Single-cell trajectories reconstruction: ideas, methods and problems

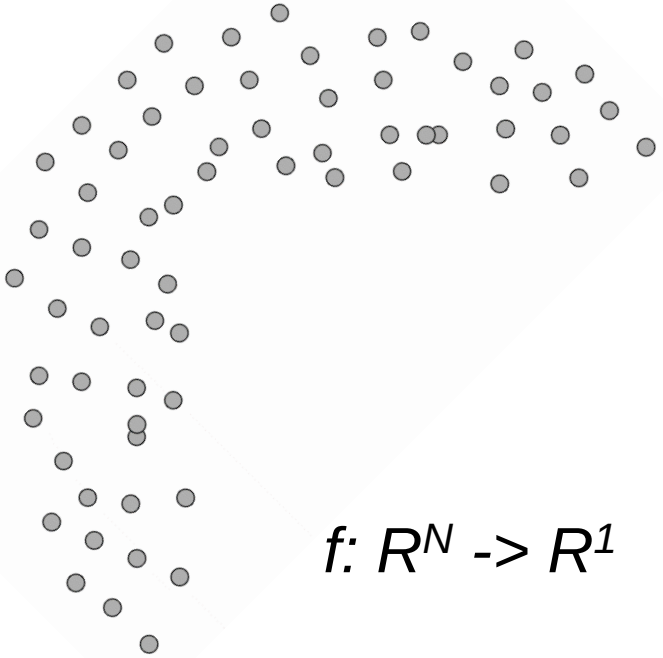
Andrei Zinovyev

Institut Curie - INSERM U900 - PSL Research University / Mines ParisTech

Computational Systems Biology of Cancer

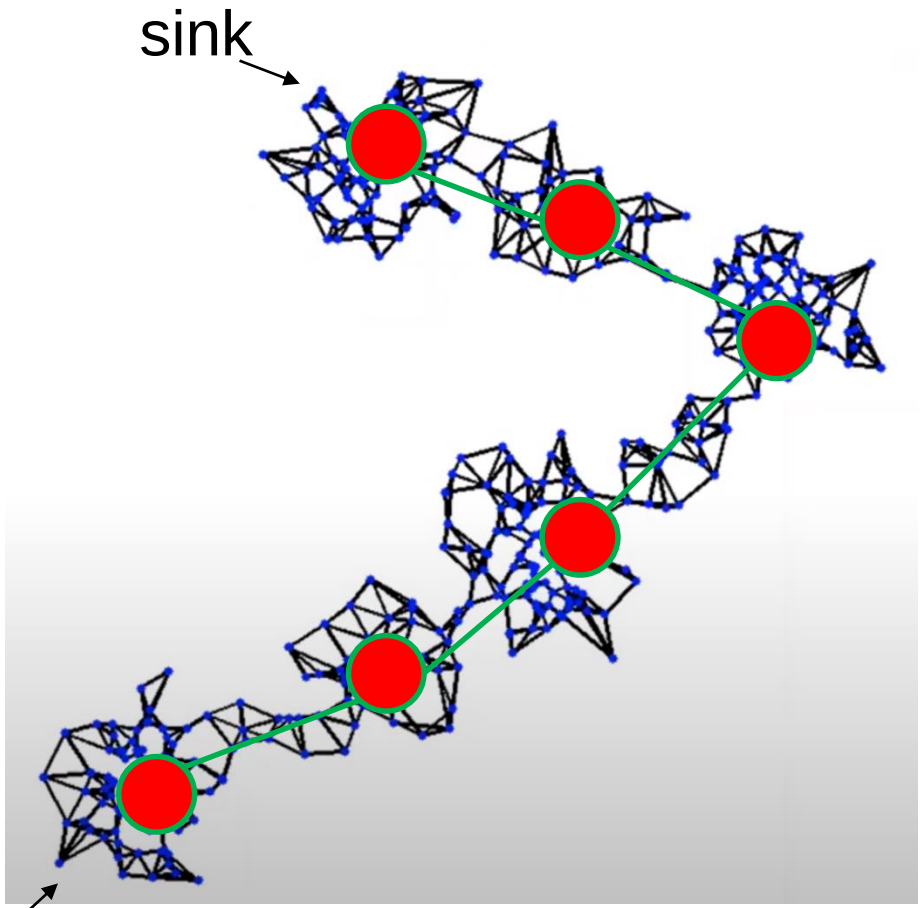
Problem : ordering object states accordingly to hypothetical progression through a dynamical process

Synchronic data (snapshot image)

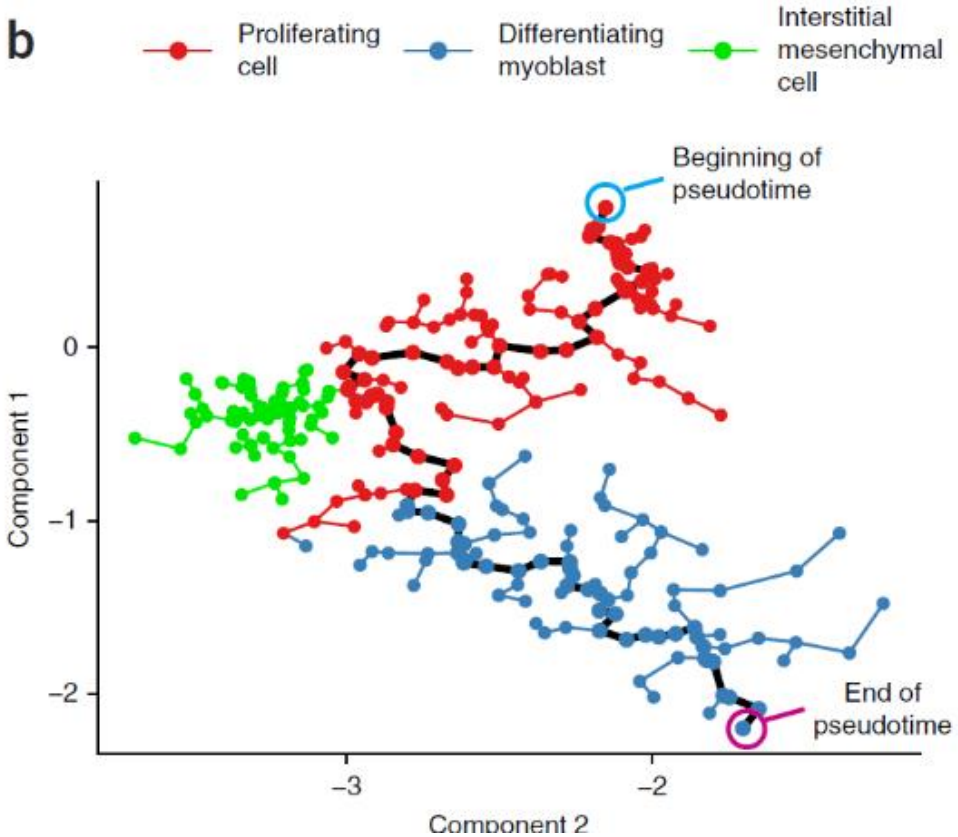


$f: R^N \rightarrow R^1$

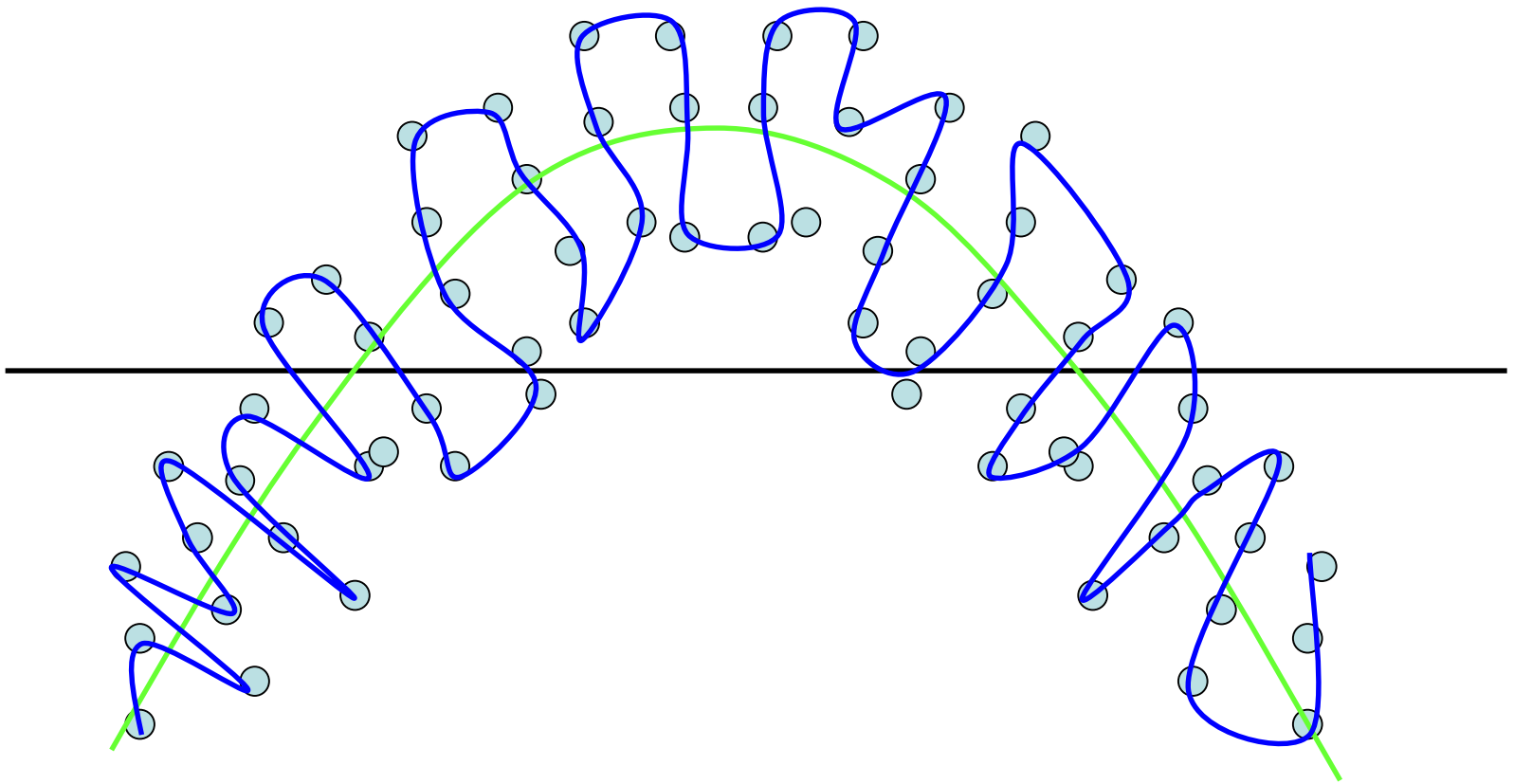
Idea 1: Traversal of kNN graph



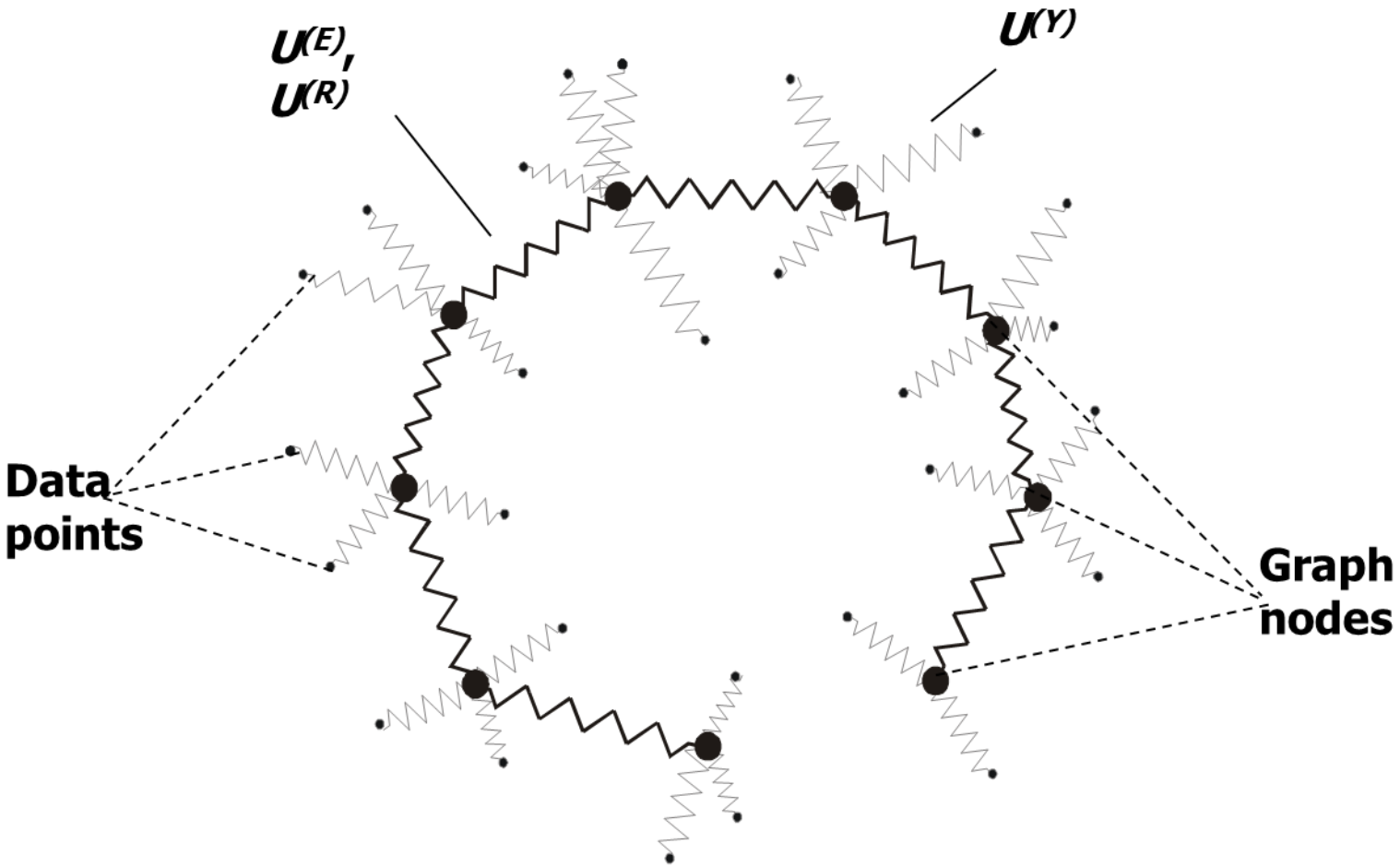
Example: Monocle 1 (suggested in 2014)



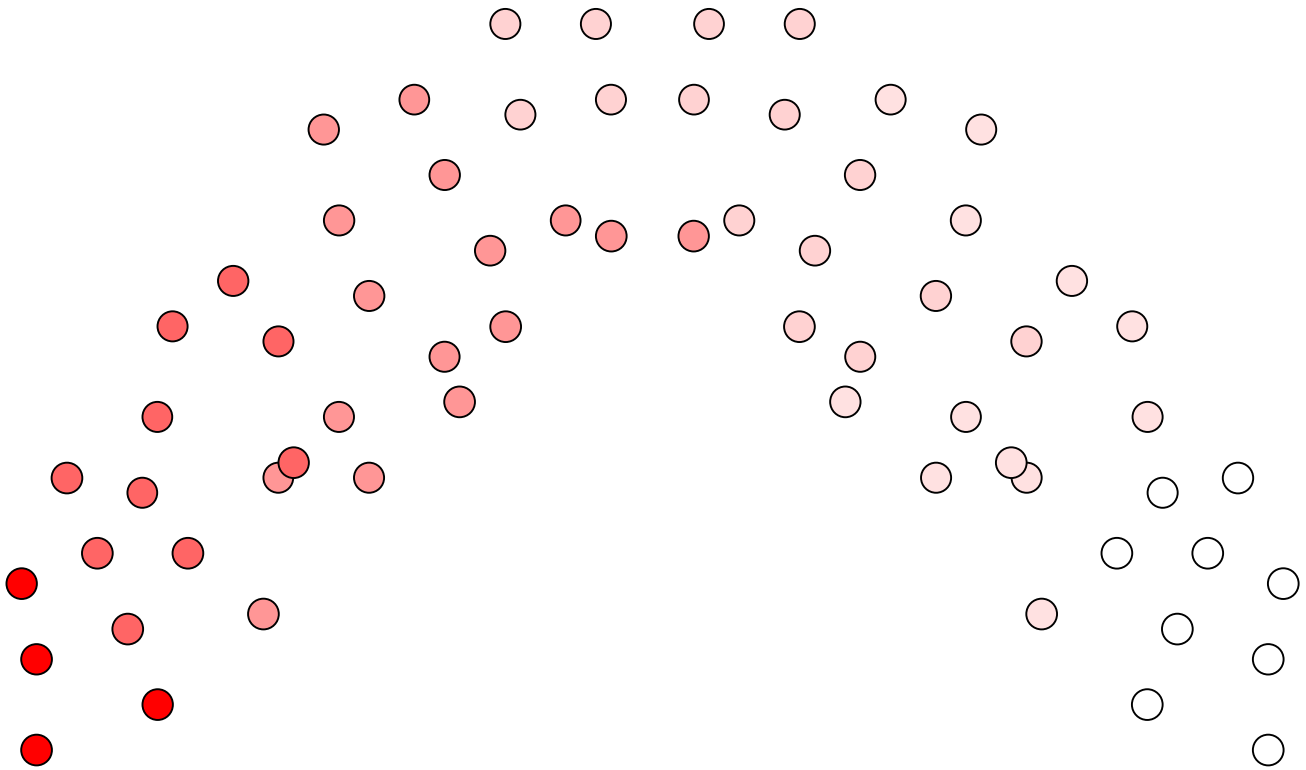
Idea 2: principal curve



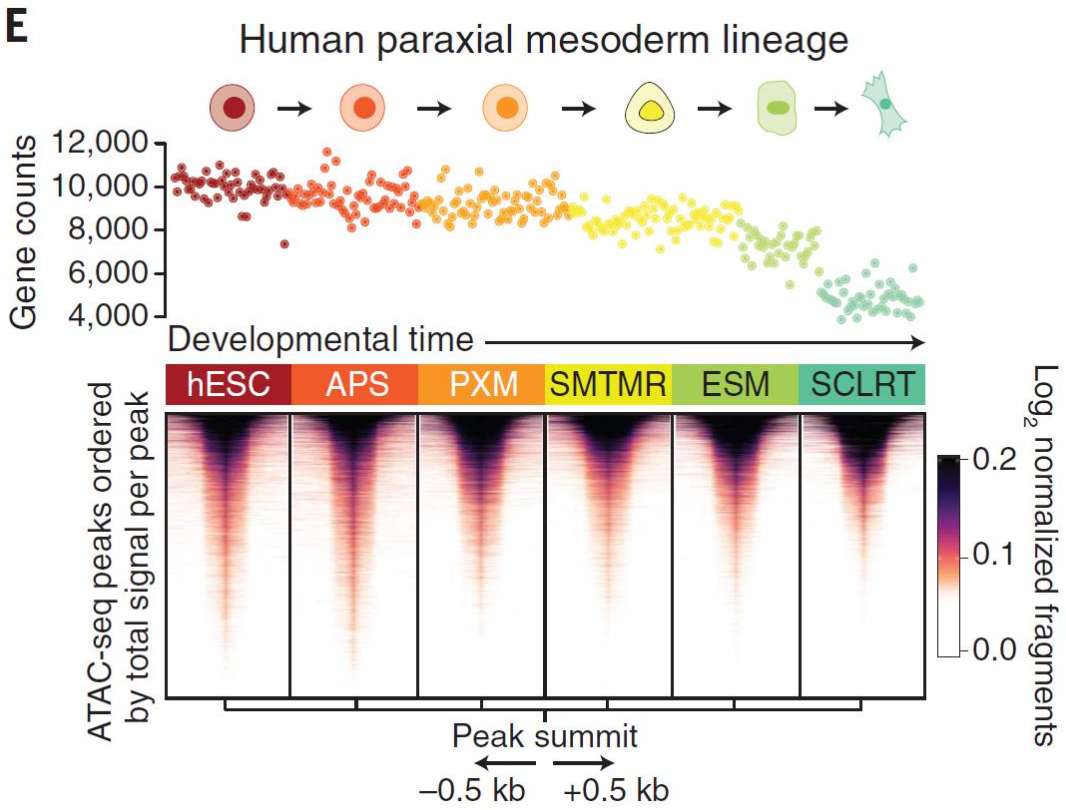
Example: 1D elastic map (suggested in 1998)



Idea 3: If we can guess a potential (Lypunov function of the process)



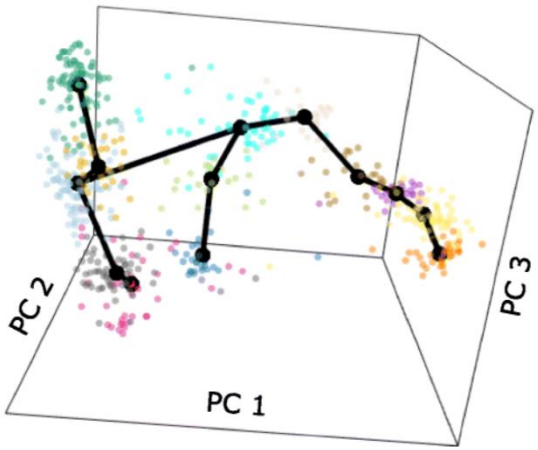
Example: CytoTRACE



Gulati et al, Science, 2020

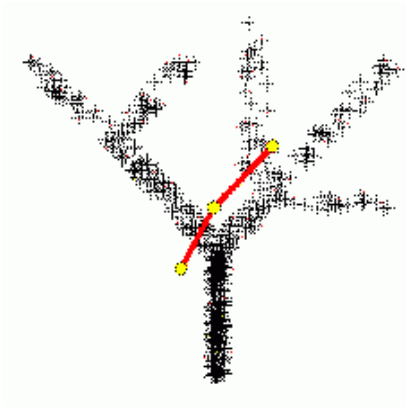
Complication 1: branching

Idea 1:
Minimal Spanning Tree



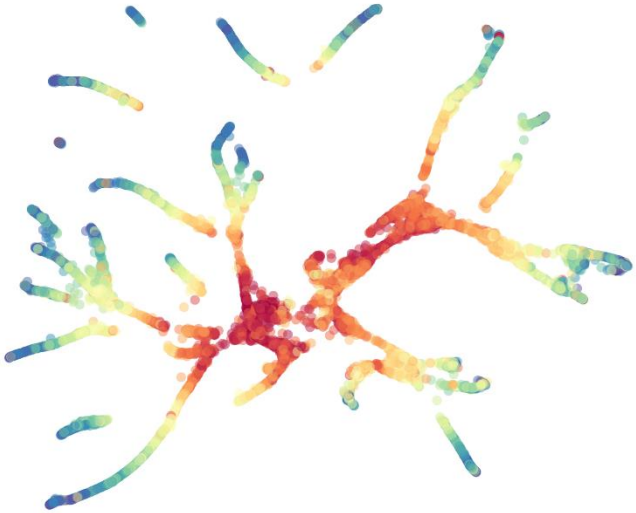
E.g., SlingShot

Idea 2:
Principal Tree



E.g., Elastic principal trees

Idea 3:
Potential



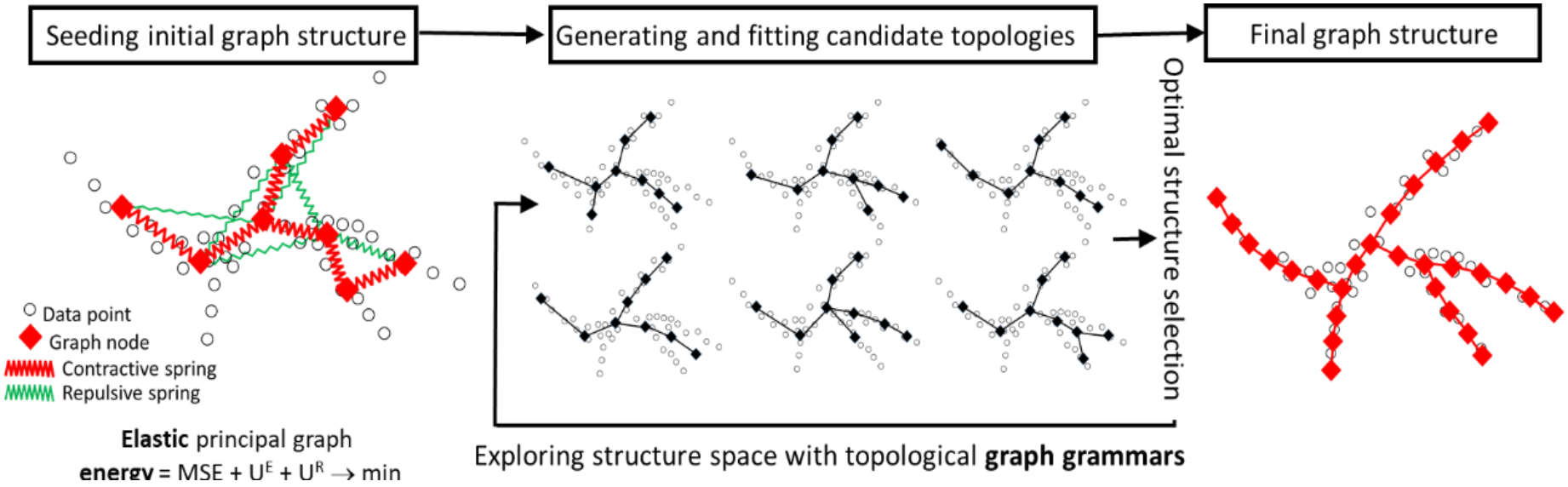
E.g., CytoTRACE

Problems with branching

1) Find the right tree structure
2) Fit the structure to the data
3) Determine position of branching points

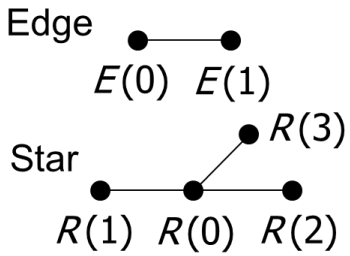
Elastic principal graphs (ElPiGraph)

(Gorban&Zinovyev,2007; Zinovyev&Mirkes, 2013; Gorban&Zinovyev, 2010; Albergante et al, 2018; Chen et al, 2019; *book* Gorban, Kegl, Wunch, Zinovyev, LNSC, 2008)



Penalty on **total length**:

Penalty on deviation from **harmonicity**:



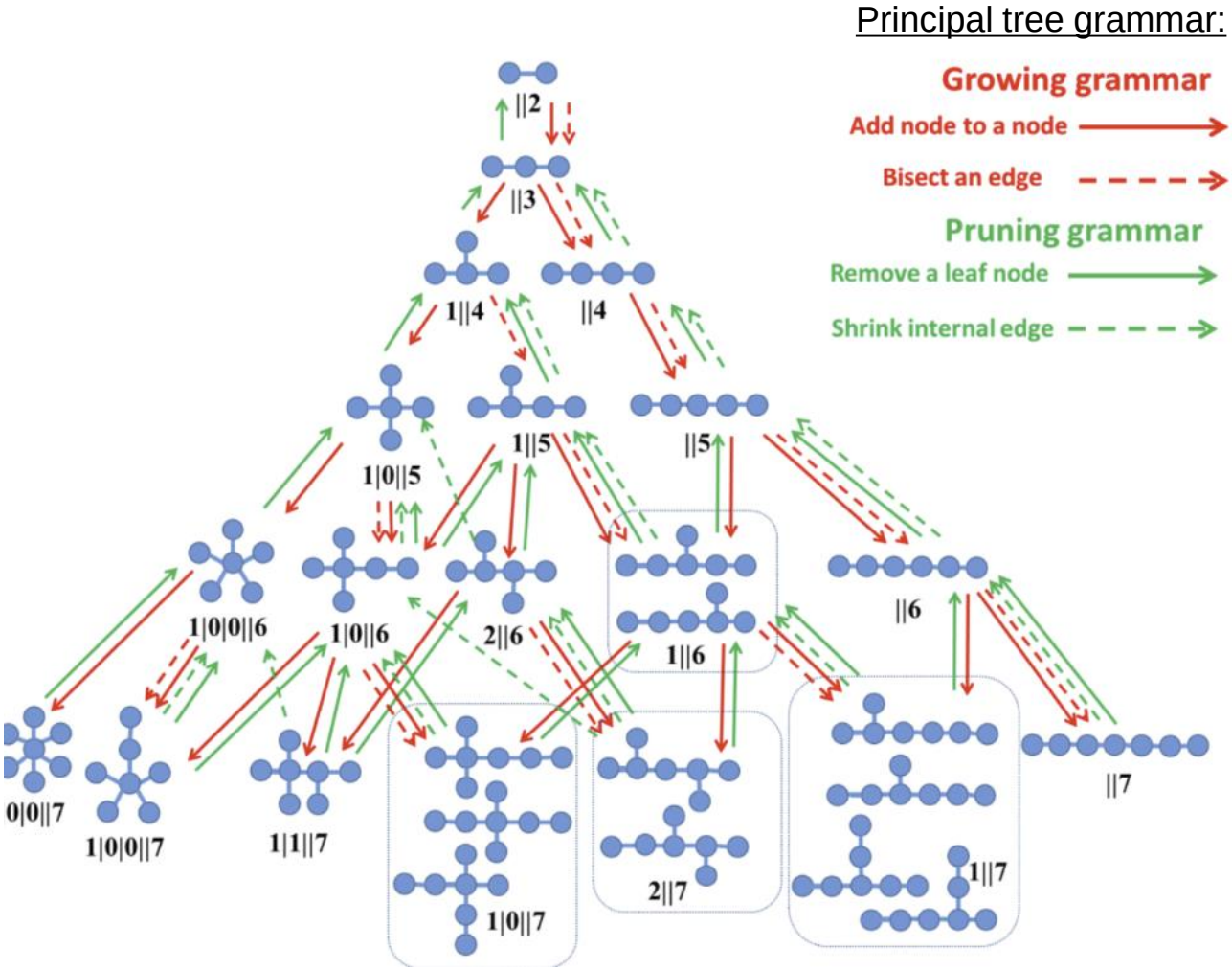
$$U^{(E)} = \sum_{i=1}^{s} \lambda_i \left\| E^{(i)}(1) - E^{(i)}(0) \right\|^2$$

$$U^{(R)} = \sum_{i=1}^{r} \mu_i \left\| R^{(i)}(0) - \frac{1}{k} \sum_{j=1..k} R^{(i)}(j) \right\|^2$$

General-purpose machine learning method

Implementations in **MATLAB, R, Python, Scala, Java, adapted to TensorFlow**

Topological grammars and gradient-based descent in the discrete space of graph structures



institutCurie

Instituts thématiques Inserm Institut national de la santé et de la recherche médicale

MINES ParisTech

PSL RESEARCH UNIVERSITY

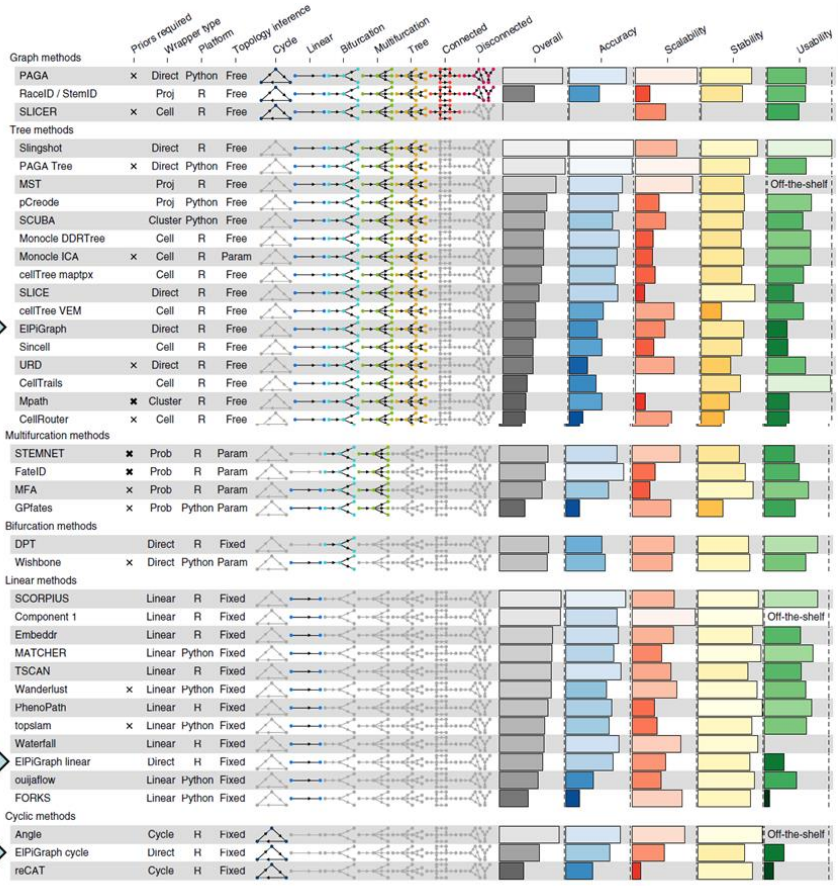
Benchmarking 45 (out of 70!) methods for cell trajectory inference

Saelens et al, Nat Biotech, 2019

ElPiGraph

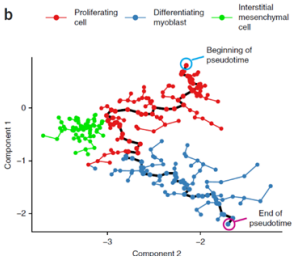
ElPiGraph linear

ElPiGraph cycle



Based on “distilling”
K nearest neighbours (kNN) graph

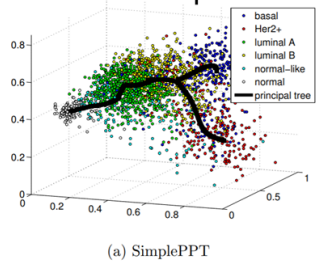
Example



From Trapnell et al, 2014

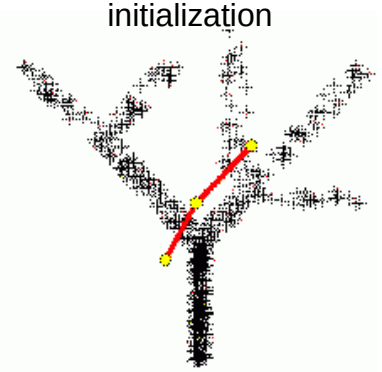
Based on
graph embedment (injection)

Example



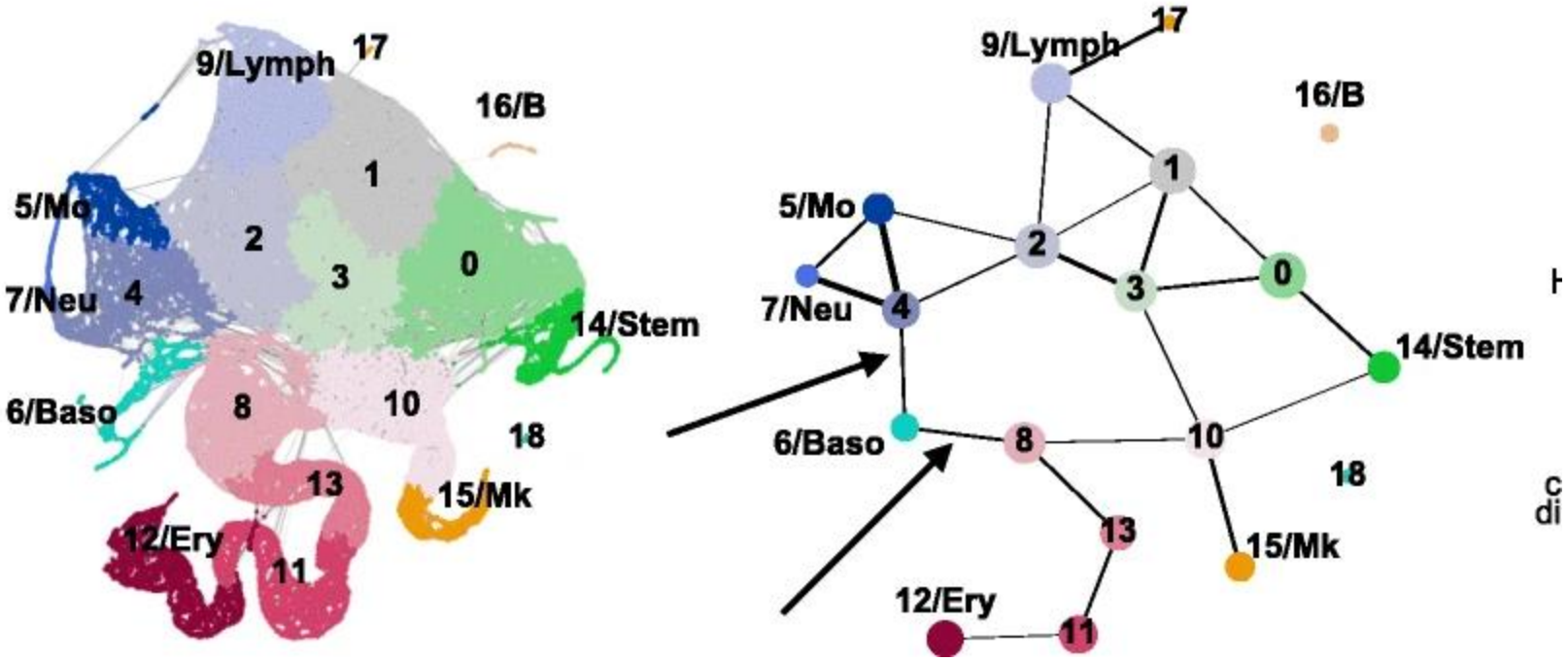
From Mao et al, 2015

Problem of graph initialization



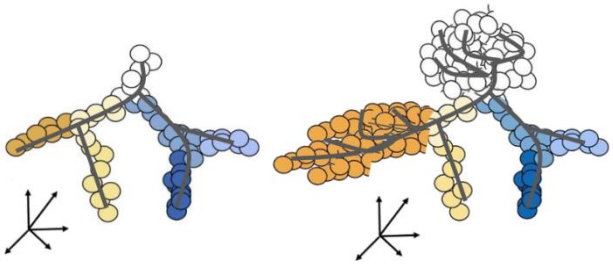
ElPiGraph is a general-purpose machine learning method
It needs to be adapted to the nature of single cell data (pre- and post-processing, clever graph initiatilization.
Currently ElPiGraph is used in two cell trajectory inference packages : **STREAM** and **MERLoT**

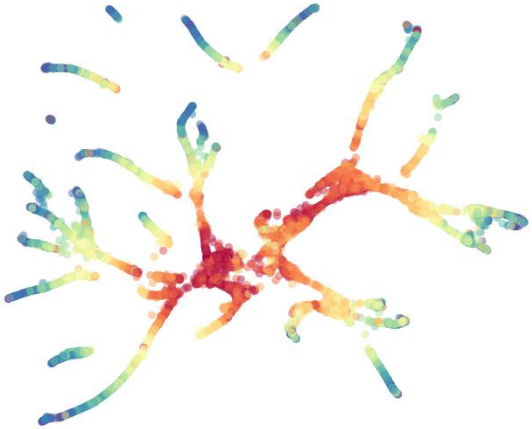
Scanpy PAGA

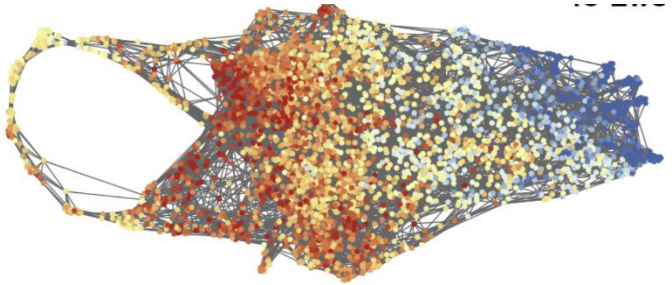


Common problems with most of trajectory inference methods

- What to do when local intrinsic dimension is >1, then the problem $R^N \rightarrow R^1$ is ill-posed
- Sensitivity to outliers
- Gaps in the data, ‘disconnected data manifold’
- Not applicability of trajectory concept (when the system is in quasi steady-state)
- Topologies more complex than a tree

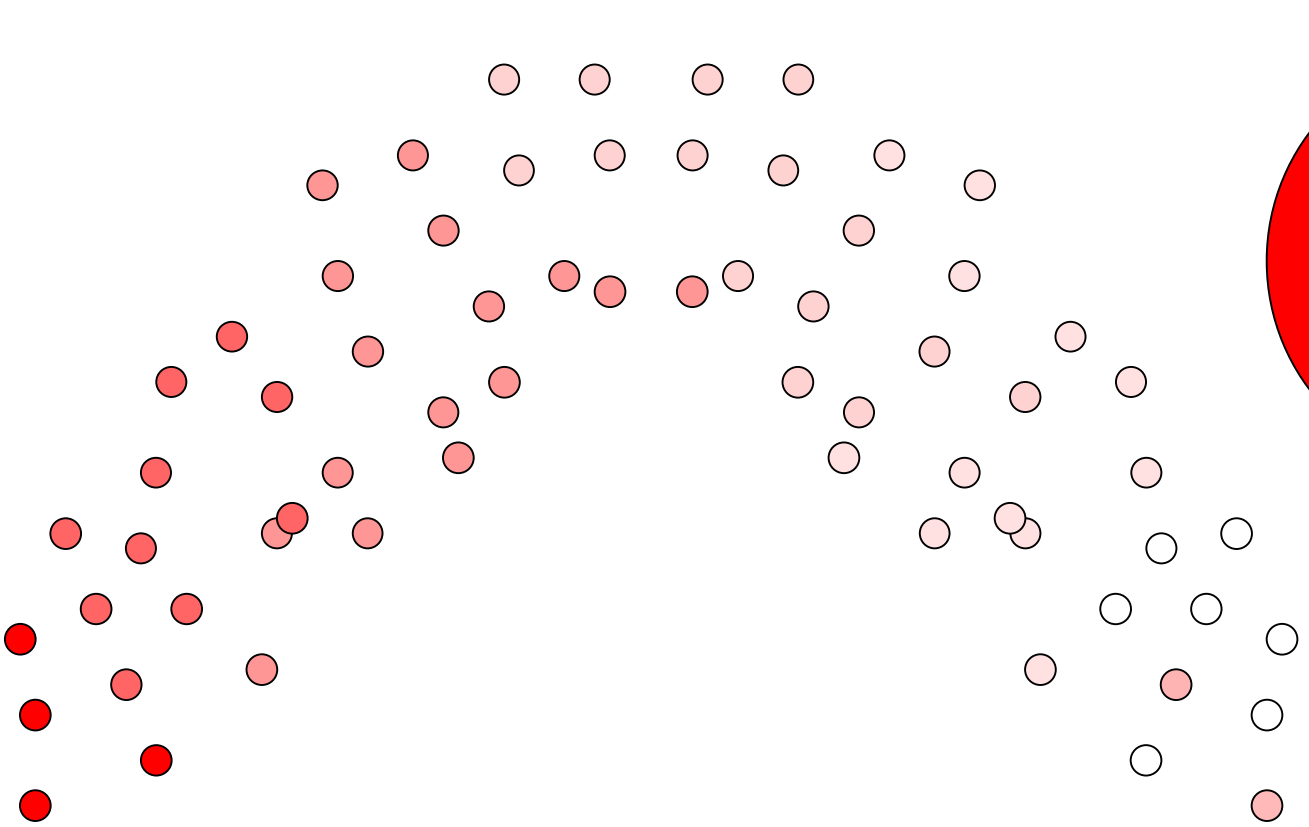


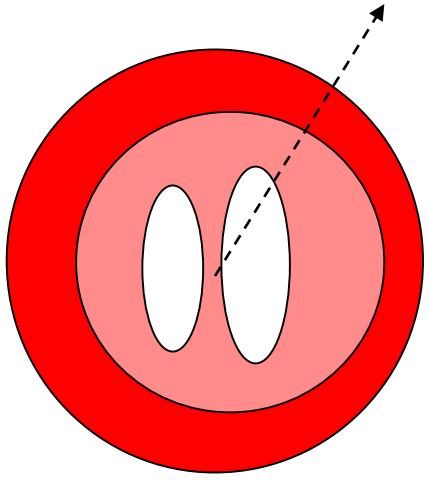




More fundamental problems… non-temporal heterogeneity

- The trajectory can represent rather spatial transcriptomic pattern than temporal

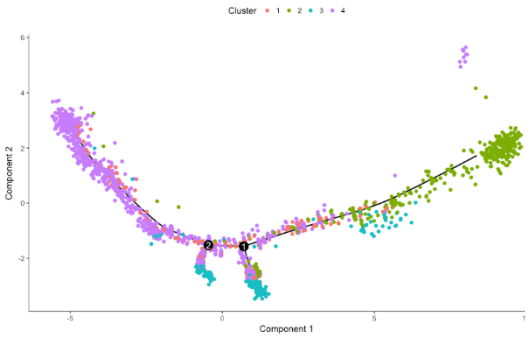


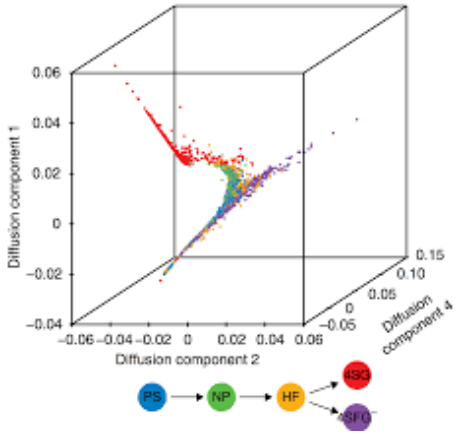


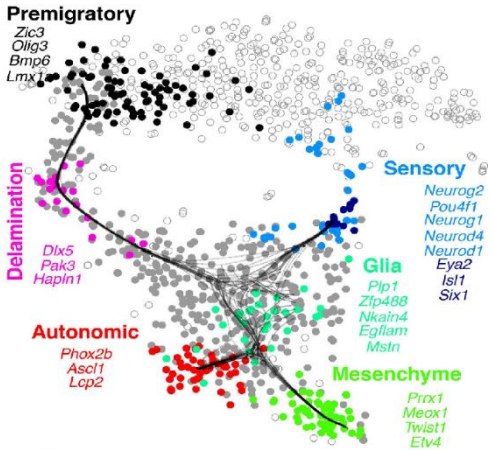
Complex tissue or organ

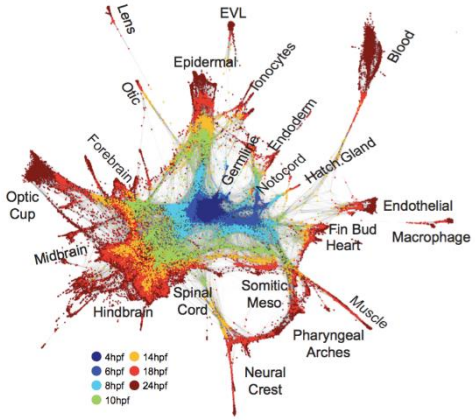
Some usual dirty tricks

- Preprocessing the data such that it becomes ‘more 1D’ locally (LLE, Diffusion maps, Force-directed layout) and ‘tree-like’
- Connecting cells only in sequential data points
- Visualizing data points close to tree edges such that the layout looks ‘tree-like’
- Drastically reduce global dimension (e.g., R^N->R^2)
- Bootstrap and resampling



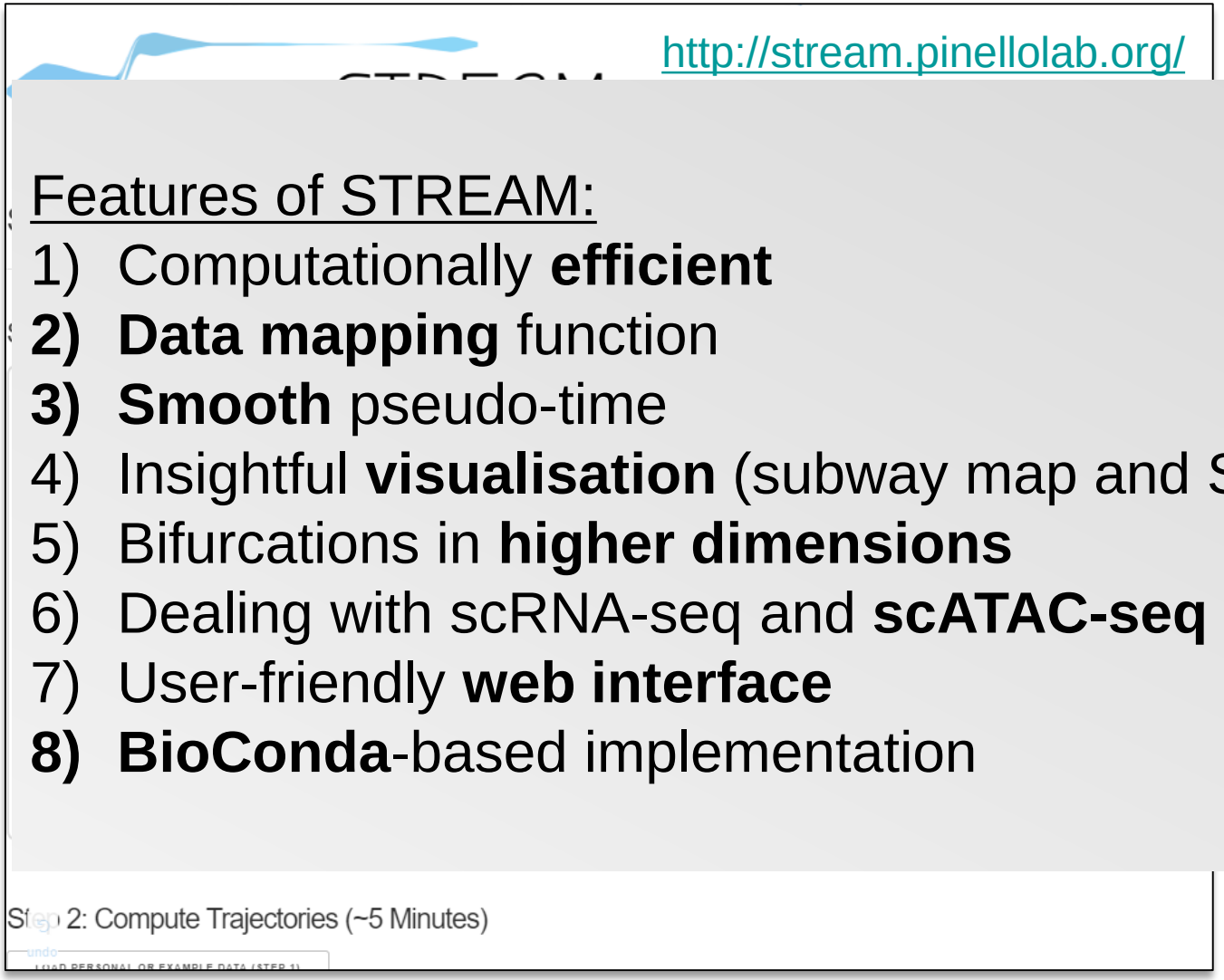






STREAM: Single-cell Trajectory REconstruction And Mapping

Chen H. et al, Nat Comm, 2019



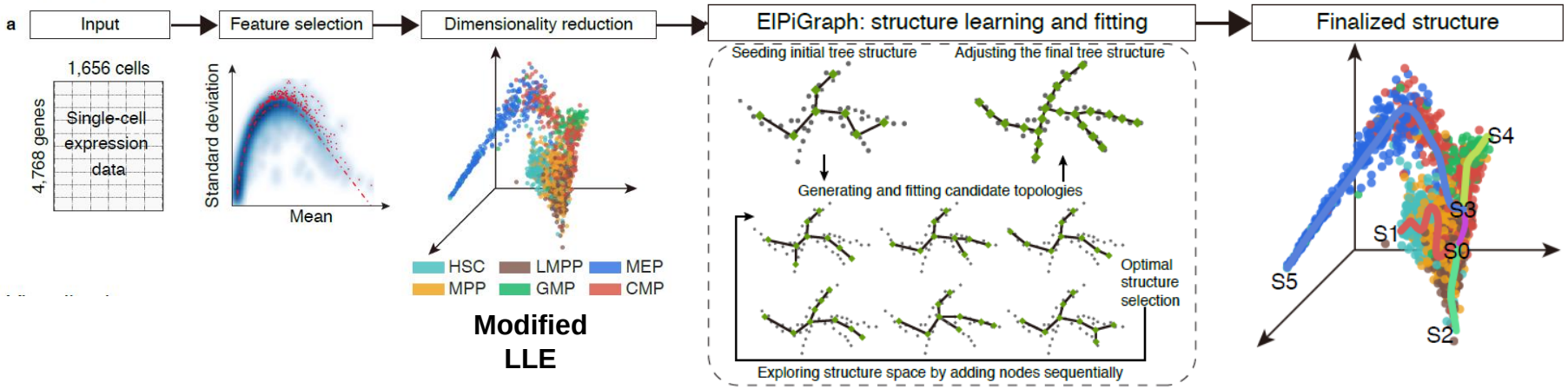
Features of STREAM:

1) Computationally **efficient**
2) **Data mapping** function
3) **Smooth** pseudo-time
4) Insightful **visualisation** (subway map and STREAM plot)
5) Bifurcations in **higher dimensions**
6) Dealing with scRNA-seq and **scATAC-seq** data
7) User-friendly **web interface**
8) **BioConda**-based implementation

Luca Albergante
INSERM U900,
Institut Curie

STREAM

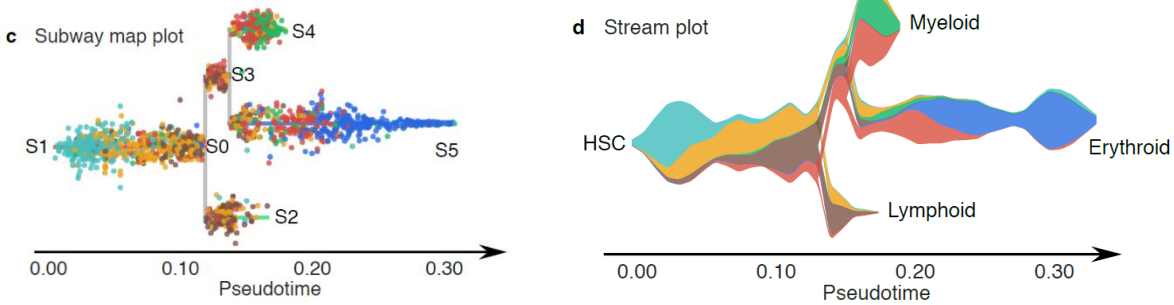
Trajectory inference

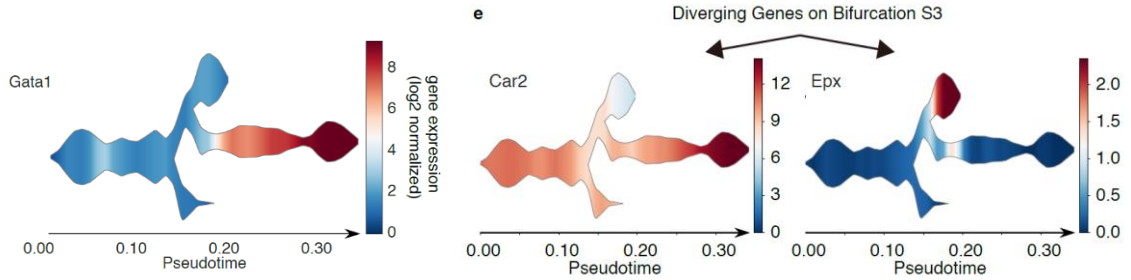


Visualization

scattering

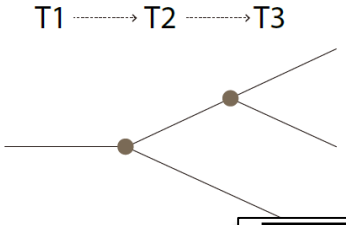
density



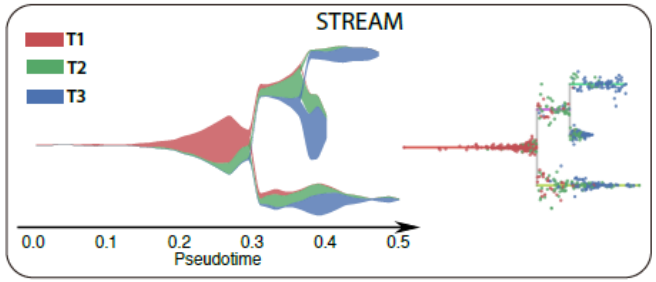


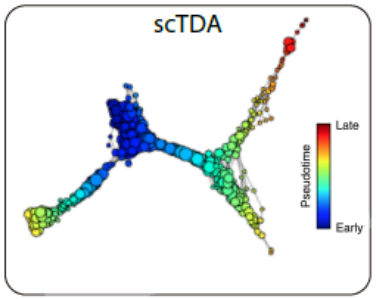
Internal benchmarking on synthetic data

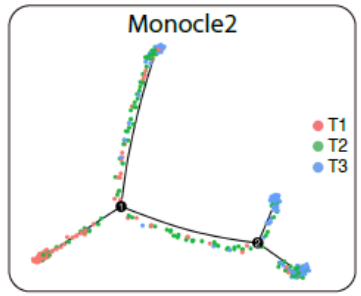
Simple topology

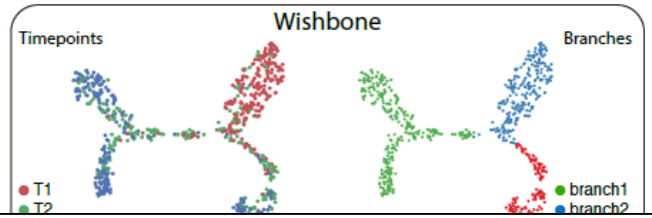


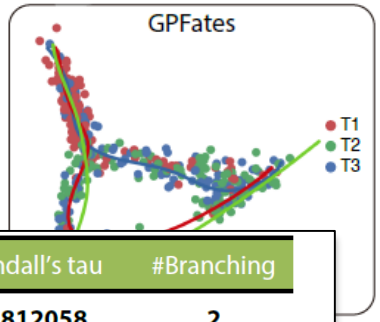
Rivzi et al, Nat Biotech,

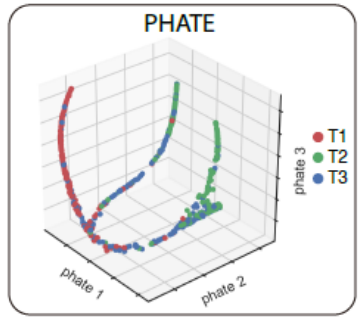




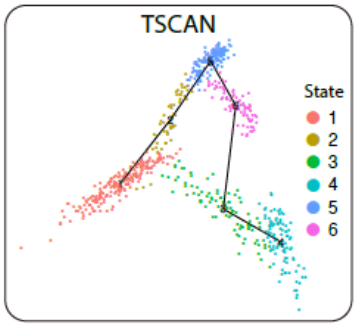


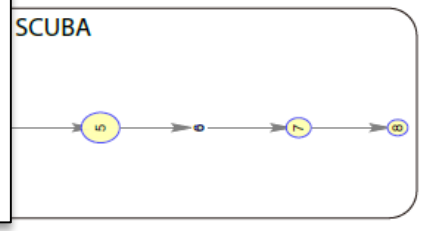




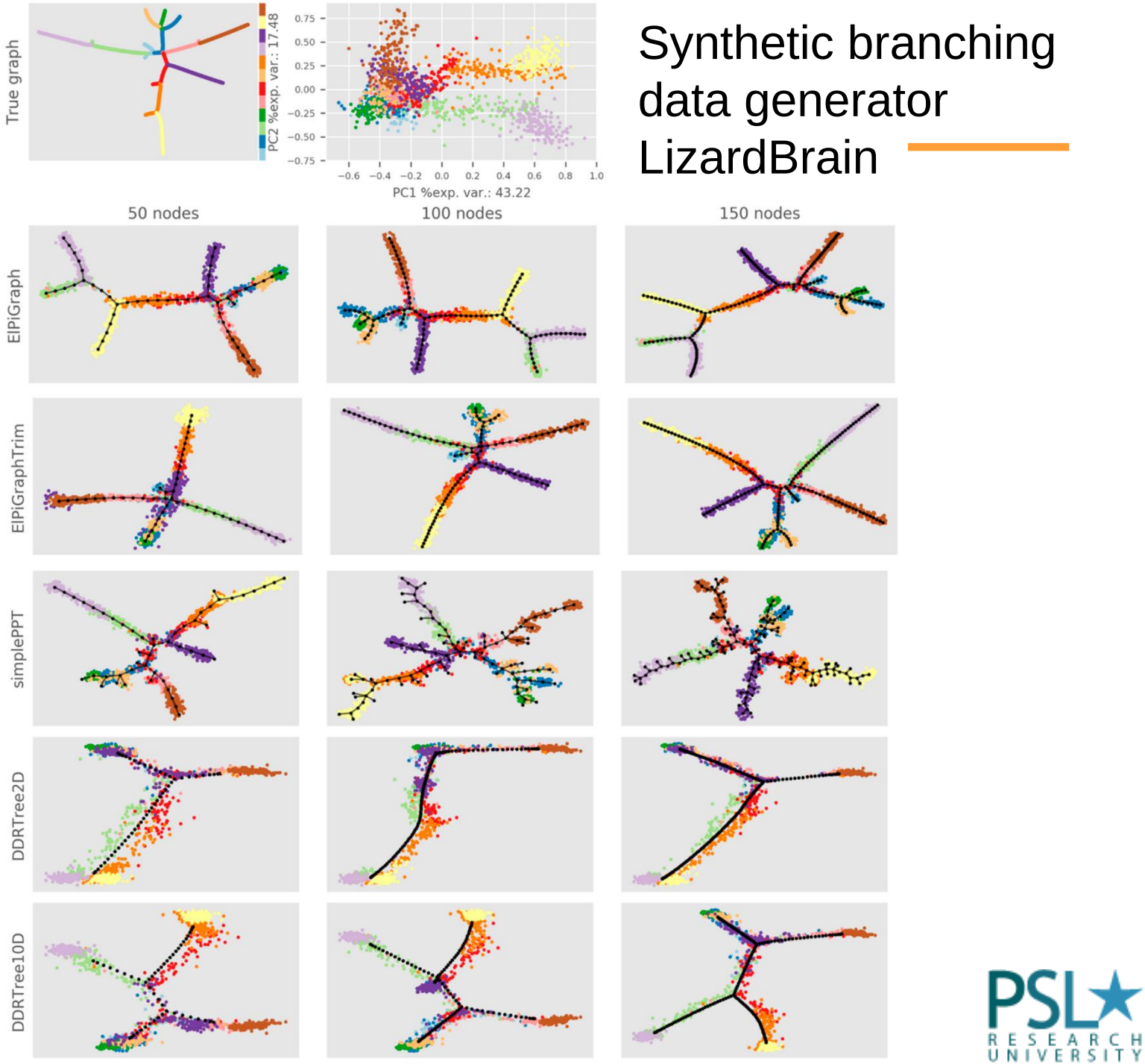


Method	Pearson(rank)	Pearson(distance)	Spearman	Kendall's tau	#Branching
STREAM	**0.950132**	**0.930923**	**0.944019**	**0.812058**	**2**
scTDA	**0.948441**	NA	**0.959412**	**0.823502**	**2**
Wishbone	0.931072	0.859669	0.922716	0.783160	1
SLICER	0.925962	0.860817	0.917089	0.771535	**2**
Monocle2	0.888202	**0.914076**	0.873133	0.727129	**2**
DPT	0.856093	0.858829	0.841390	0.671817	1
TSCAN	0.781034	NA	0.767569	0.618827	0
SCUBA	0.24301	0.150043	0.240983	0.158902	0
Mpath	NA	NA	NA	NA	**2**
GPfates	NA	NA	NA	NA	**2**
PHATE	NA	NA	NA	NA	**2**





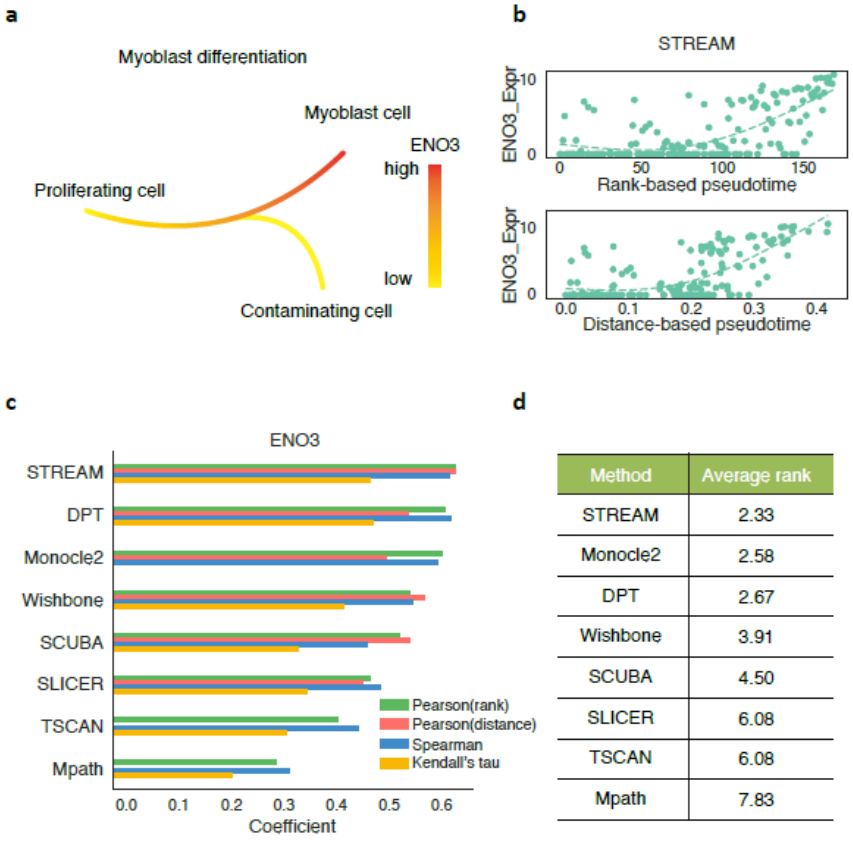
Synthetic branching data generator LizardBrain



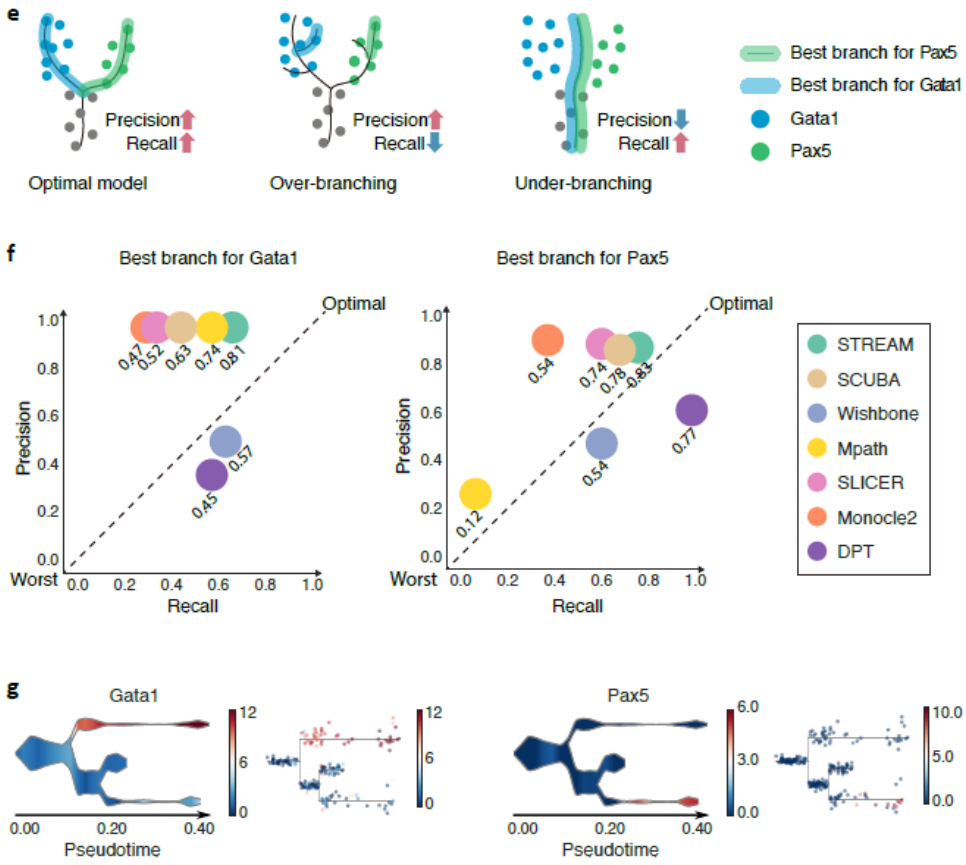
institutCurie

PSL RESEARCH UNIVERSITY

Internal benchmarking on real data:
smoothness of pseudo-time and accuracy of branches



Method	Average rank
STREAM	2.33
Monocle2	2.58
DPT	2.67
Wishbone	3.91
SCUBA	4.50
SLICER	6.08
TSCAN	6.08
Mpath	7.83



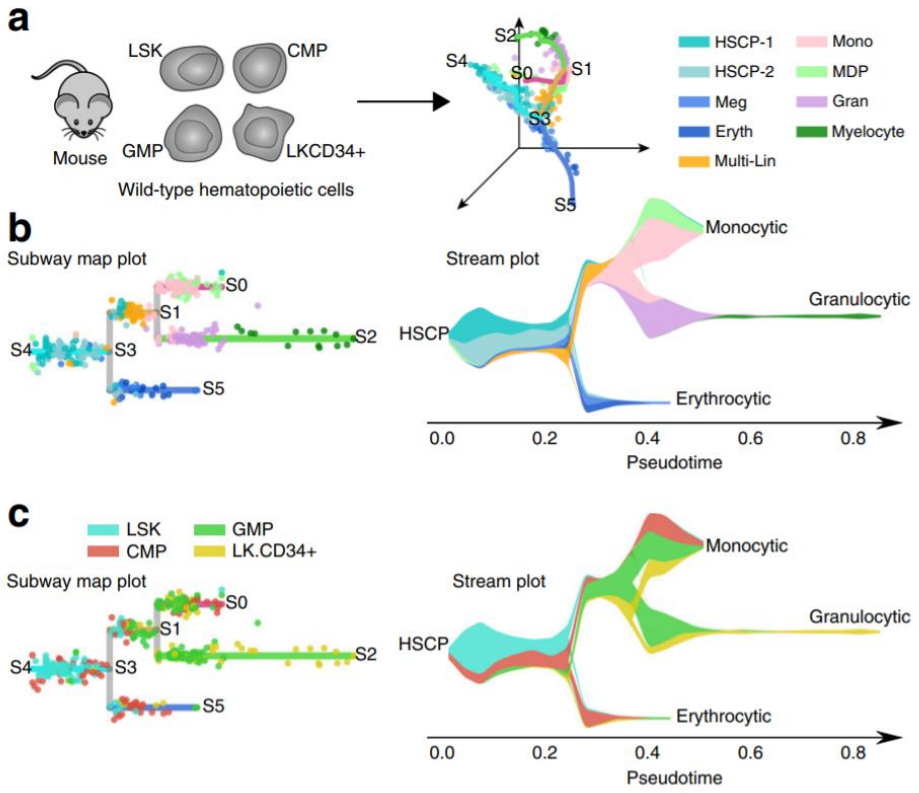
(data from Trapnell et al, Nat Biotech, 2014)

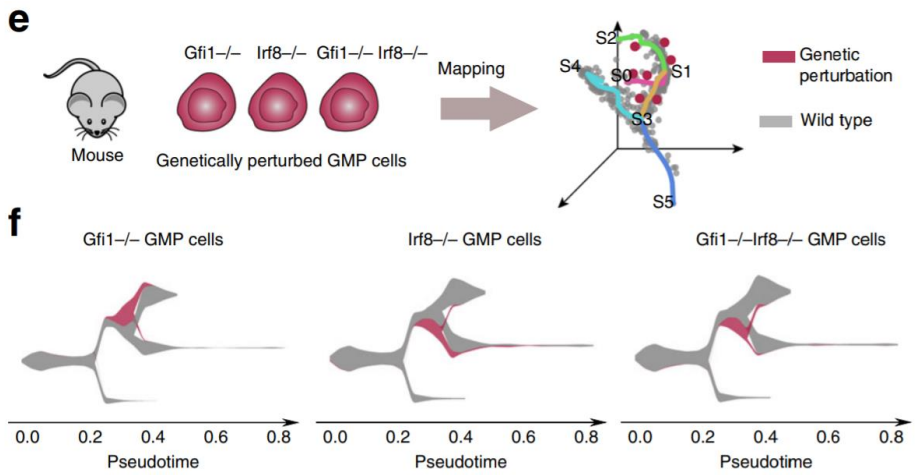
(data from Guo et al, Cancer Stem Cell, 2013)

STREAM is the only trajectory inference tool with explicit mapping function

Wild type (reference) analysis

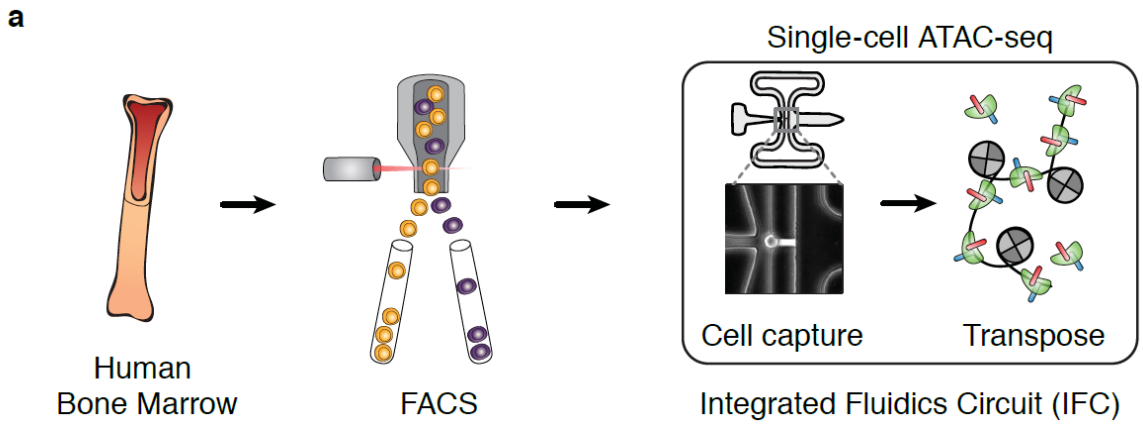
Genetically modified (new) data

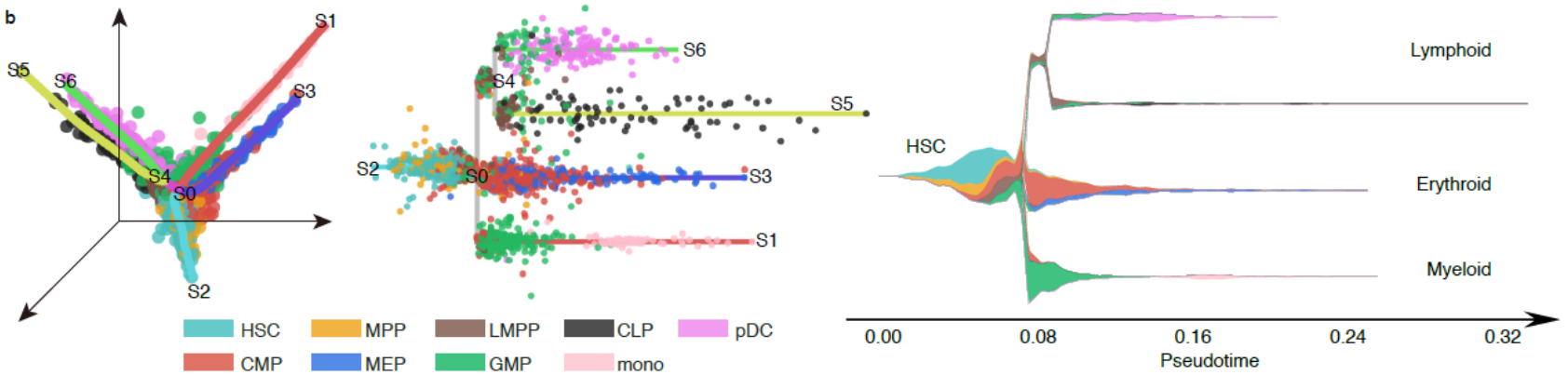




(data from Olsson et al, Nature, 2016)

STREAM is one of the first tools enabling trajectory inference from scATAC-seq data

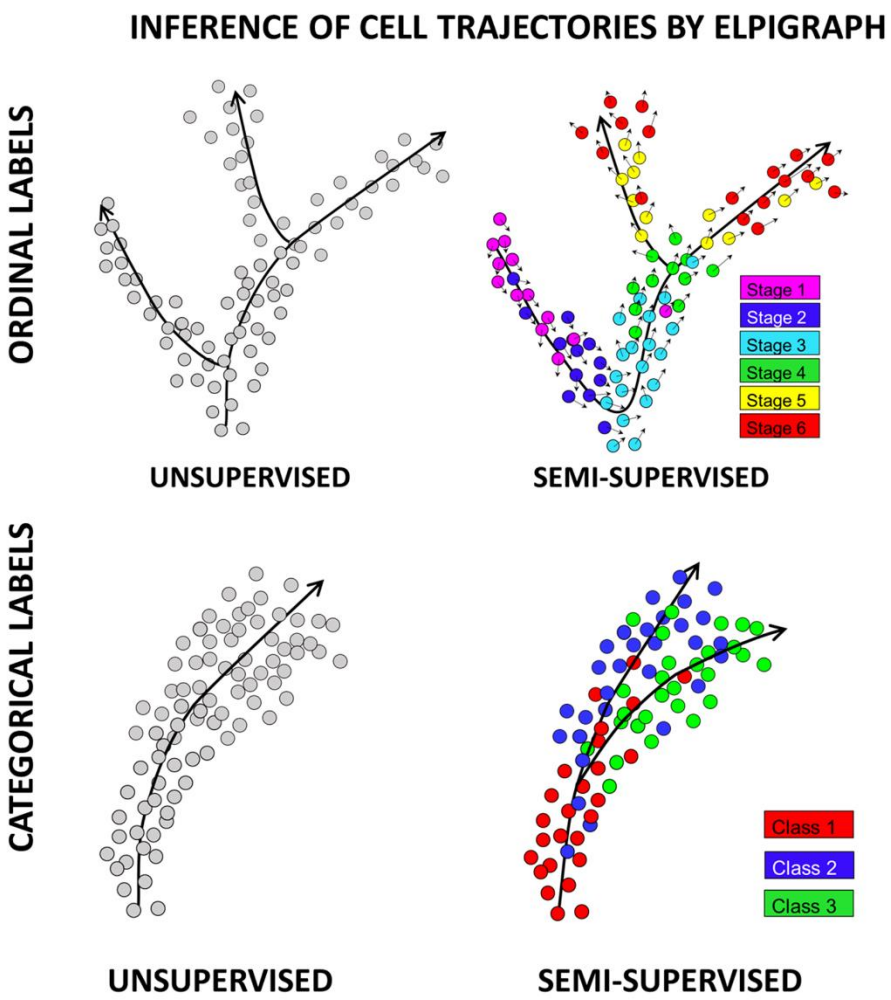




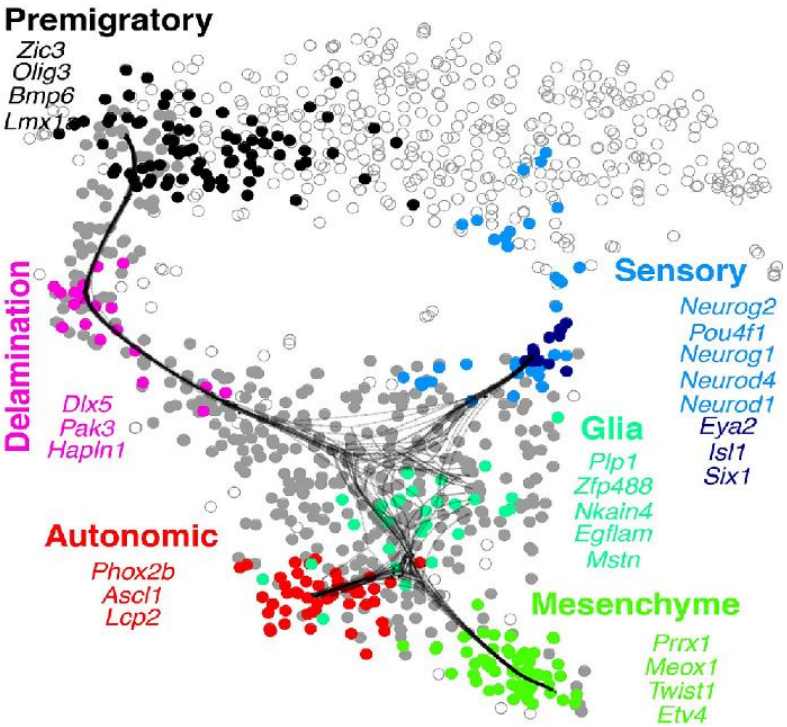
(data from Buenrostro et al, Cell, 2018)

Next steps for ElPiGraph and STREAM

Semi-supervised learning



Dealing with variable local intrinsic dimensionality

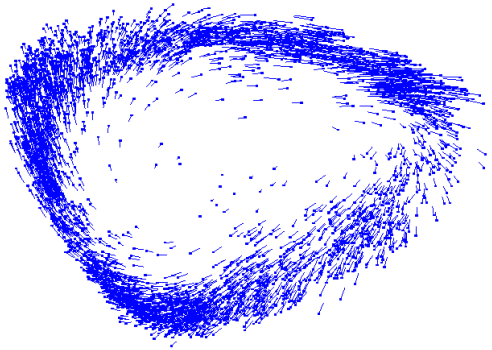


(from Soldatov et al, Science, 2019)

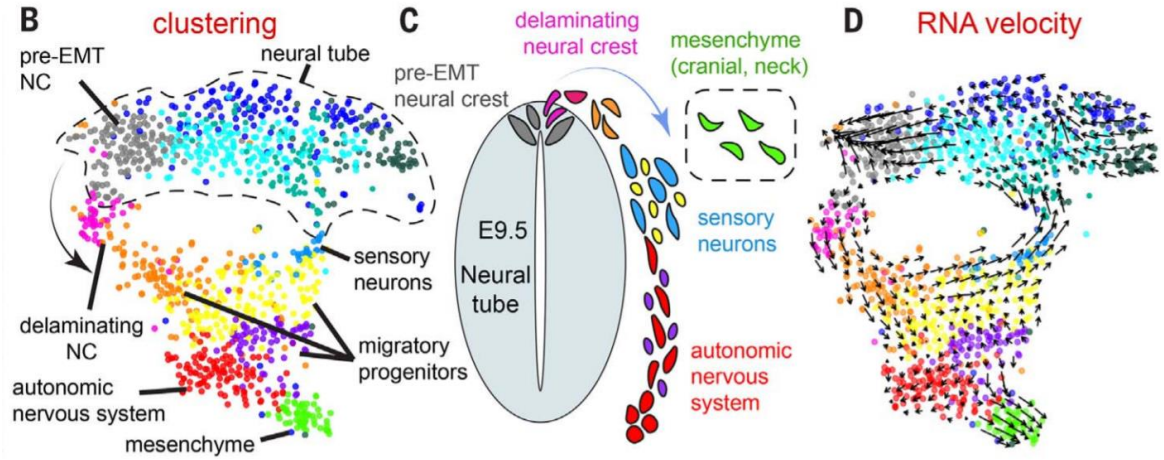
Instead of conclusion:
from “geometry” to “physics”

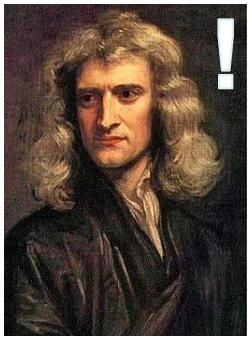
Learning cellular dynamics from cell snapshot data is an interesting but already yesterday’s idea

Instead of multi-dimensional data point clouds, we now have possibility to deal with more interesting objects: **multi-dimensional vector fields**



CHLA9 cell line, PCA projection





(from Soldatov et al, Science, 2019)

institutCurie

Instituts thématiques Inserm Institut national de la santé et de la recherche médicale

MINES ParisTech

PSL RESEARCH UNIVERSITY

Acknowledgements

ElPiGraph development

Luca Albergante
INSERM U900,
Institut Curie

Alexander Gorban
University
of Leicester
UK

Evgeny Mirkes
University
of Leicester
UK

Emmanuel Batillot
Institut Curie

Jonathan Bac
Centre de Recherche
Interdisciplinaire (CRI)

Louis Faure
Adameyko's lab
University of Vienna

STREAM development

Luca Pinello
Harvard Medical
School

Huidong Chen
Harvard Medical
School

Grants:

Chan Zuckerberg Initiative

National Human Genome Research Institute (NHGRI) Career Development Award

ITMO Cancer SysBio program (MOSAIC)

INCa PLBIO program (CALYS, INCA_11692)

Ministry of Education and Science of Russia (Project No. 14.Y26.31.0022)

institutCurie

Instituts thématiques Inserm Institut national de la santé et de la recherche médicale

MINES ParisTech

PSL RESEARCH UNIVERSITY