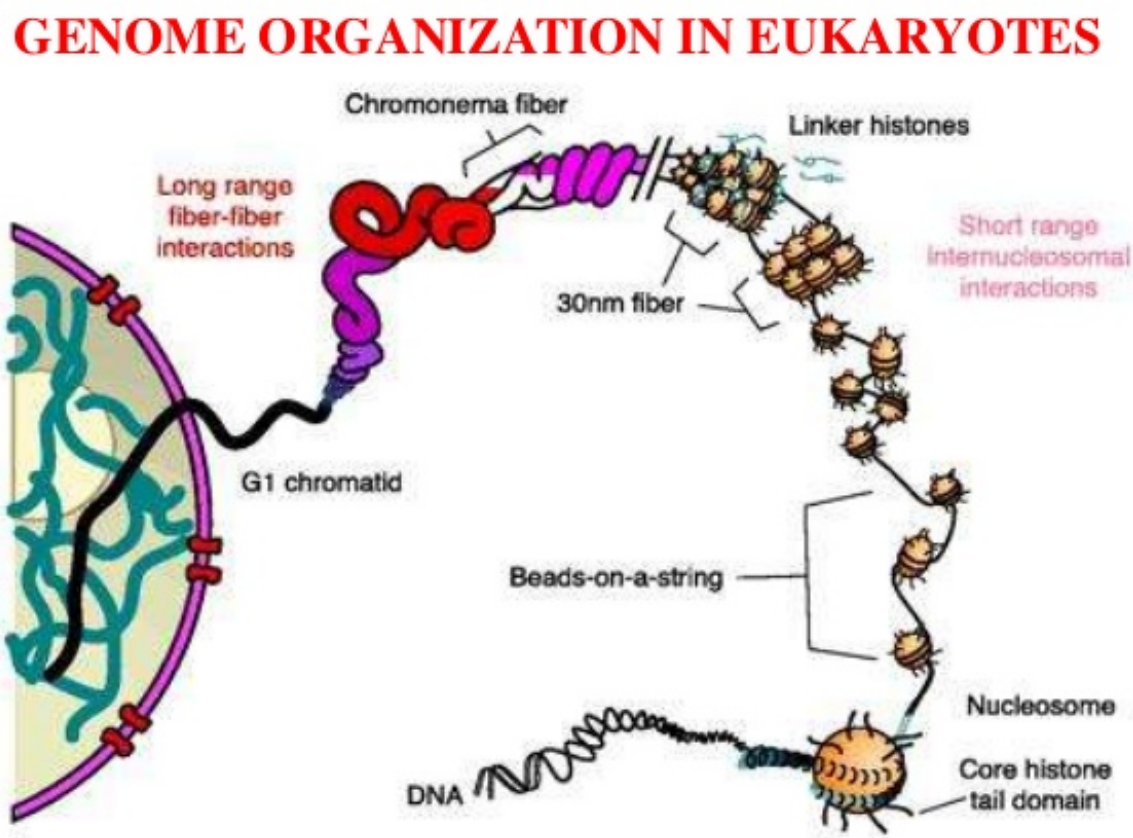


The central dogma

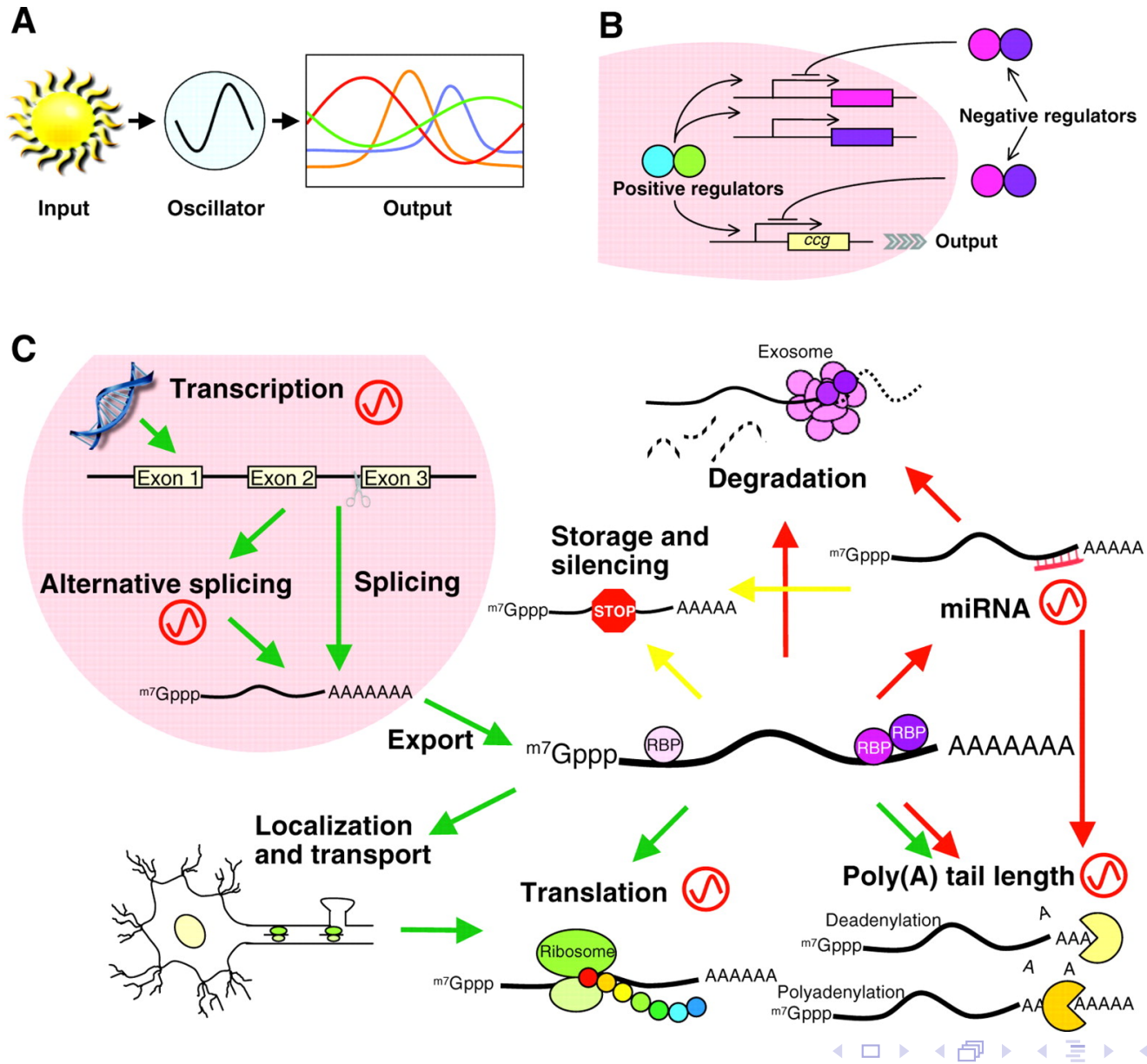


How is this regulated? Where does variability come into play?

Even more complicated



Also during and after transcription



A big challenge

- Biophysical modelling really hard!

A big challenge

- Biophysical modelling really hard!
- Systems biology: mechanistic/ semimechanistic models of cellular behaviour. Two families: ODE/ stochastic models of biological processes; metabolic models
- Bioinformatics: analysis and prediction algorithms on large biological data sets

A big challenge

- Biophysical modelling really hard!
- Systems biology: mechanistic/ semimechanistic models of cellular behaviour. Two families: ODE/ stochastic models of biological processes; metabolic models
- Bioinformatics: analysis and prediction algorithms on large biological data sets
- Stats/ unsupervised/ supervised learning on large-scale measurements

Roadmap of today

Some bio-concepts

- A *gene* is a stretch of DNA with a defined function (e.g. code for a protein)
- The collection of all genes and genetic material is the *genome* of the organism (all proteins $\rightarrow$ proteome, etc)
- Animals generally have two identical copies of the genome in each cell (*diploid*)
- When cells divide, they copy the DNA and random errors may happen. *Evolution* is the accrual of these changes, selected to increase *fitness*
- DNA is tightly wrapped around protein complexes. DNA + proteins = *chromatin*. A continuous stretch of chromatin is a *chromosome*

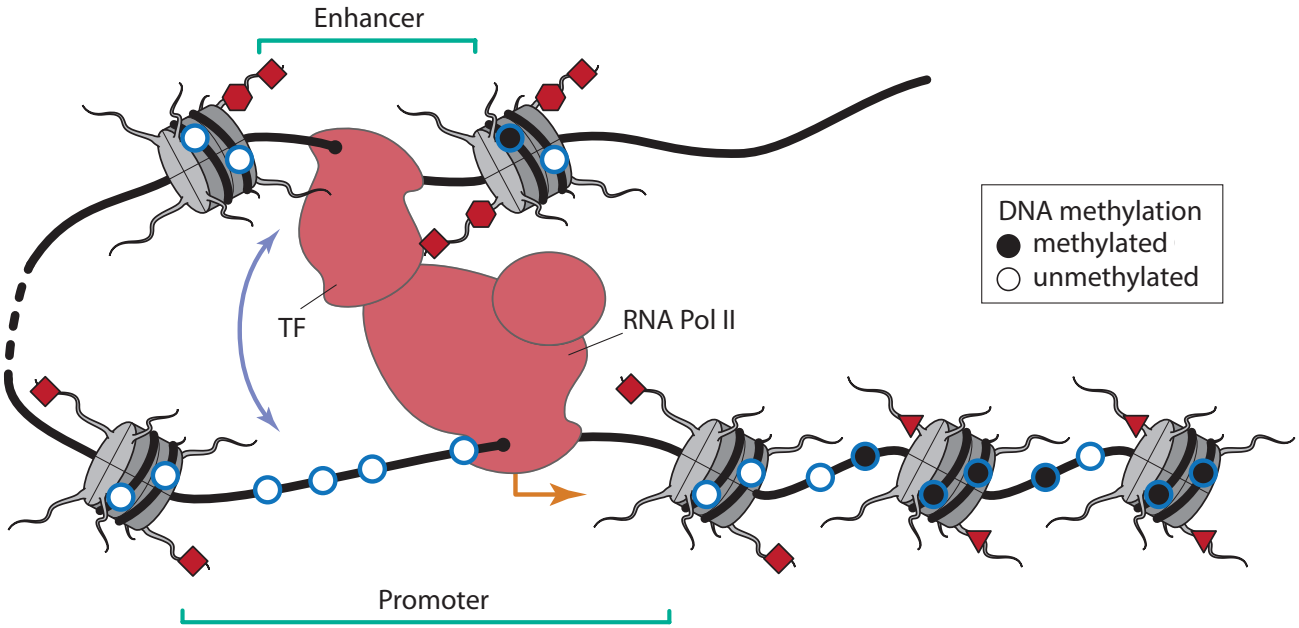
Some bio-numbers

- Number of genes $= 0.4\text{-}3 \times 10^4$ (humans $\sim 20K$)
- Number of human proteins $\sim 70K$
- Length of human genomes $\sim 3 \times 10^9 bp \sim 2m$. Total length of human DNA per individual?

Some bio-numbers

- Number of genes $= 0.4\text{-}3 \times 10^4$ (humans $\sim 20K$)
- Number of human proteins $\sim 70K$
- Length of human genomes $\sim 3 \times 10^9 bp \sim 2m$. Total length of human DNA per individual?
- Total number of annotated functions (gene ontology terms) $\sim 38K$

The sequencing multiverse



DNA methylation
- WGBS
- RRBS
- NOMe-seq
- MeDIP-seq
- MBD-seq

Histone modifications, TF binding
- ChIP-seq

Accessible Chromatin
- DNaseI-seq
- ATAC-seq
- NOMe-seq
- MNase-seq

Chromatin Interactions
- Hi-C

RNA expression
- RNA-seq

The kind of questions we want to ask

- Single-cell 'omics regularly measure hundreds/ hundreds of thousands of cells, each with thousands of features (genes/ loci)
- Variability in features (gene expression) and interactions can be explained as technical/ biological/ intrinsic

The kind of questions we want to ask

- Single-cell 'omics regularly measure hundreds/ hundreds of thousands of cells, each with thousands of features (genes/ loci)
- Variability in features (gene expression) and interactions can be explained as technical/ biological/ intrinsic
- The law of total variance

$$\mathrm{Var}[Y] = E[\mathrm{Var}[Y|X]] + \mathrm{Var}[E[Y|X]]$$

decomposes observed variance in *unexplained* and *explained* components

The kind of questions we want to ask

- Single-cell 'omics regularly measure hundreds/ hundreds of thousands of cells, each with thousands of features (genes/ loci)
- Variability in features (gene expression) and interactions can be explained as technical/ biological/ intrinsic
- The law of total variance

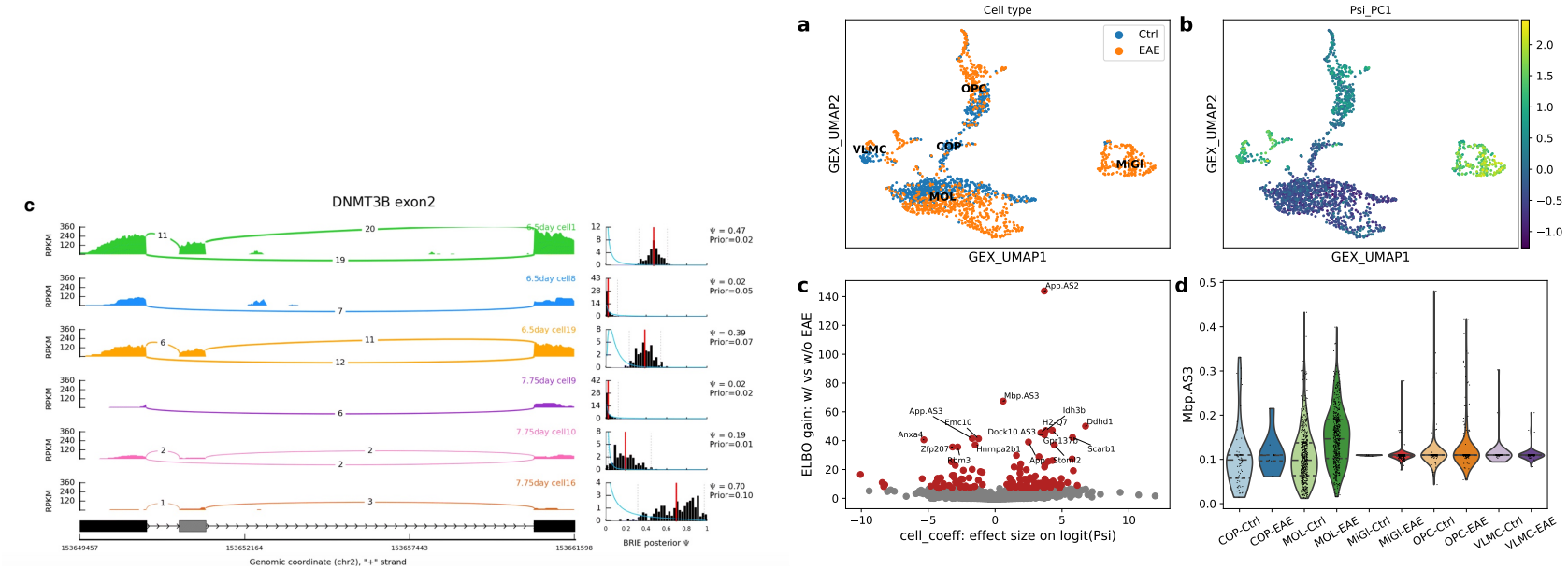
$$\mathrm{Var}[Y] = E[\mathrm{Var}[Y|X]] + \mathrm{Var}[E[Y|X]]$$

decomposes observed variance in *unexplained* and *explained* components

- You need to tease apart what is shared from what is individual

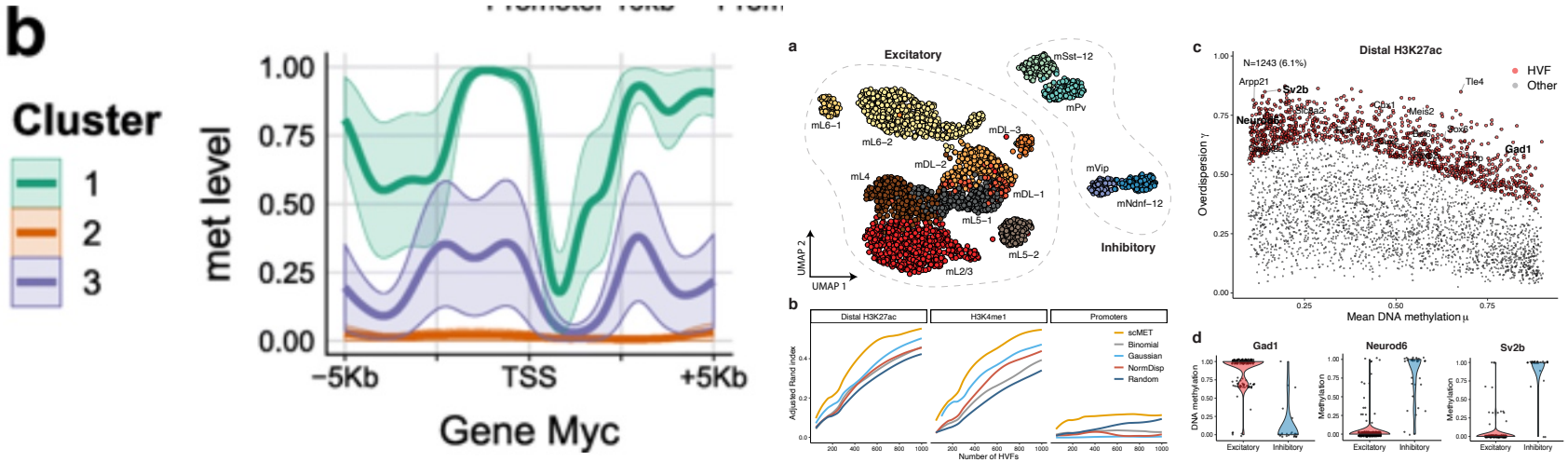
Roadmap of today

BRIE: splicing quantification in scRNA-seq



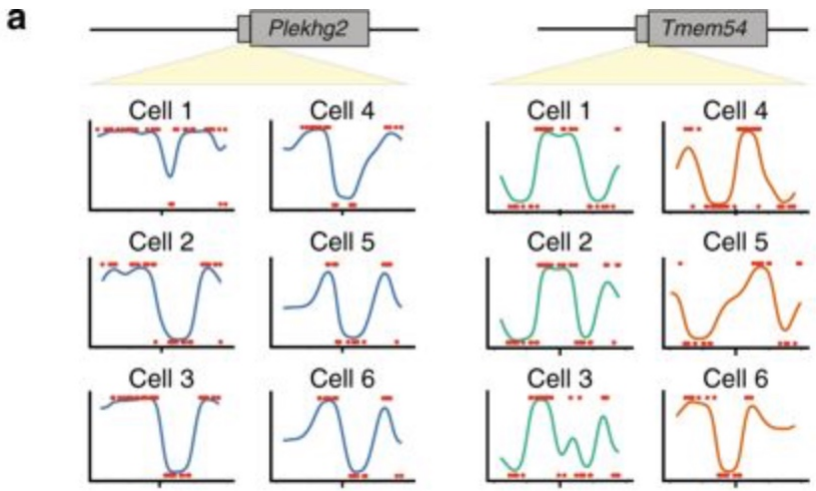
Splicing quantification in scRNA-seq leveraging sequence (Huang and Sanguinetti, *Genome Biology* 2017,2021)

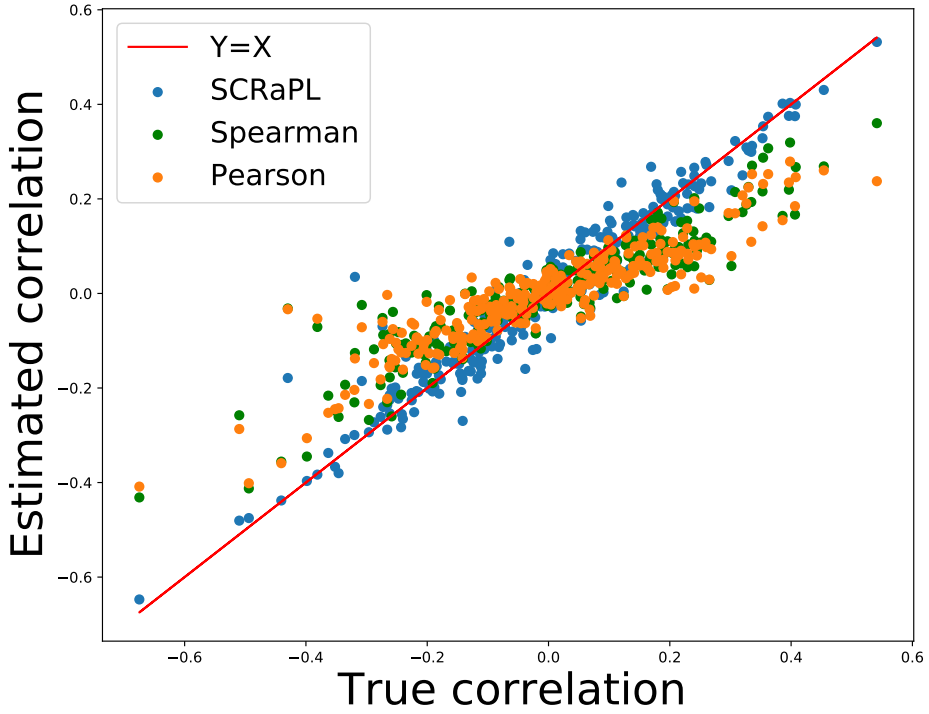
Single-cell methylation



Clustering and variability in scBS-seq leveraging chromosomal location (Kapourani and Sanguinetti *Genome Biology* 2017, Kapourani et al *Genome Biology* 2021)

Single-cell multi-omics





Studying and modelling correlations between different layers in single cells (Clark et al *Nat. Commun.* 2018, Maniatis et al *PLoS CompBio* 2022).

Talk outline

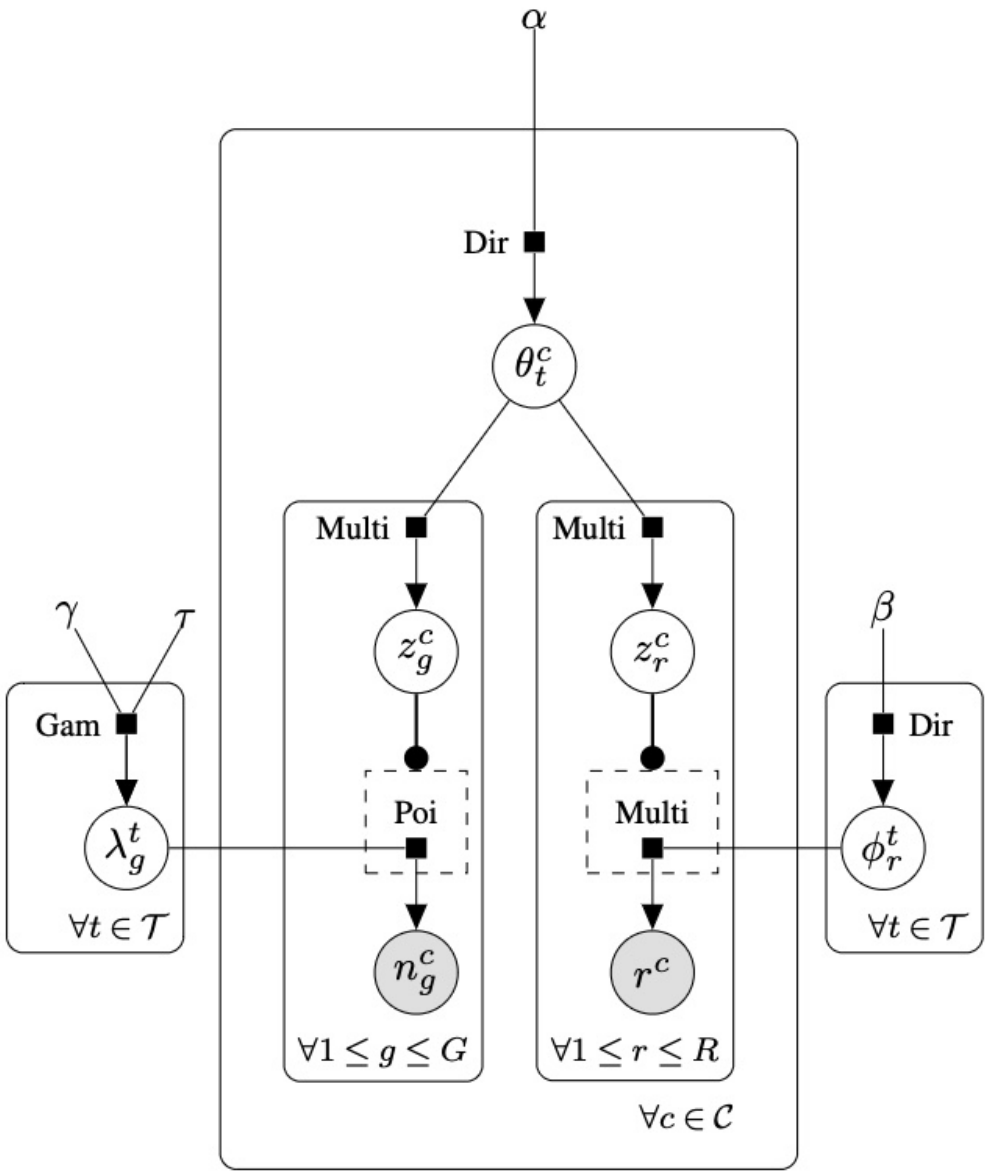
Intepretable cell-level models

- Single-cell multi-omics: two large feature vectors for each cell (e.g. accessible regions by ATAC, gene expression)
- Each feature vector ¿10K dim, each vector ¿90% zeros, about 10K cells

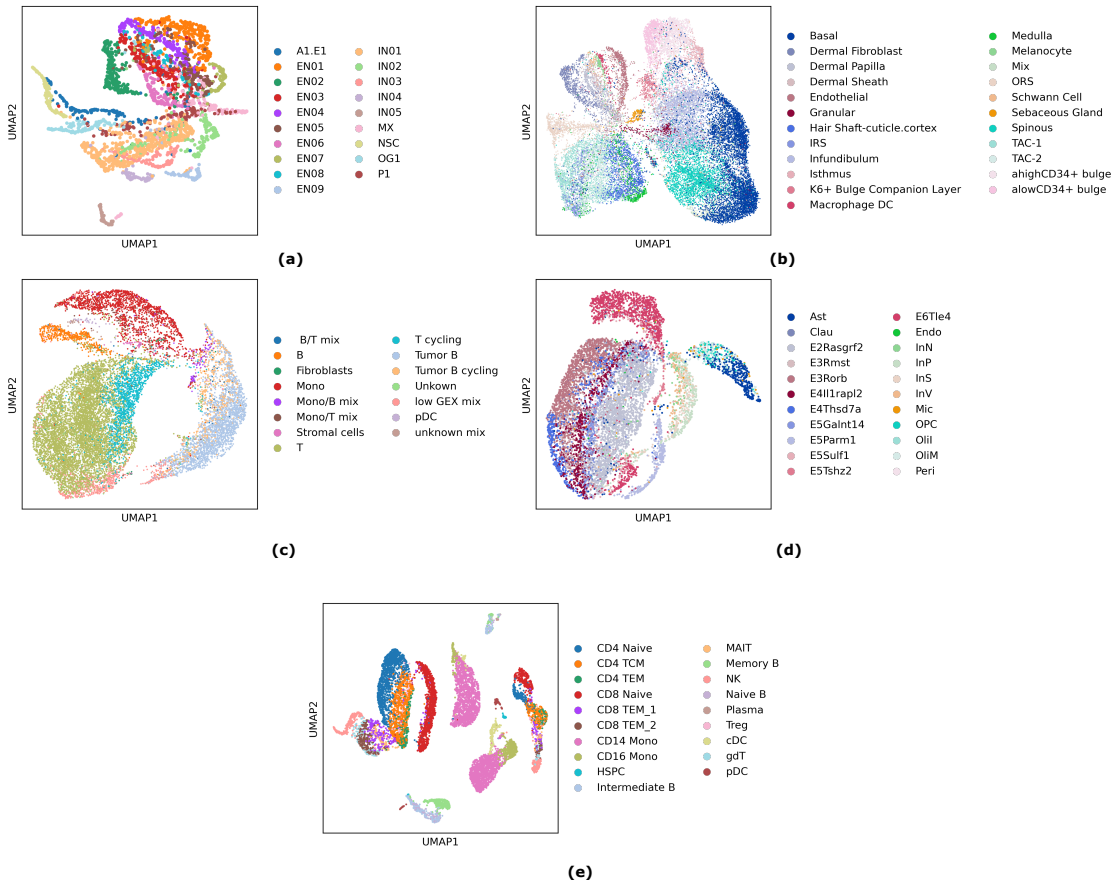
Intepretable cell-level models

- Single-cell multi-omics: two large feature vectors for each cell (e.g. accessible regions by ATAC, gene expression)
- Each feature vector ¿10K dim, each vector ¿90% zeros, about 10K cells
- Can we provide latent representations of cells which are interpretable at the level of interactions between individual features (genes/ regions)?

SHARE-topic (Nour El-Kazwini)

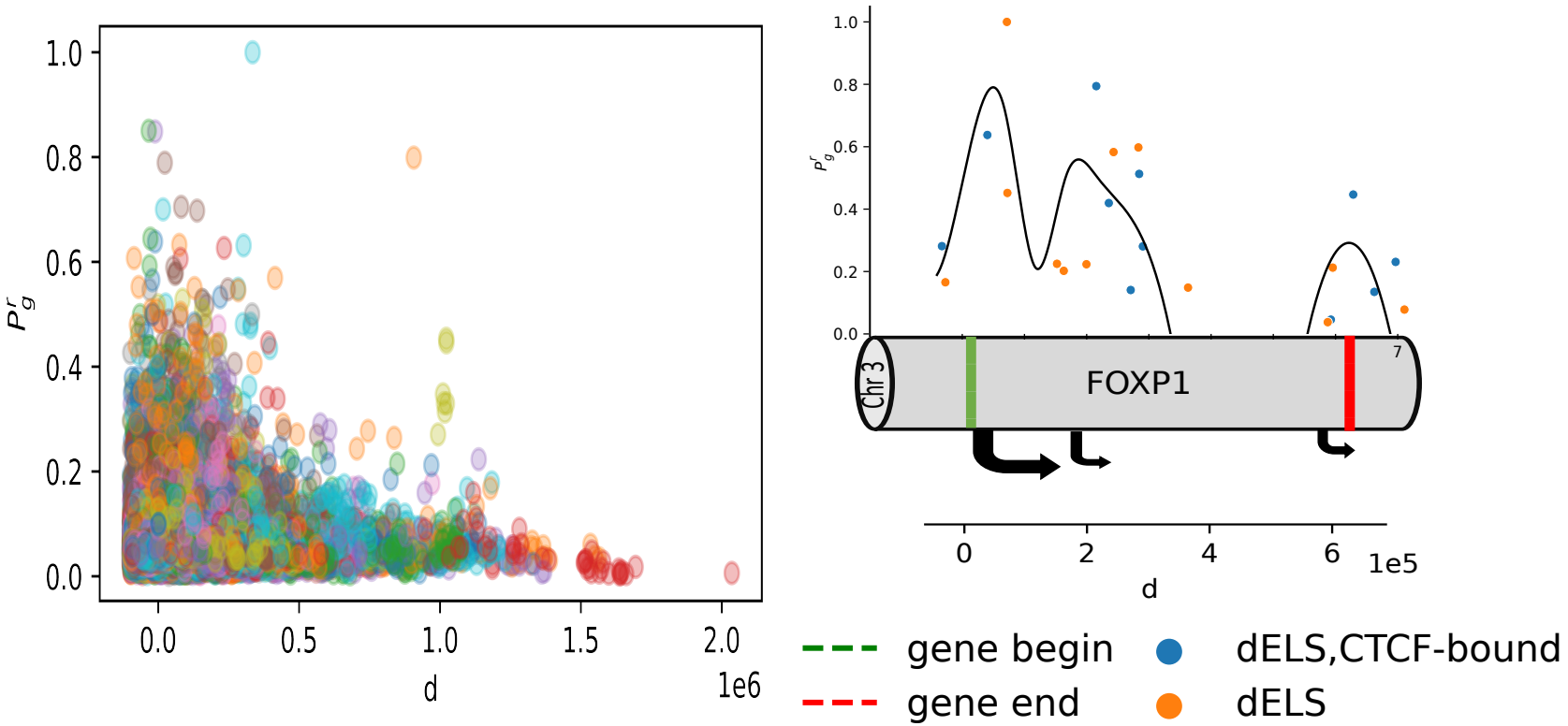


SHARE-topic: cell level results



Topics can be associated with cell types by enrichment. Also topics can be interpreted biologically by looking at which genes are highly expressed in each topic.

SHARE-topic: gene-level insights



Can create local region-gene associations (joint probability having marginalised topic)

Roadmap of today

1. Biology from an ML perspective
2. Some core concepts
3. Some earlier work
4. Topic modelling for single-cell multi-omics
5. Dynamics from static data

Uncovering dynamics: RNA velocity

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

Uncovering dynamics: RNA velocity

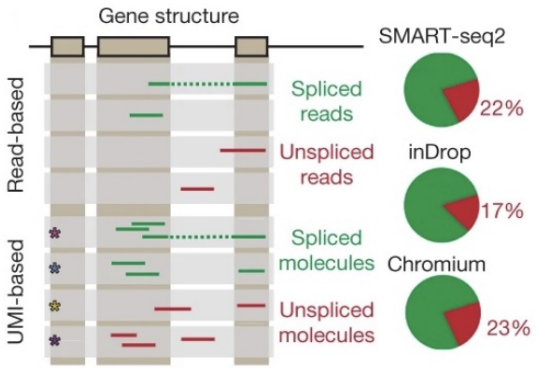
- scRNA-seq is destructive $\rightarrow$ static snapshots from a dynamic process
- **IDEA** (La Manno et al, 2018): use spliced/ unspliced reads to derive *rate of change* of RNA levels

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

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

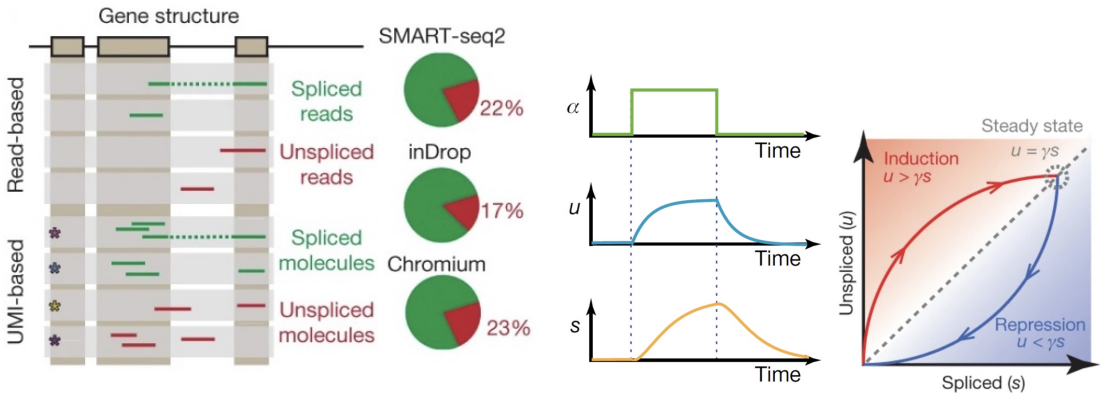
$$\frac{dx_u}{dt} = \alpha - \beta x_u \qquad \frac{dx_s}{dt} = \beta x_u - \gamma x_s$$



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

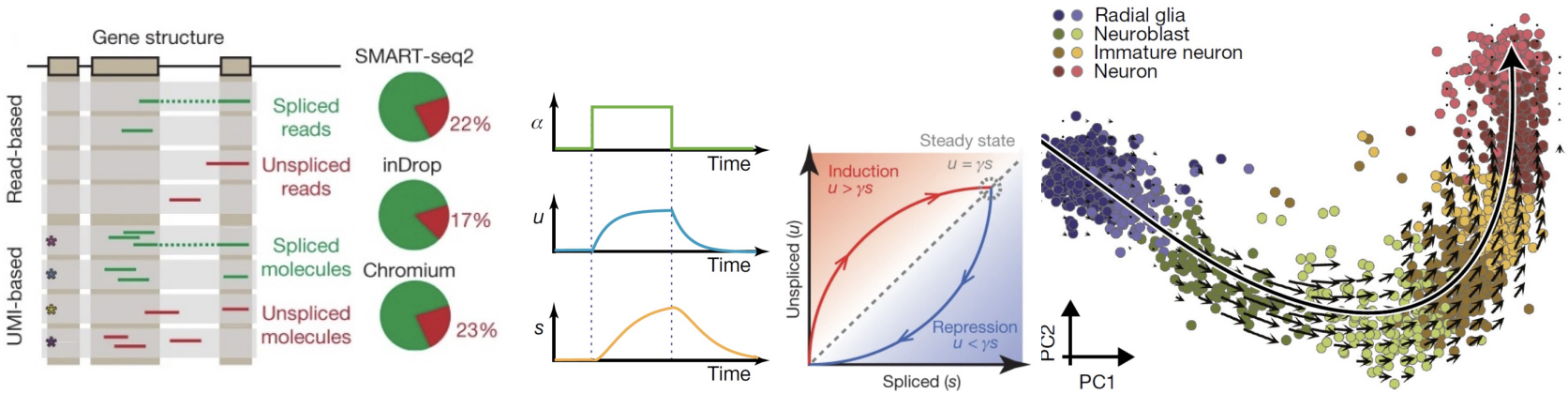
$$\frac{dx_u}{dt} = \alpha - \beta x_u \qquad \frac{dx_s}{dt} = \beta x_u - \gamma x_s$$



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

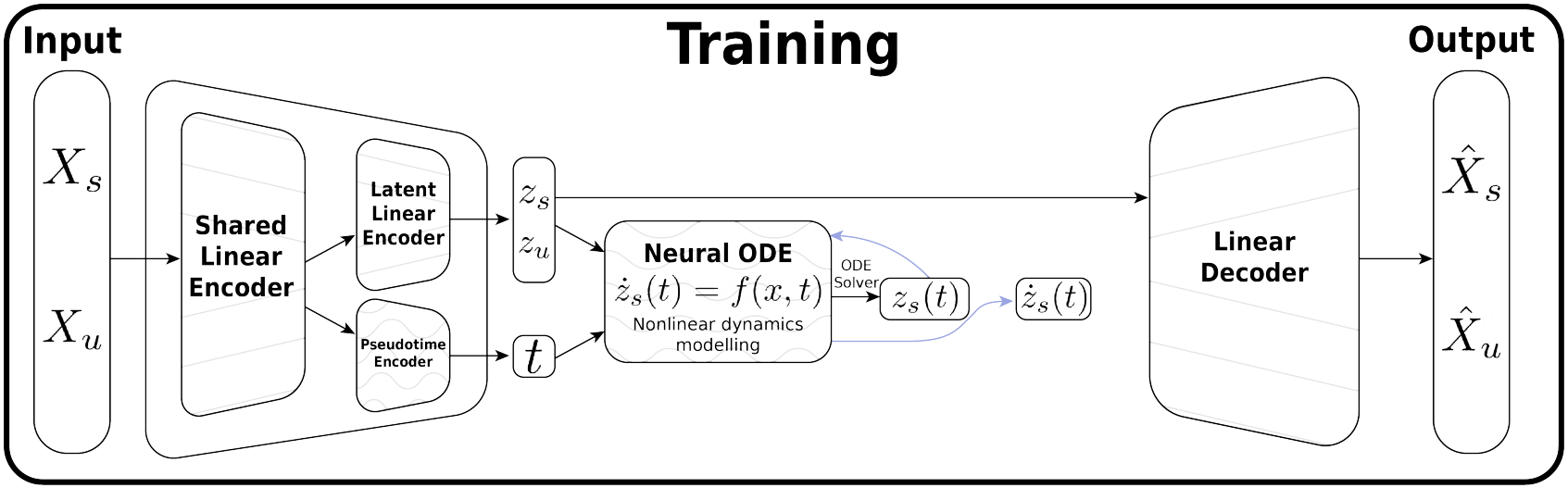
Problems and solutions

- 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*

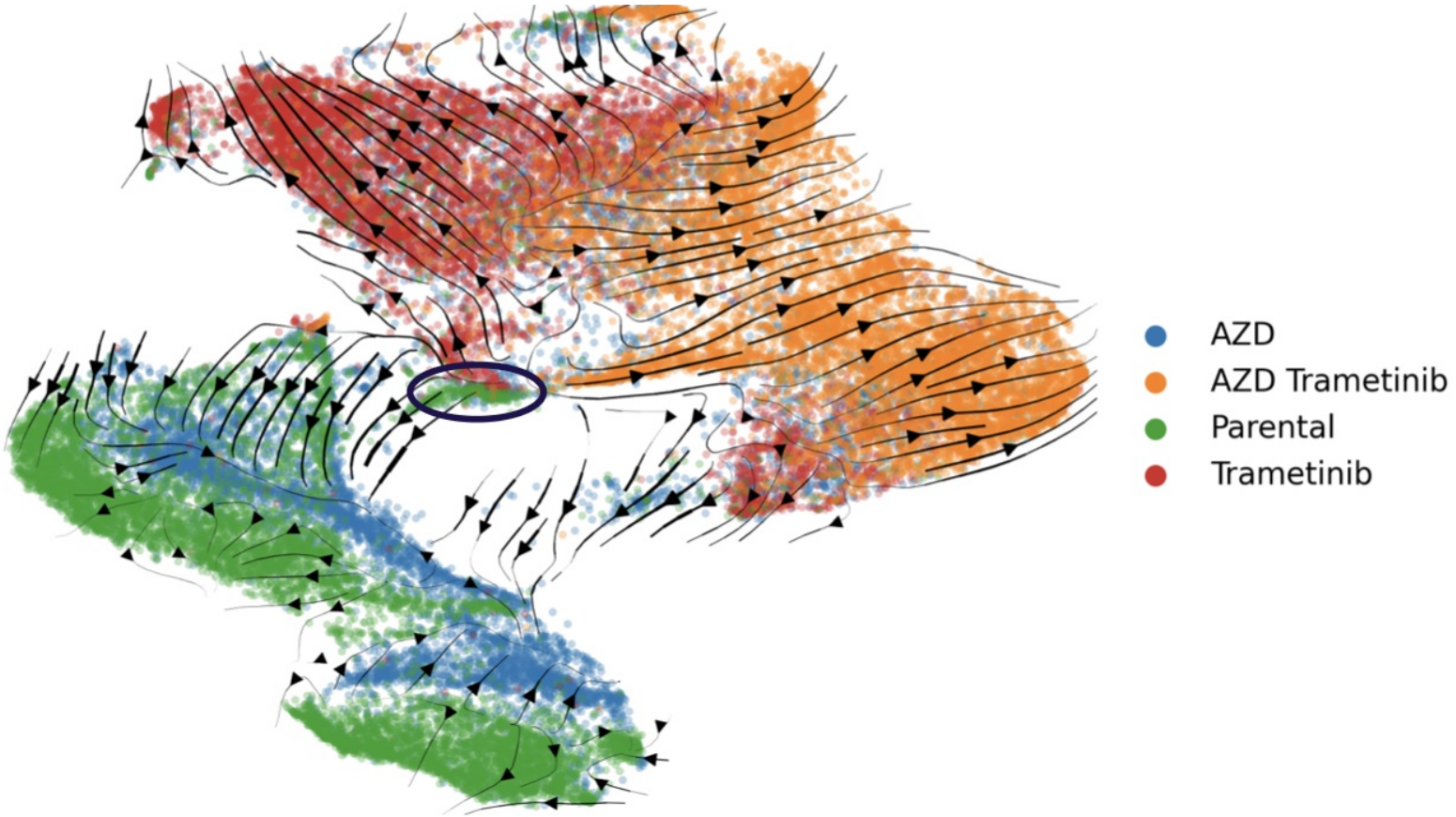
NeuroVelo (Idris Kouadri Boudjelthia)



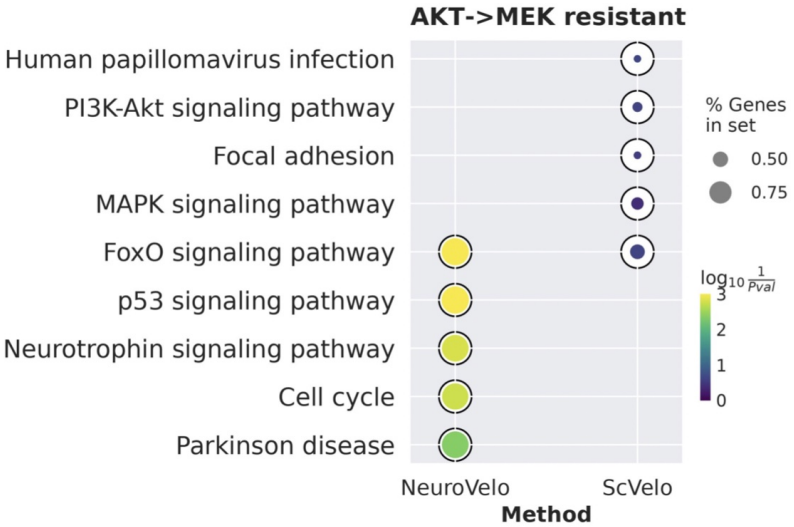
$$\mathcal{L} = \mathrm{MSE}(X, \hat{X}) + \mathrm{RNAvelocityterm}$$

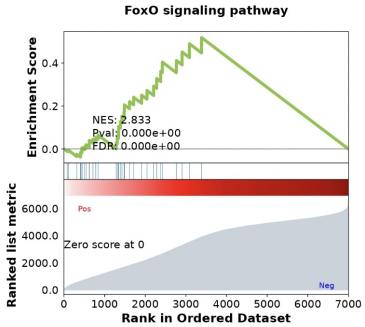
Nonlinear dynamical system interpretable via standard spectral techniques

NeuroVelo on CRC

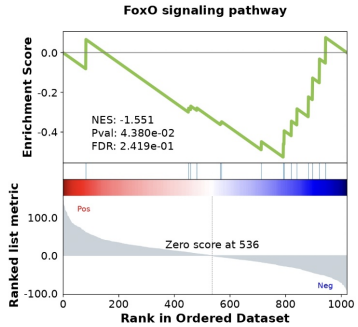


Interpreting NeuroVelo: enrichment





NeuroVelo



ScVelo

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