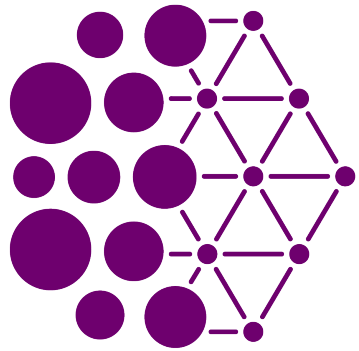
Geometry-based Data Exploration with Manifold Learning & Diffusion Geoemtry

Guy Wolf

Université de Montréal
Mila – Quebec AI Institute

guy.wolf@umontreal.ca

2020



Exploratory data analysis

One of the main challenges in modern data science is that:

- **Big high-dimensional data** are being produced everywhere
- **Limited numbers of domain scientists** have to process such data into useful knowledge

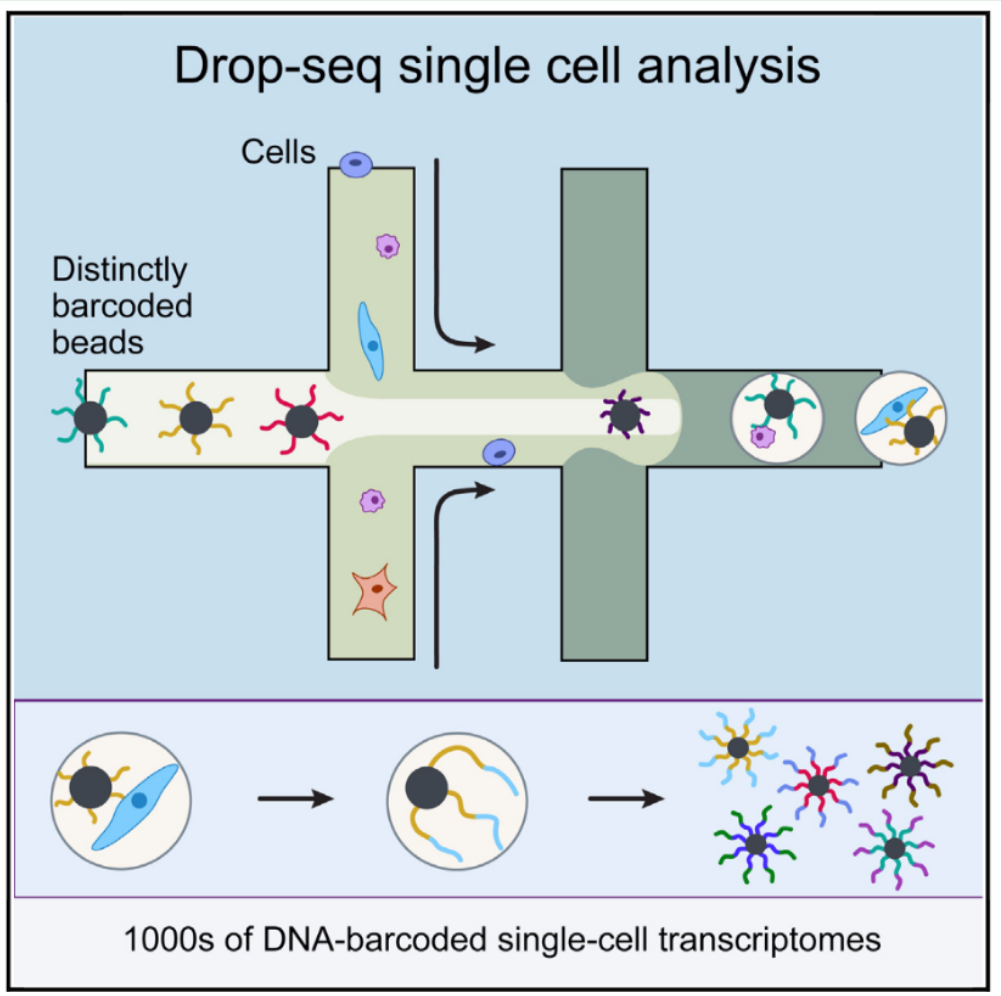
This challenge requires **exploratory data analysis** to produce human-interpretable data representations by

1. **Inferring structure** from collected data
2. Using this structure to **process data features** to become accessible for analysis

New frontier for data science & machine learning, beyond traditional predictive & generative tasks.

Exploratory data analysis

Example (high throughput single cell technologies)



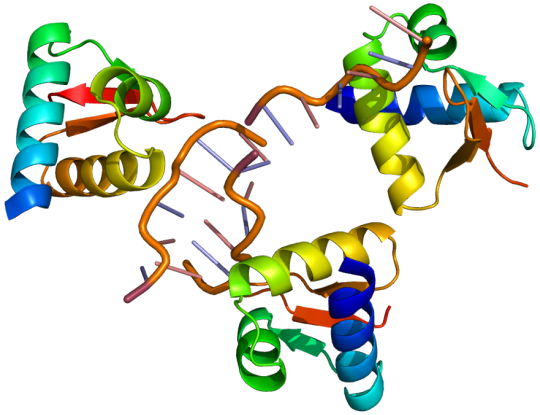
scRNA-seq: cells × genes
CyTOF: cells × proteins

- Big volumes of data
- High dimensional feature space
- Nontrivial noise & collection artifacts
- Multiresolution structures & processes
- Exploration often targets sparse data regions

Descriptive exploration in genomics & proteomics

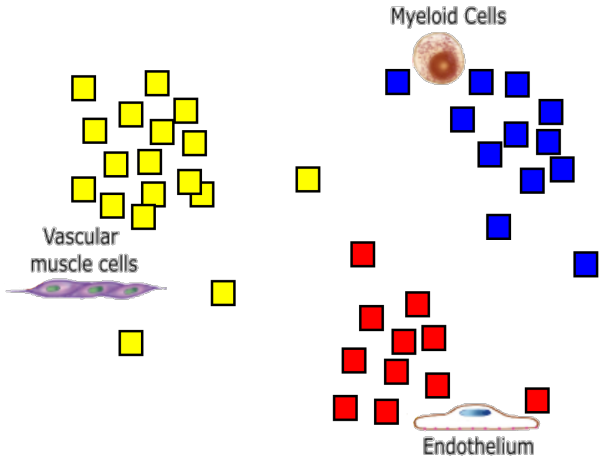
Single-cell data:

Gene counts

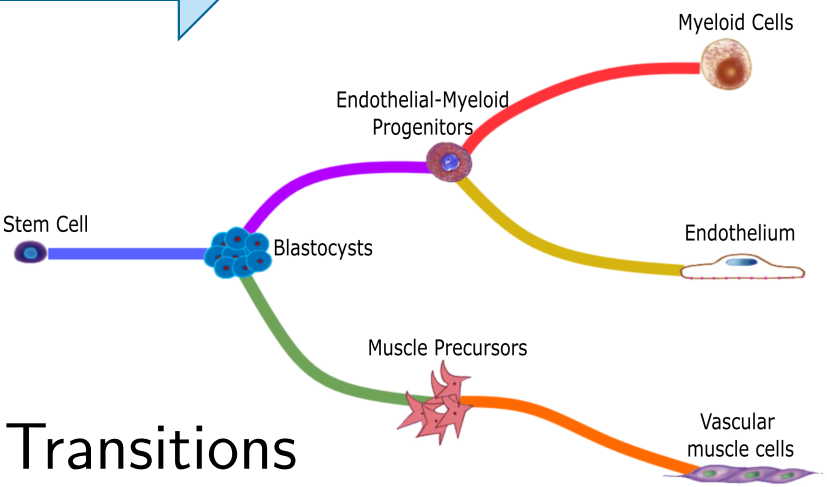


Protein counts

Clusters

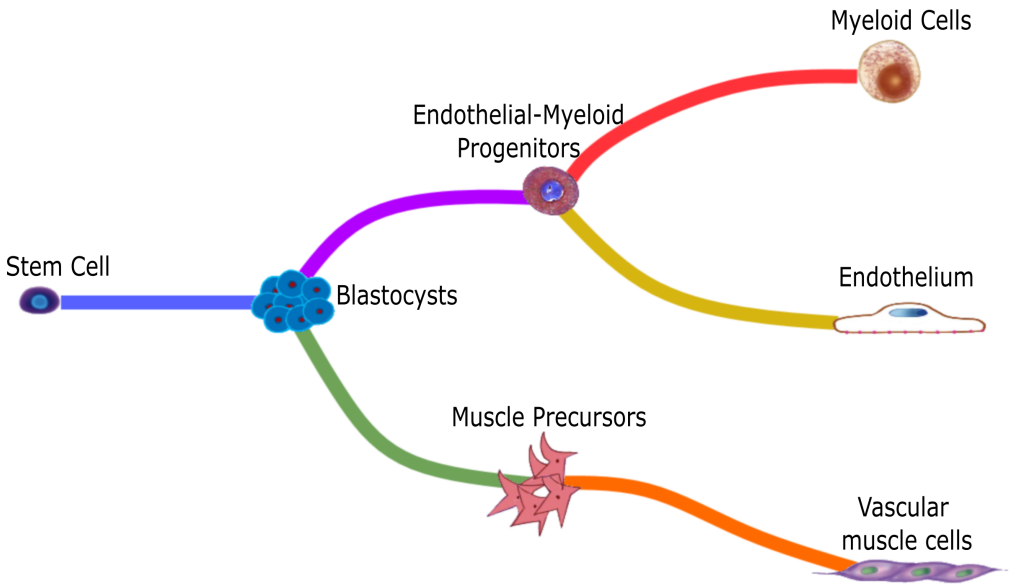


Transitions

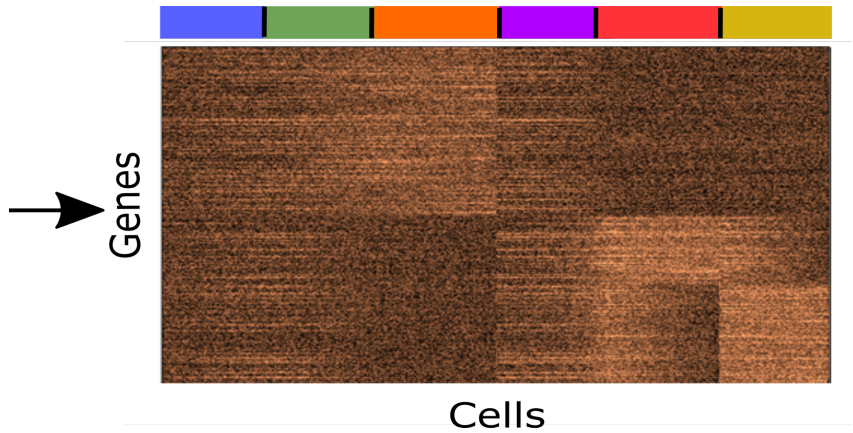


Visualizing progression & transitions in data

Progression & Transition Structures



High-dimensional Measurements

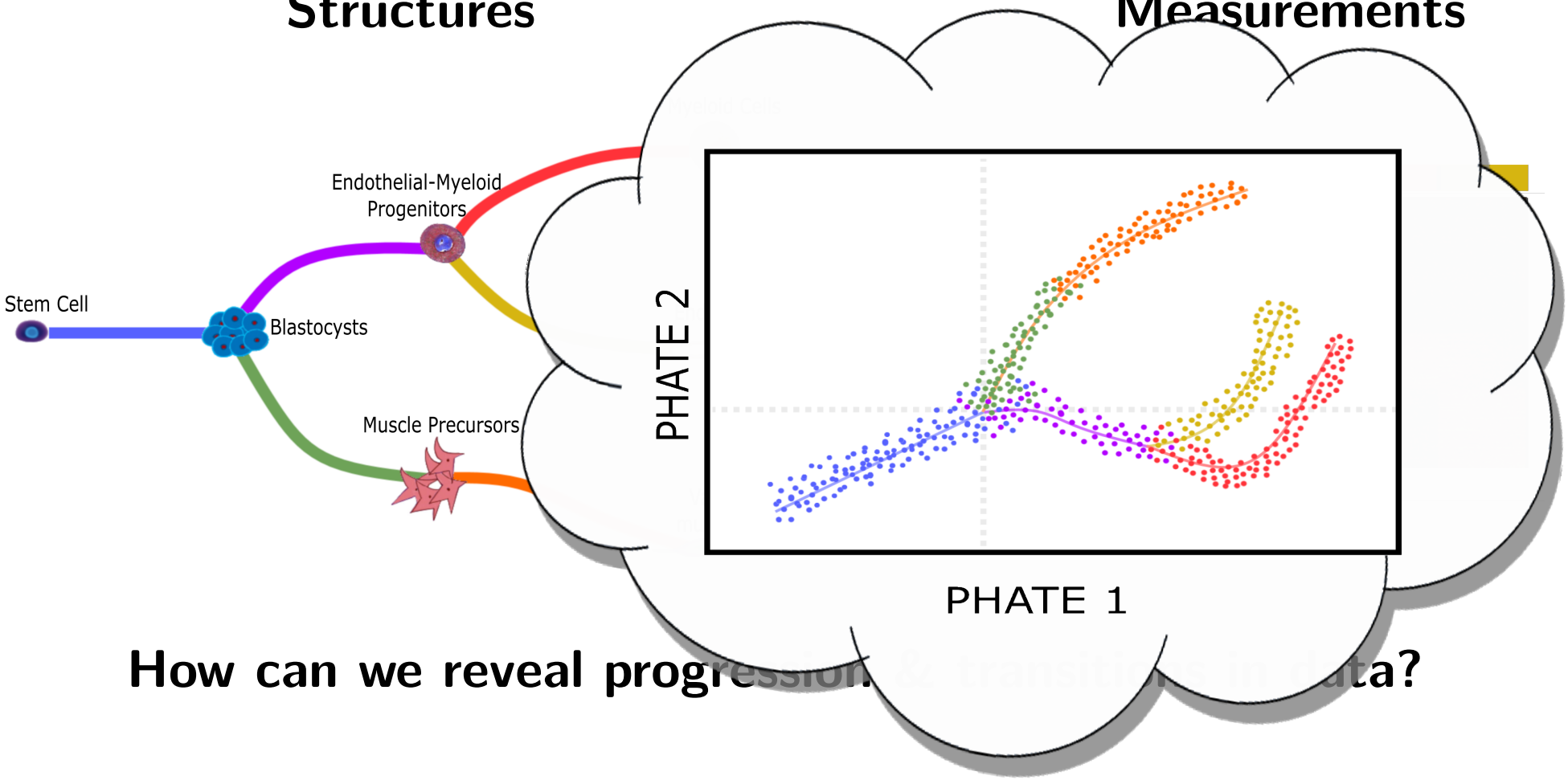


How can we reveal progression & transitions in data?

Visualizing progression & transitions in data

Progression & Transition Structures

High-dimensional Measurements



How can we reveal progression & transitions in data?

Manifold learning

Question: is cellular development really a high-dimensional process?

Consider the following key properties:

1. Cells develop progressively via **small incremental steps** (e.g., differentiation and mutation)
2. Variations in each step have **limited degrees of freedom**

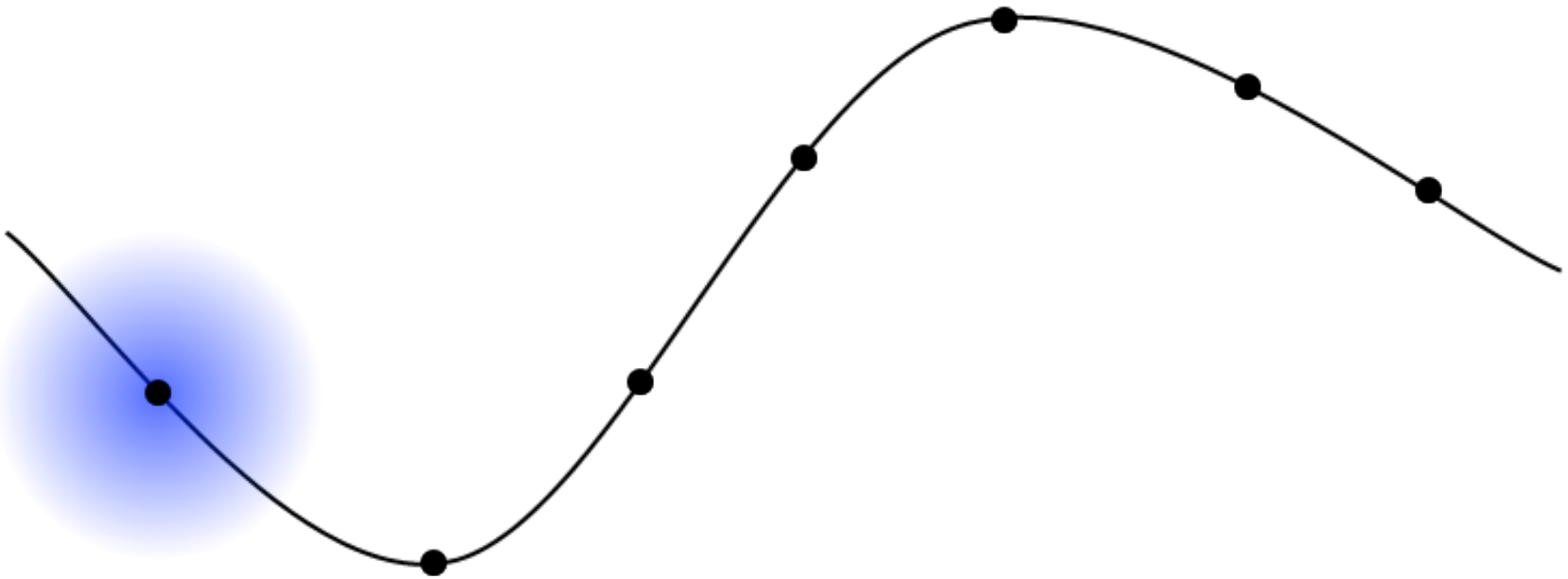
Conclusion: this progression can be modeled as a collection of **smoothly varying**, **locally low-dimensional**, **data patches**.

Such models are similar to the mathematical formulation of a manifold, and can be **inferred by manifold learning methods**.

More details in *"Manifold learning-based methods for analyzing single-cell RNA-sequencing data"* by K.R. Moon, J.S. Stanley, D. Burkhardt, D. van Dijk, G.W., and S. Krishnaswamy, *Current Opinion in Systems Biology*, 7:36–46, 2018.

Diffusion geometry

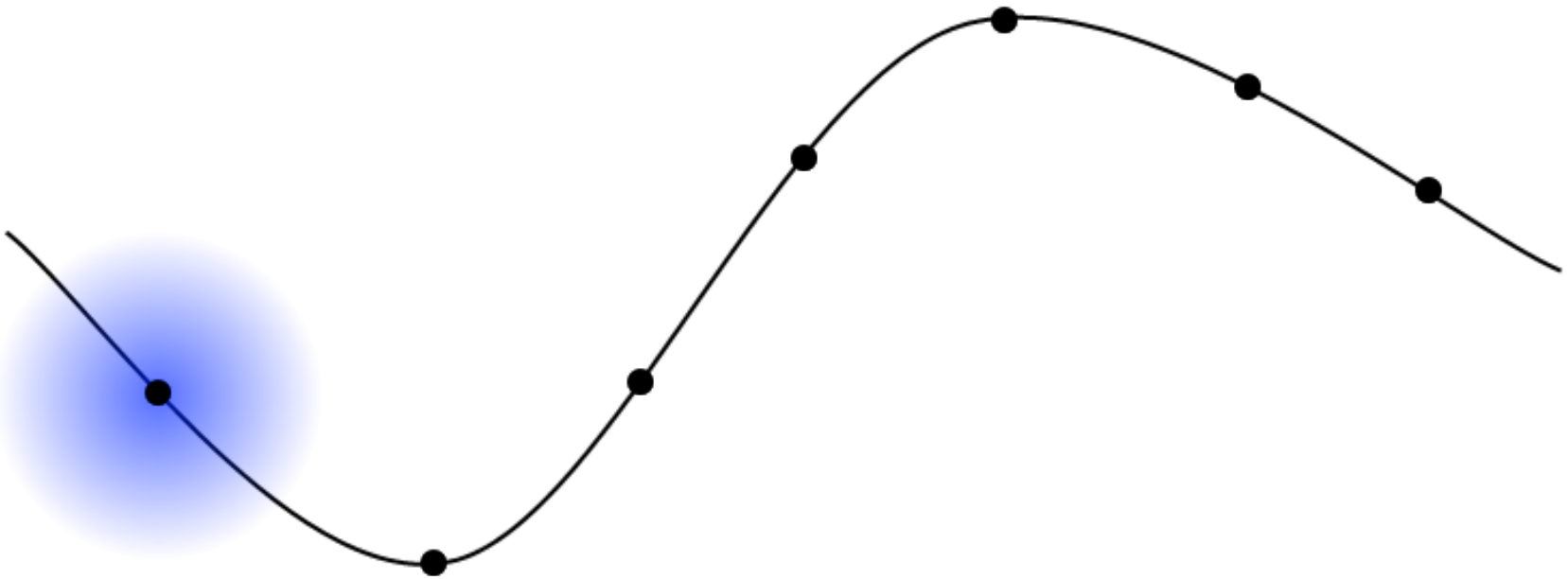
Manifold learning with random walks



- Local affinities $g(x,y) \Rightarrow$ transition probs. $\Pr[x \rightsquigarrow y] = \frac{g(x,y)}{\|g(x,\cdot)\|_1}$

Diffusion geometry

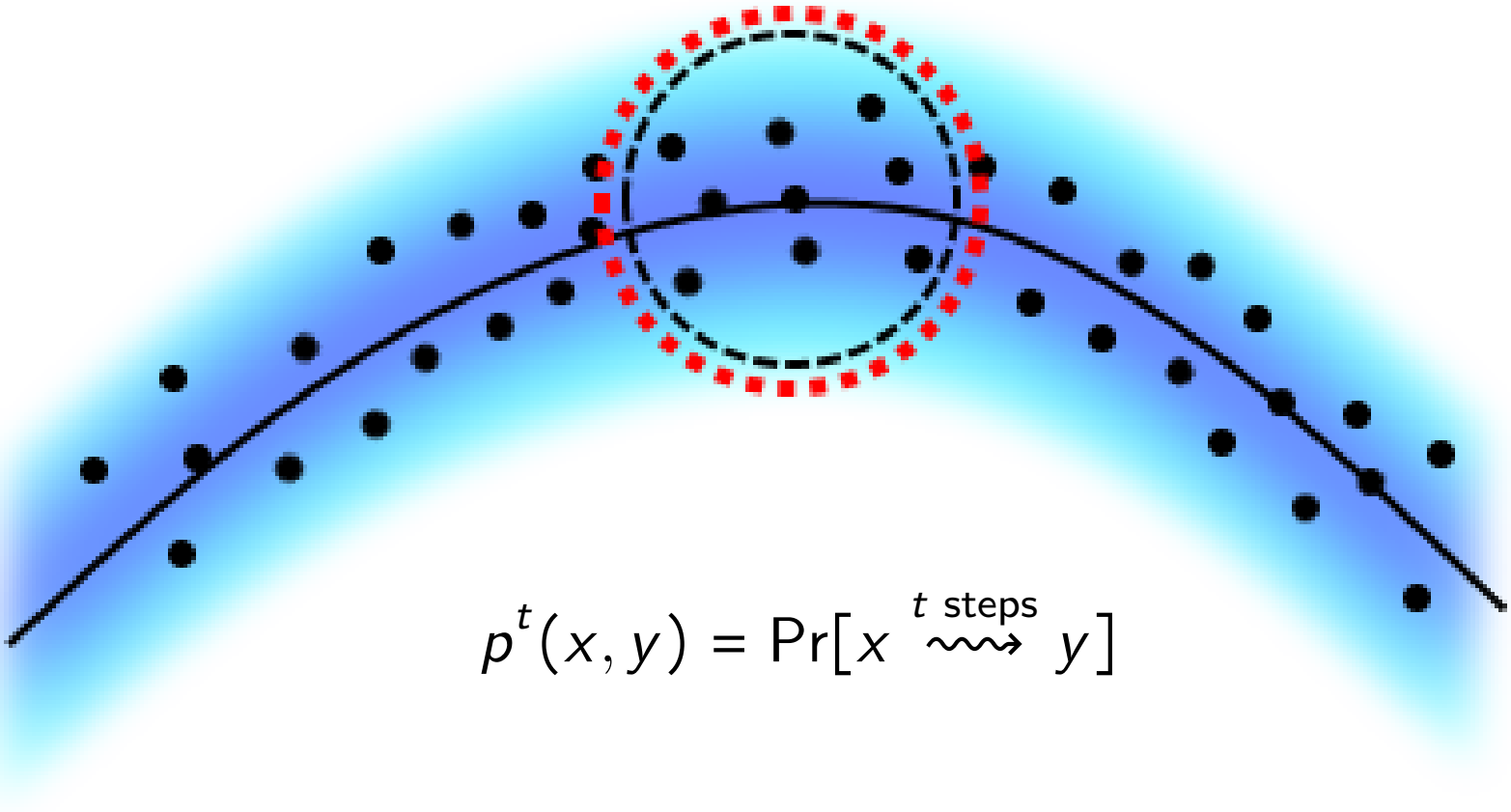
Manifold learning with random walks



- Local affinities $g(x,y) \Longrightarrow$ transition probs. $\Pr[x \rightsquigarrow y] = \frac{g(x,y)}{\|g(x,\cdot)\|_1}$
- Markov chain/process $\Longrightarrow$ random walks on data manifold

Diffusion geometry

Random walks reveal intrinsic neighborhoods

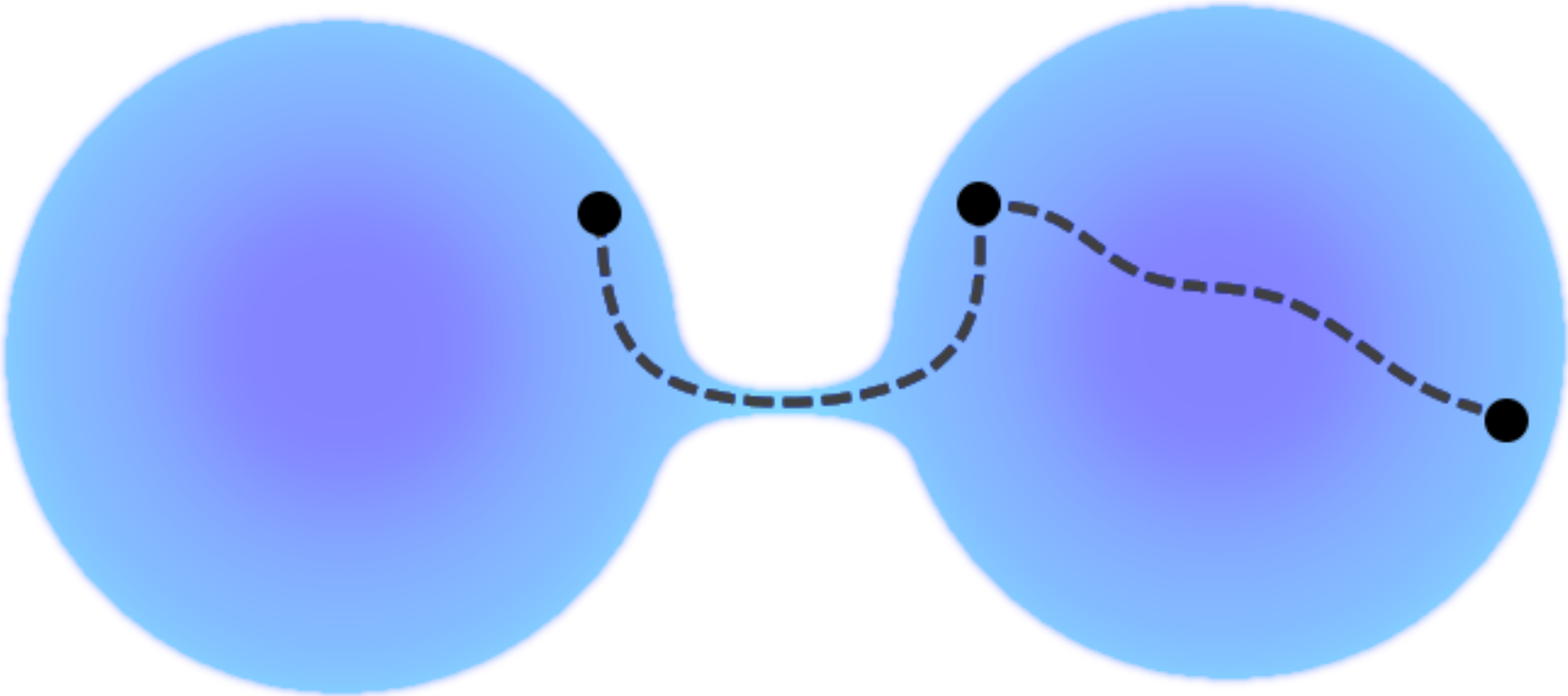


$$p^t(x,y) = \Pr[x \overset{t \text{ steps}}{\rightsquigarrow} y]$$

Diffusion geometry

Geodesic vs. diffusion distances

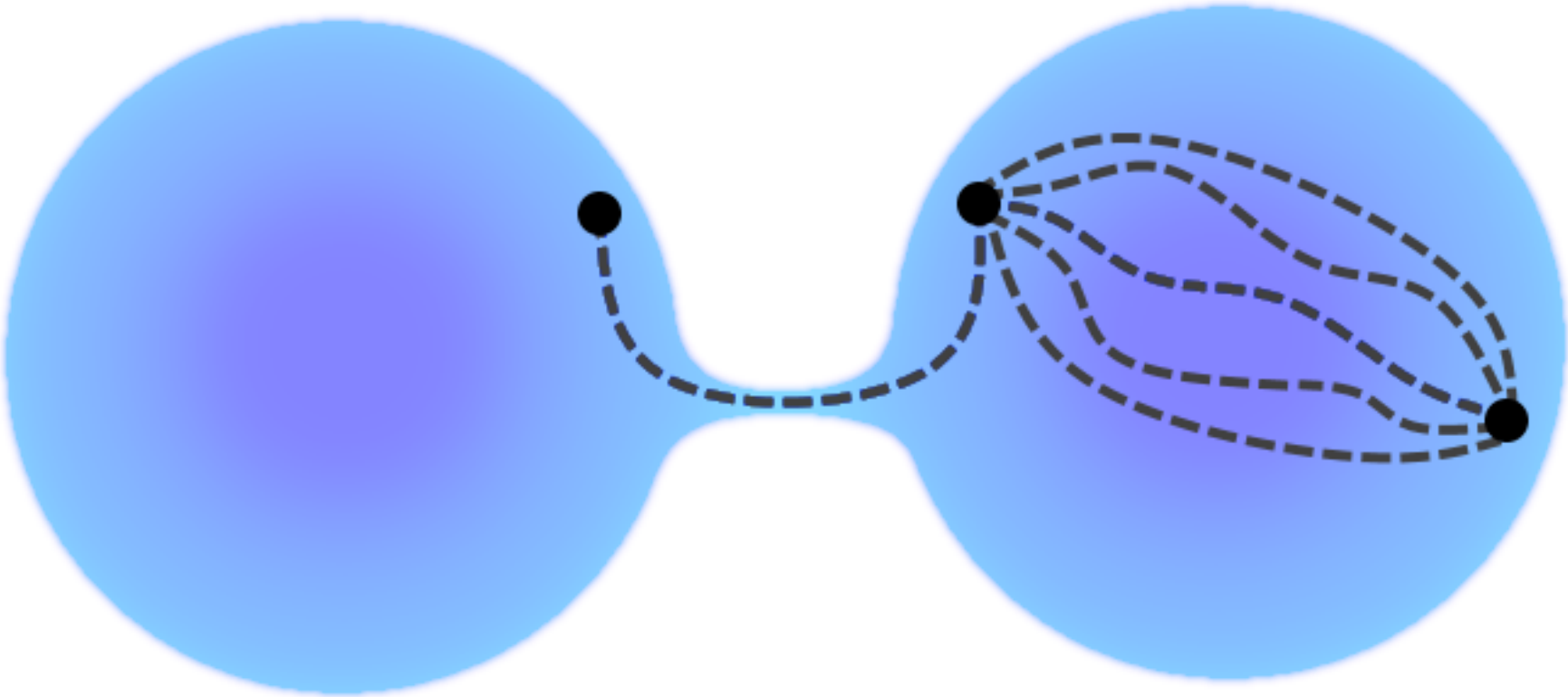
Are geodesic distances sufficient for faithful intrinsic embedding?



Diffusion geometry

Geodesic vs. diffusion distances

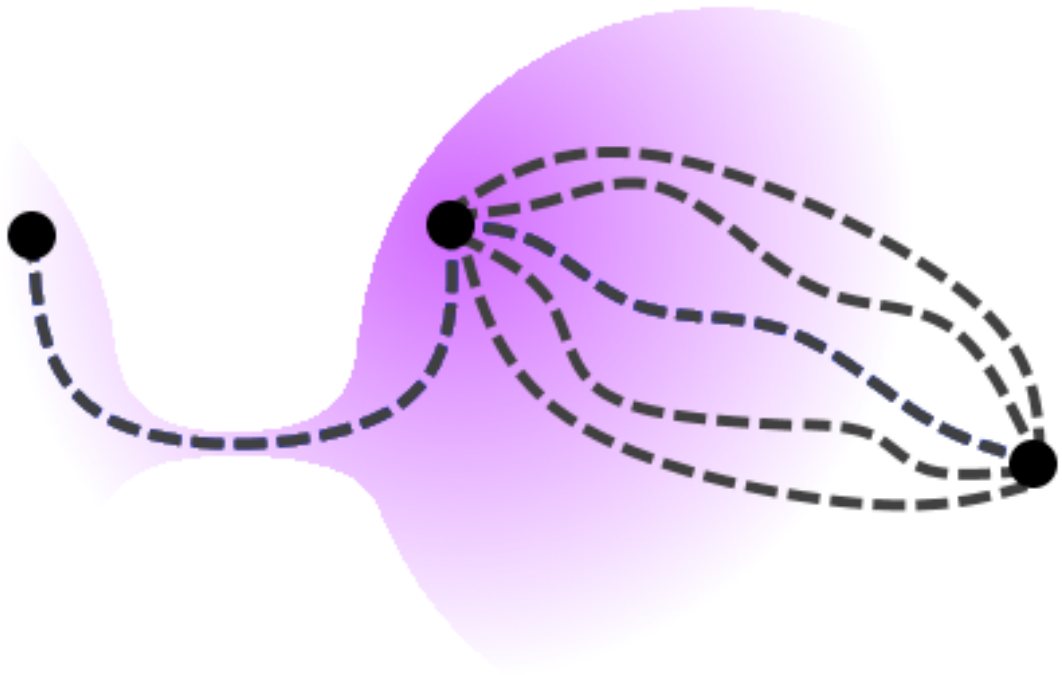
Are geodesic distances sufficient for faithful intrinsic embedding?



Diffusion geometry

Geodesic vs. diffusion distances

Are geodesic distances sufficient for faithful intrinsic embedding?

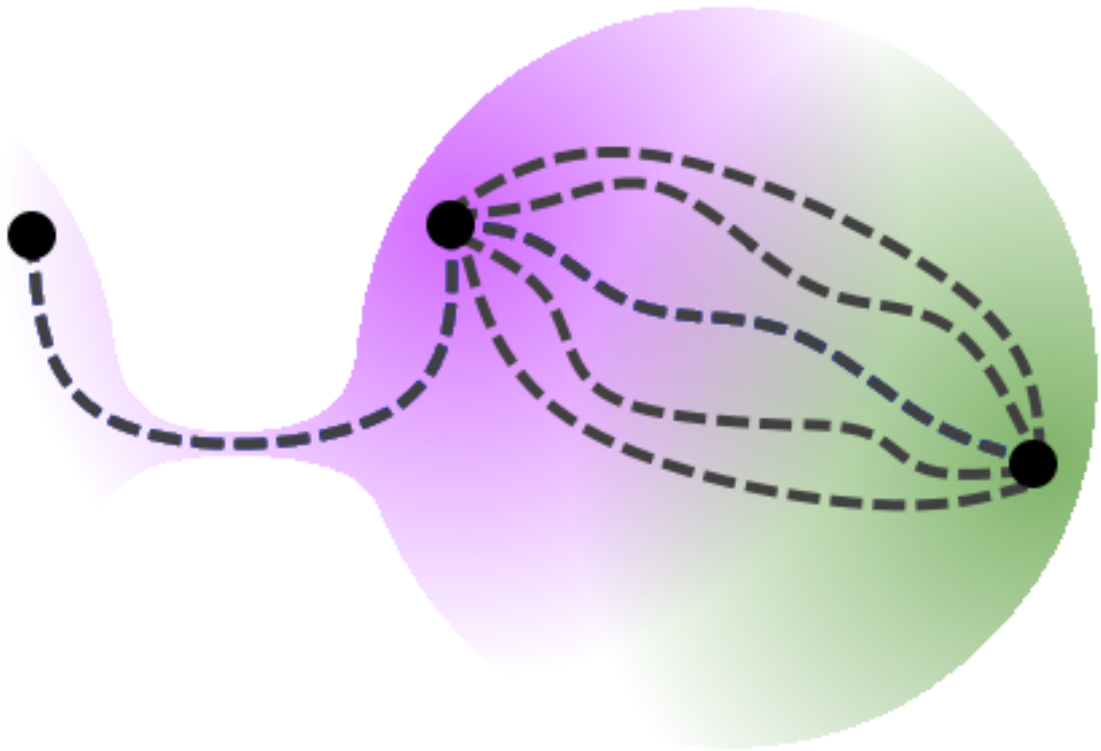


Diffusion-based notions enable robust intrinsic data geometry.

Diffusion geometry

Geodesic vs. diffusion distances

Are geodesic distances sufficient for faithful intrinsic embedding?

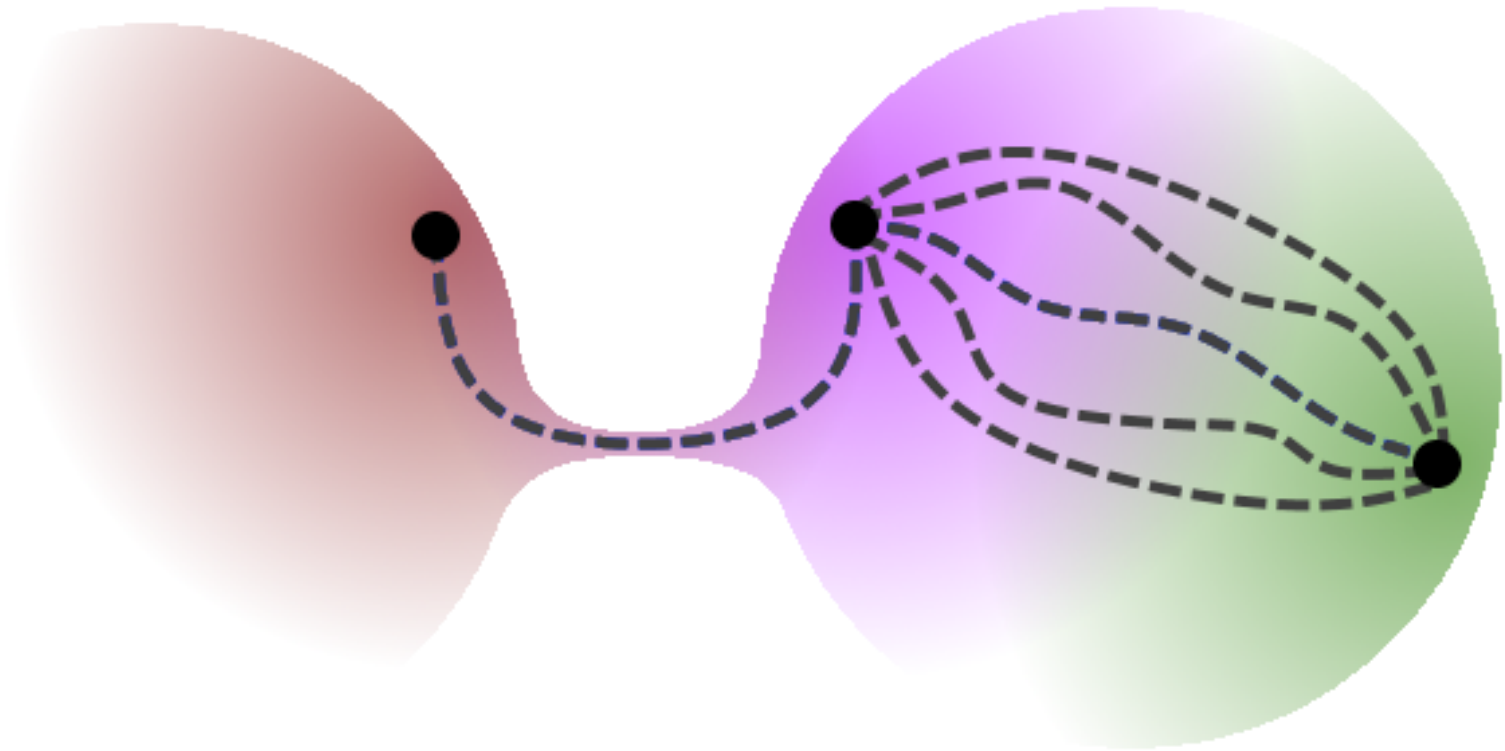


Diffusion-based notions enable robust intrinsic data geometry.

Diffusion geometry

Geodesic vs. diffusion distances

Are geodesic distances sufficient for faithful intrinsic embedding?



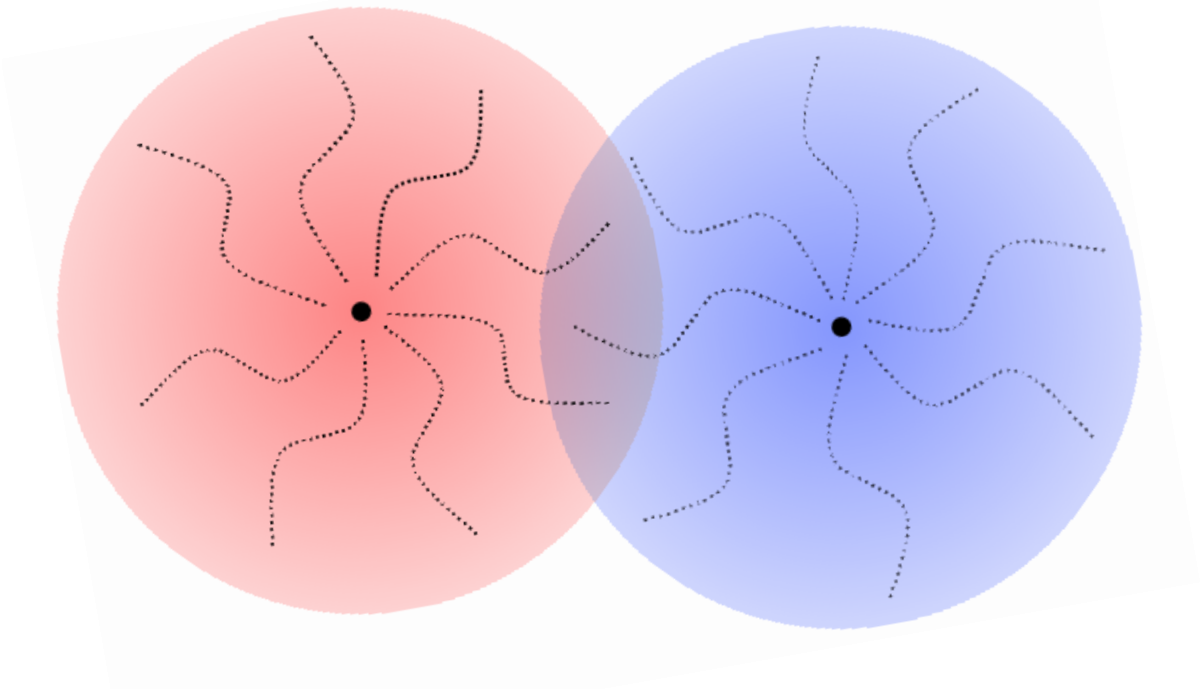
Diffusion-based notions enable robust intrinsic data geometry.

Diffusion geometry

Diffusion & potential distances

DM (Coifman & Lafon): $\overbrace{\left\|\Phi^t(x) - \Phi^t(y)\right\|}^{\text{embedded distance}} \approx \overbrace{\left\|P^t_{(x,\cdot)} - P^t_{(y,\cdot)}\right\|_{L^2(\|q\|_1/q)}}^{\text{diffusion distance}}$

PHATE (Moon et al.): $\underbrace{\left\|\Phi^t(x) - \Phi^t(y)\right\|}_{\text{embedded distance}} \approx \underbrace{\left\|\log P^t_{(x,\cdot)} - \log P^t_{(y,\cdot)}\right\|}_{\text{potential distance}}$



$g(u,v)$ = local affinity

$q(u) = \|g(u,\cdot)\|_1$

$P_{(u,v)} = g(u,v)/q(u)$

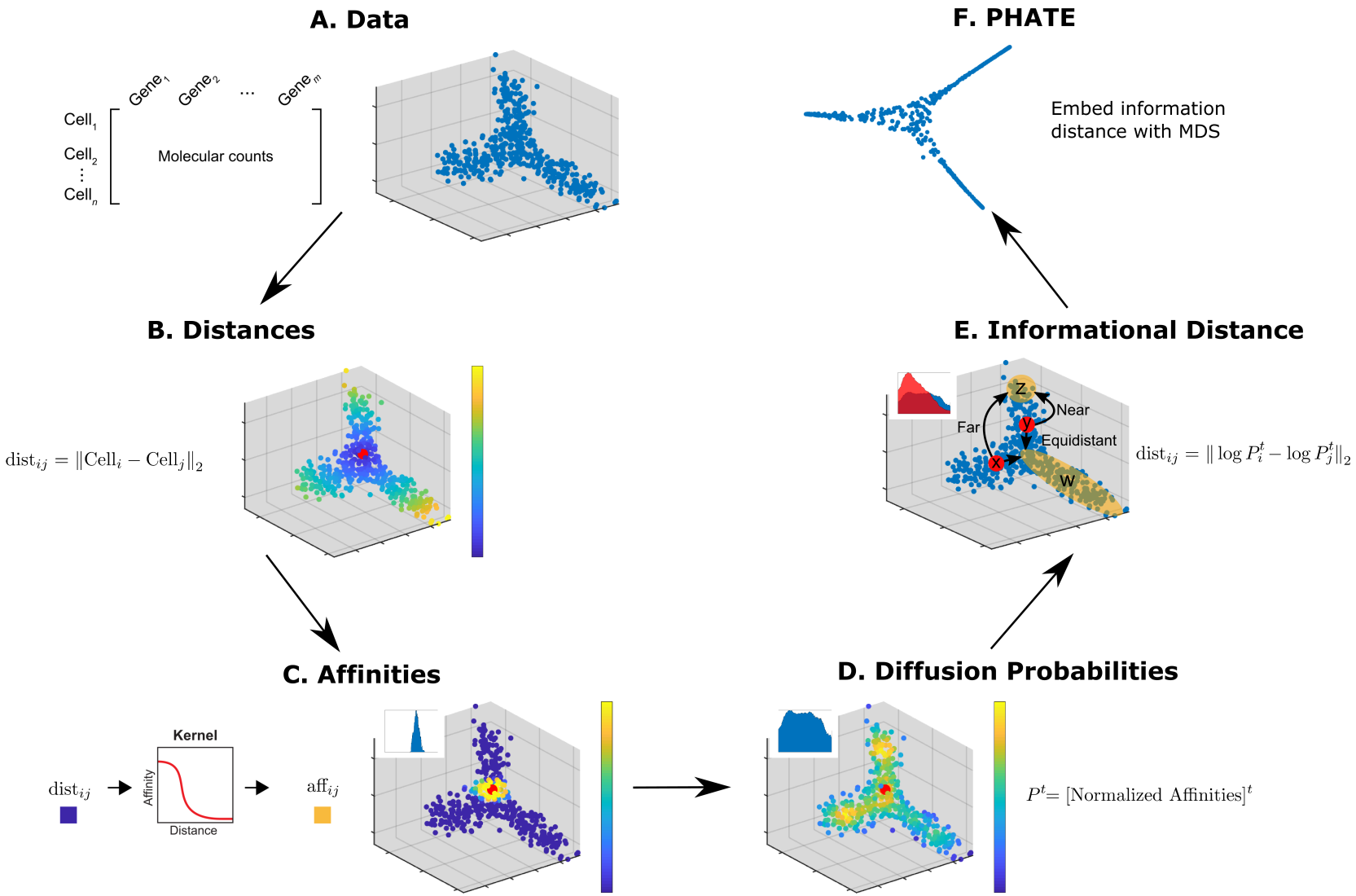
$P^t_{(u,v)} = \Pr[u \overset{t \text{ steps}}{\rightsquigarrow} v]$

Φ^t : data $\to \mathbb{R}^d$ (small d)

Data visualization

PHATE (Moon et al., Nat. Biotech. 2019)

Overview



A. Data

F. PHATE

Affinities via adaptive α-decaying kernel

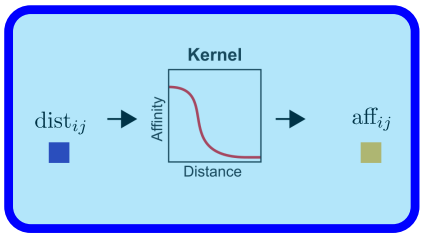
$$\tilde{g}(x,y) = \exp\left[-\left(\frac{\|x-y\|}{\varepsilon_x}\right)^{\alpha}\right] \mapsto g(x,y) = \frac{\tilde{g}(x,y)+\tilde{g}(y,x)}{2}$$

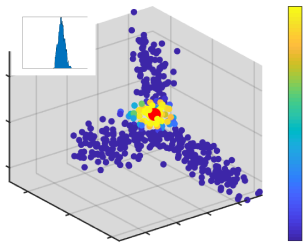
Where

- ε_x = **distance** from x to its ***k*-th nearest neighbor**
- α controls the **decay rate** of $\tilde{g}$ w.r.t $\|x-y\|/\varepsilon_x$

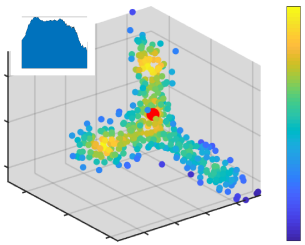
$\text{dist}_{ij} = \|\text{Cell}_i$

C. Affinities



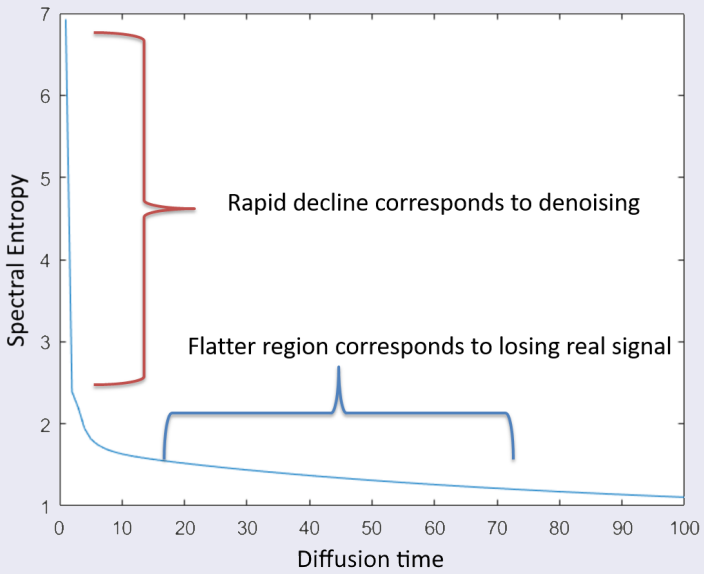


D. Diffusion Probabilities



P^t= [Normalized Affinities]^t

Diffusion time tuning with spectral entropy



Spectral entropy at time t:

$$-\sum_j \frac{\lambda_j^t}{\|\lambda^t\|_1} \log \frac{\lambda_j^t}{\|\lambda^t\|_1}$$

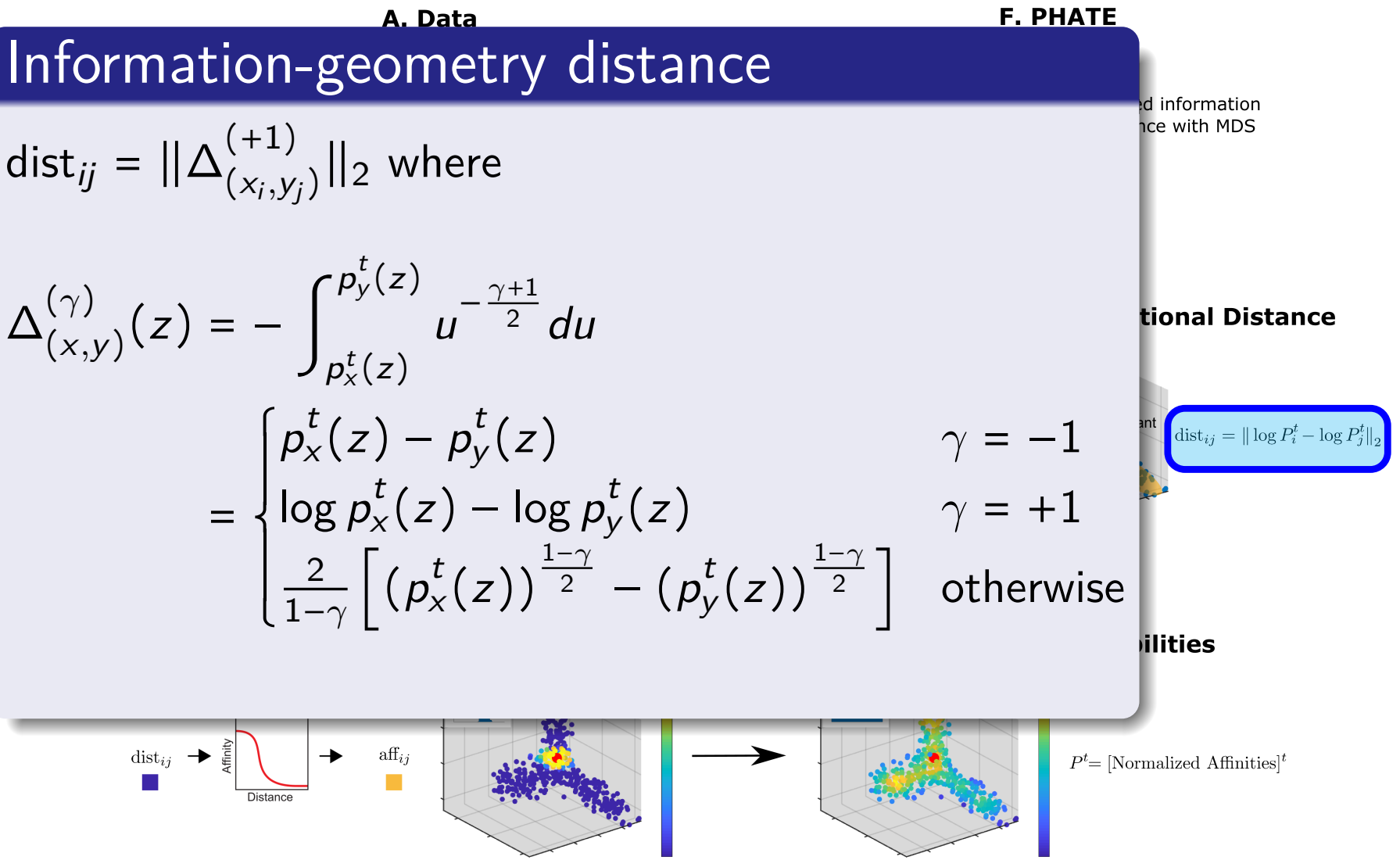
where $\boldsymbol{\lambda^t} = \{\lambda_0^t, \lambda_1^t, \lambda_2^t, \ldots\}$ are the eigenvalues of P^t.

Information-geometry distance

$$\text{dist}_{ij} = \|\Delta^{(+1)}_{(x_i,y_j)}\|_2 \text{ where}$$

$$\Delta^{(\gamma)}_{(x,y)}(z) = -\int_{p^t_x(z)}^{p^t_y(z)} u^{-\frac{\gamma+1}{2}}\, du$$

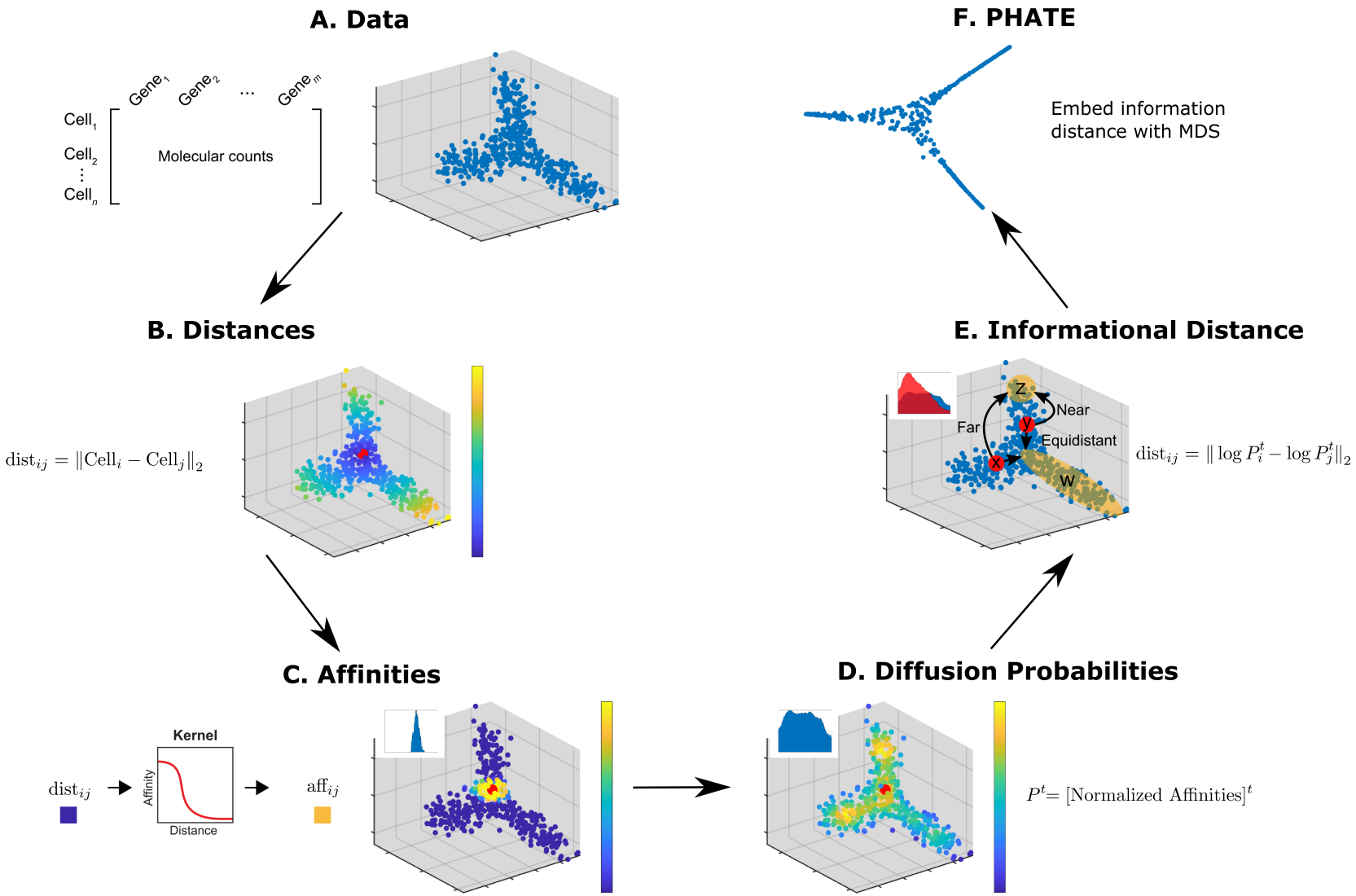
$$= \begin{cases} p^t_x(z) - p^t_y(z) & \gamma = -1 \\ \log p^t_x(z) - \log p^t_y(z) & \gamma = +1 \\ \frac{2}{1-\gamma}\left[(p^t_x(z))^{\frac{1-\gamma}{2}} - (p^t_y(z))^{\frac{1-\gamma}{2}}\right] & \text{otherwise} \end{cases}$$



Data visualization

PHATE (Moon et al., Nat. Biotech. 2019)

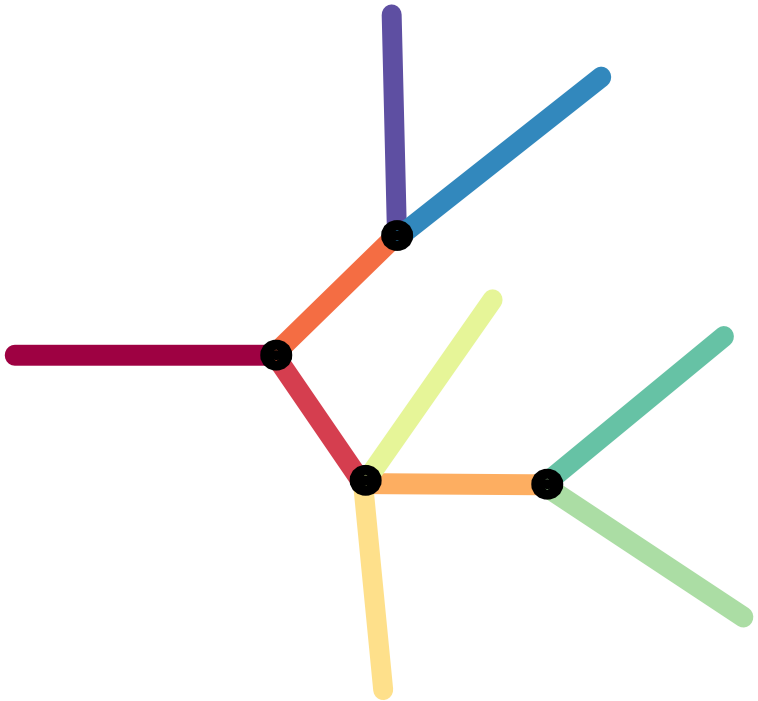
Overview



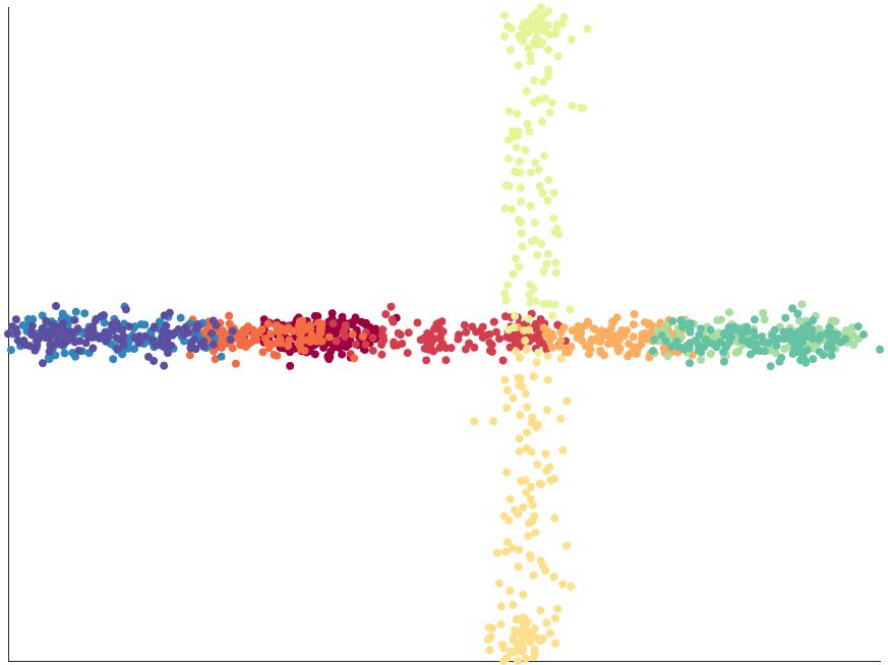
Data visualization

Example #1: artificial tree

PHATE (Moon et al., Nat. Biotech. 2019)



PCA:

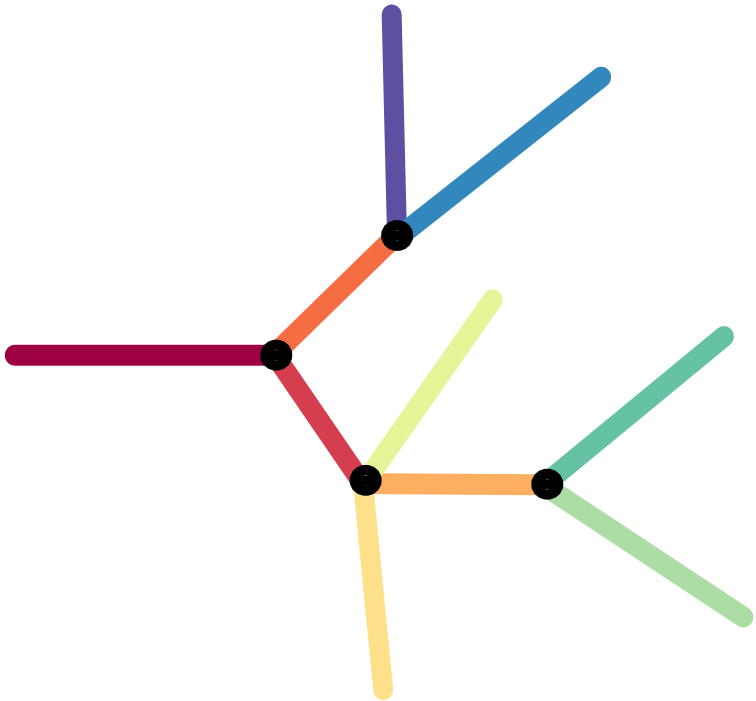


- 40 dimensions, dense regions at branch- and end-points

Data visualization

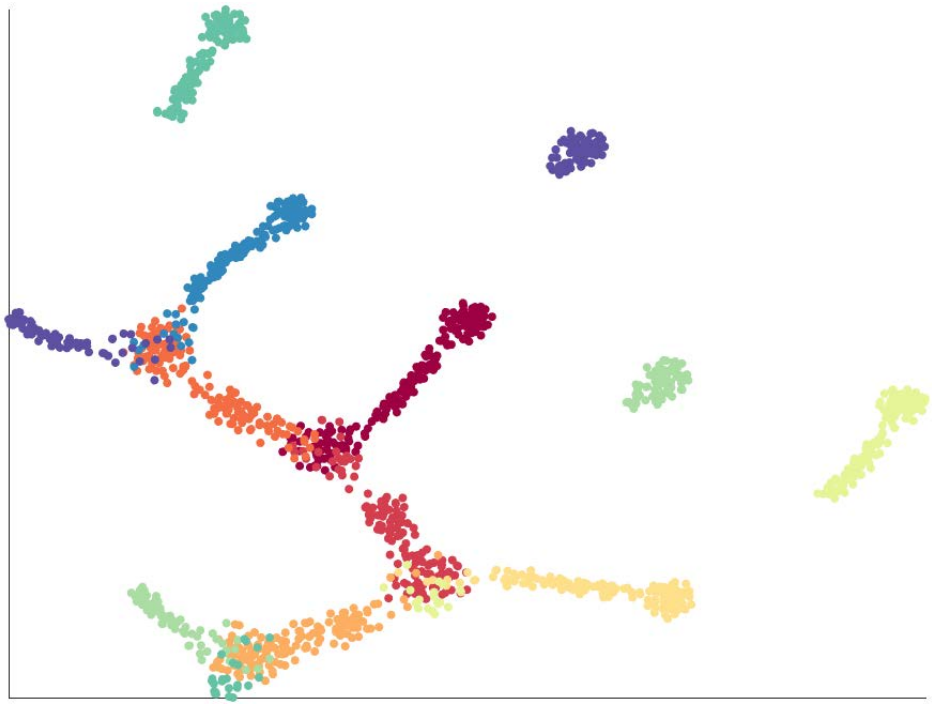
PHATE (Moon et al., Nat. Biotech. 2019)

Example #1: artificial tree



- 40 dimensions, dense regions at branch- and end-points

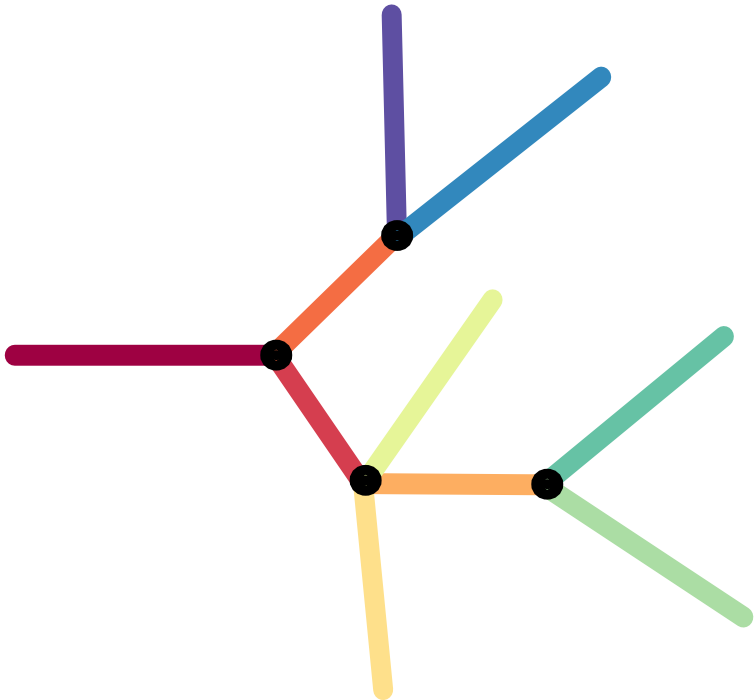
tSNE:



Data visualization

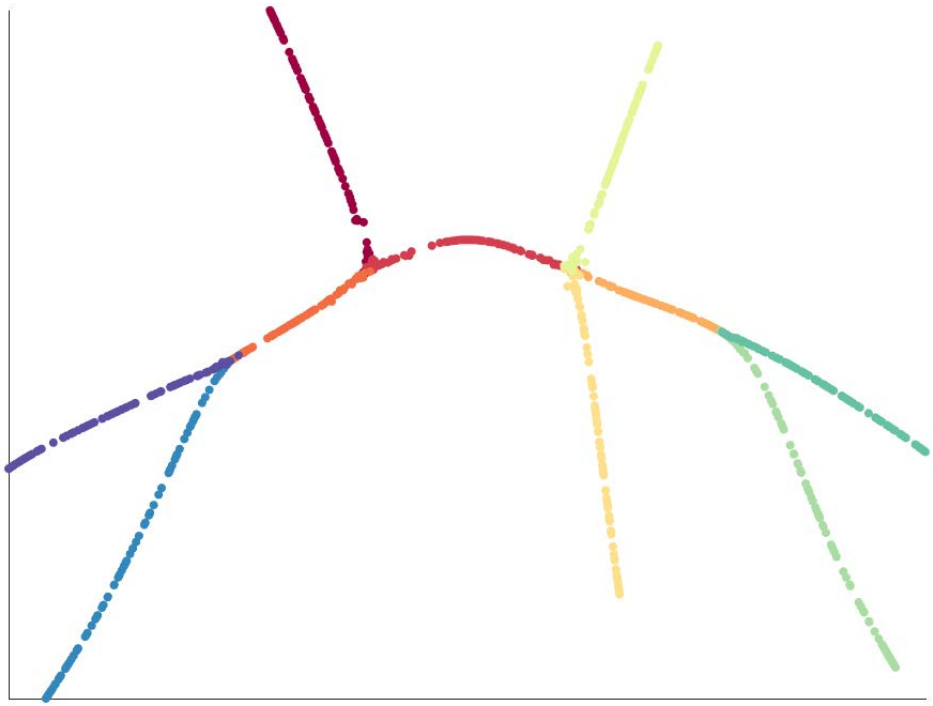
PHATE (Moon et al., Nat. Biotech. 2019)

Example #1: artificial tree



- 40 dimensions, dense regions at branch- and end-points

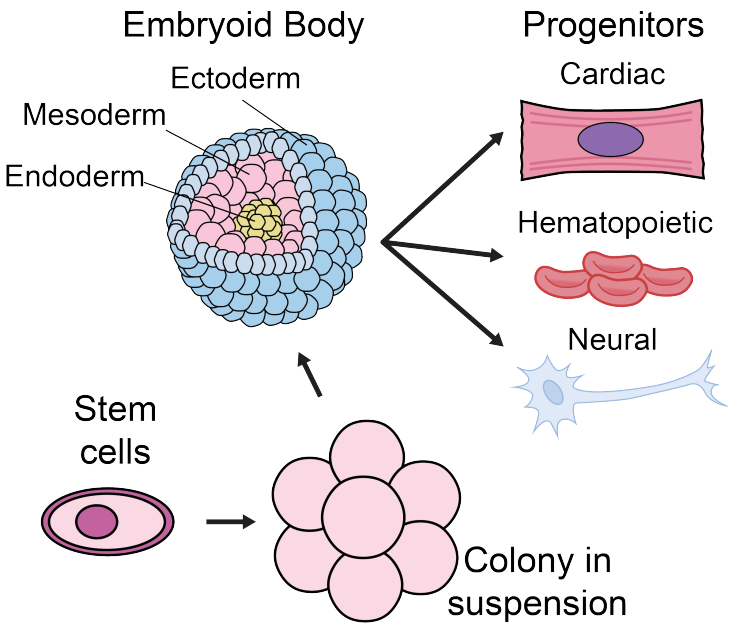
PHATE:



Data visualization

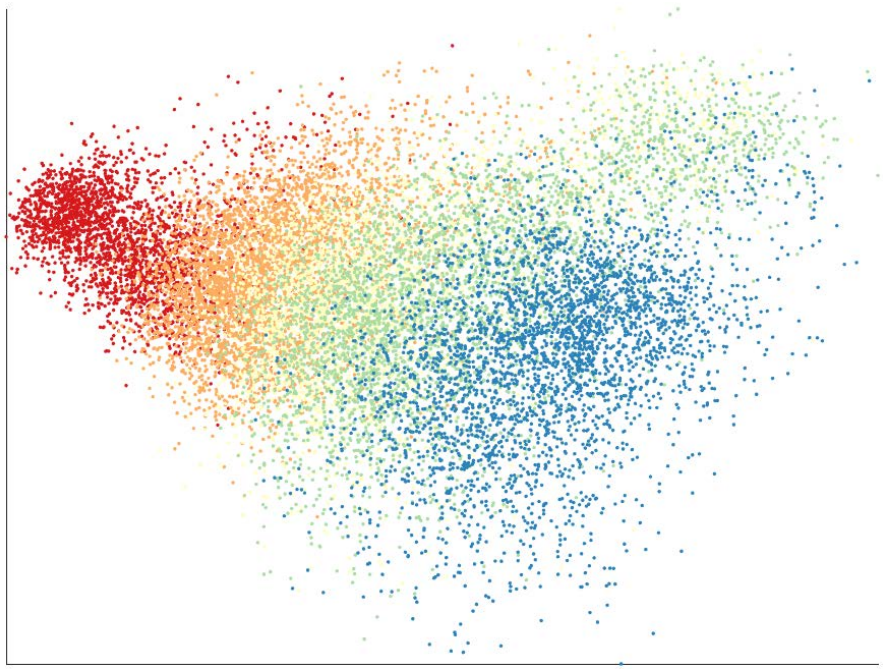
PHATE (Moon et al., Nat. Biotech. 2019)

Example #2: exploring differentiation trajectories in Embryoid Bodies



- New single-cell RNA-sequencing measured over 27-day timecourse

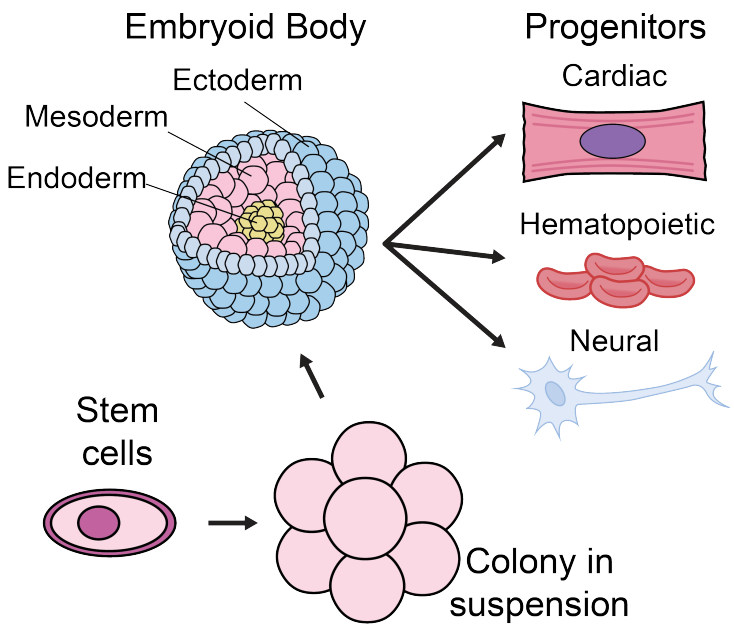
PCA:



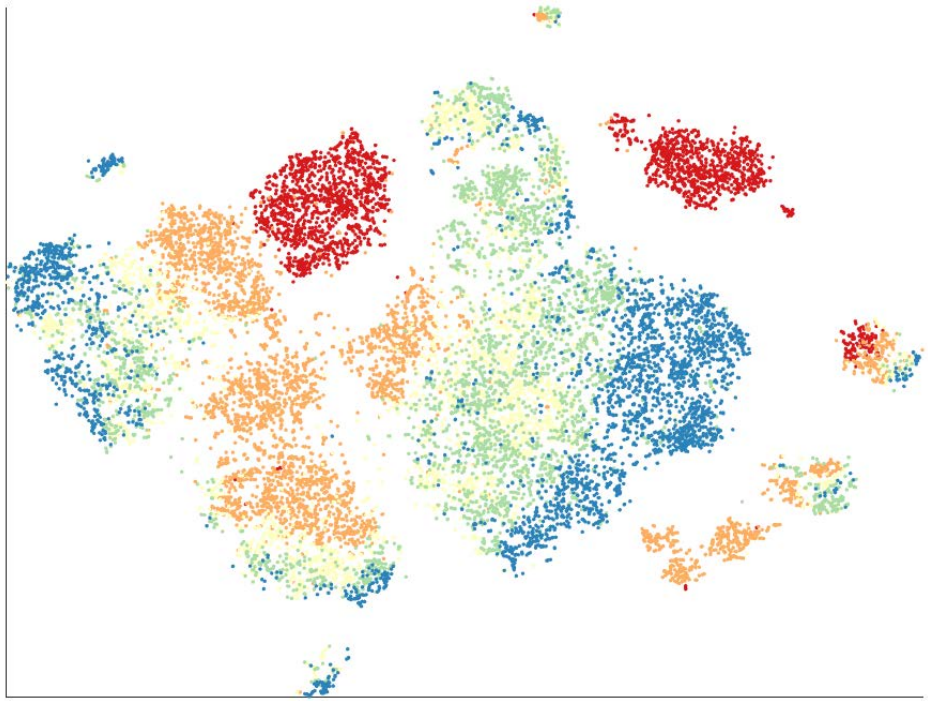
Data visualization

PHATE (Moon et al., Nat. Biotech. 2019)

Example #2: exploring differentiation trajectories in Embryoid Bodies



tSNE:

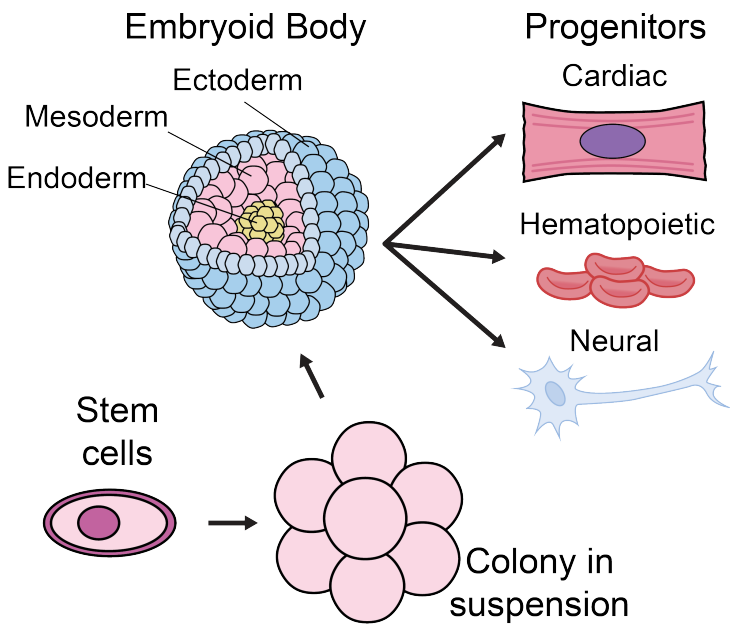


- New single-cell RNA-sequencing measured over 27-day timecourse

Data visualization

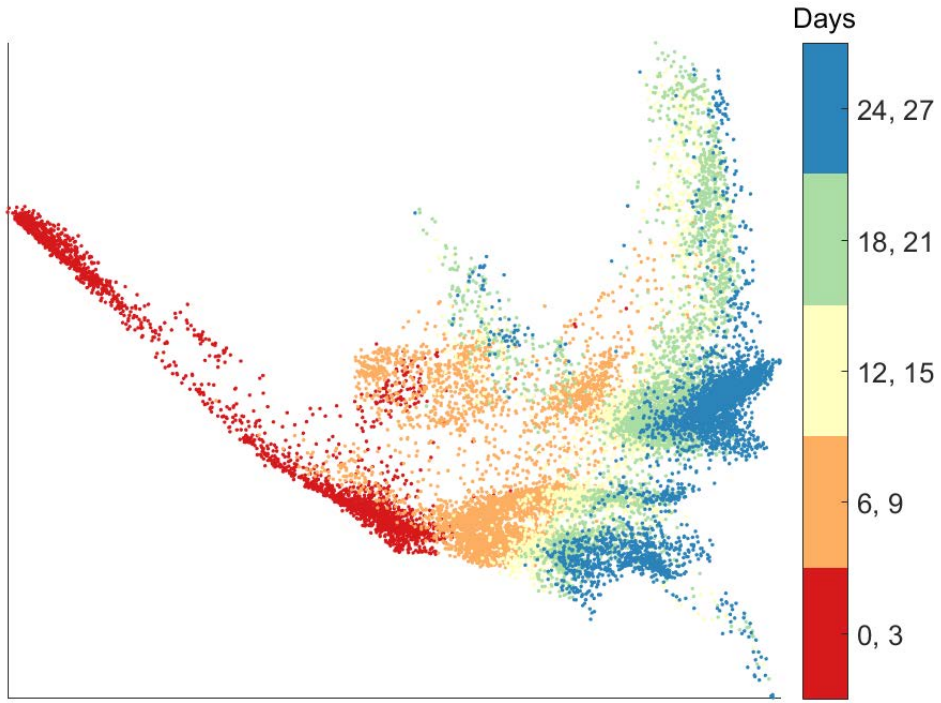
PHATE (Moon et al., Nat. Biotech. 2019)

Example #2: exploring differentiation trajectories in Embryoid Bodies



- New single-cell RNA-sequencing measured over 27-day timecourse

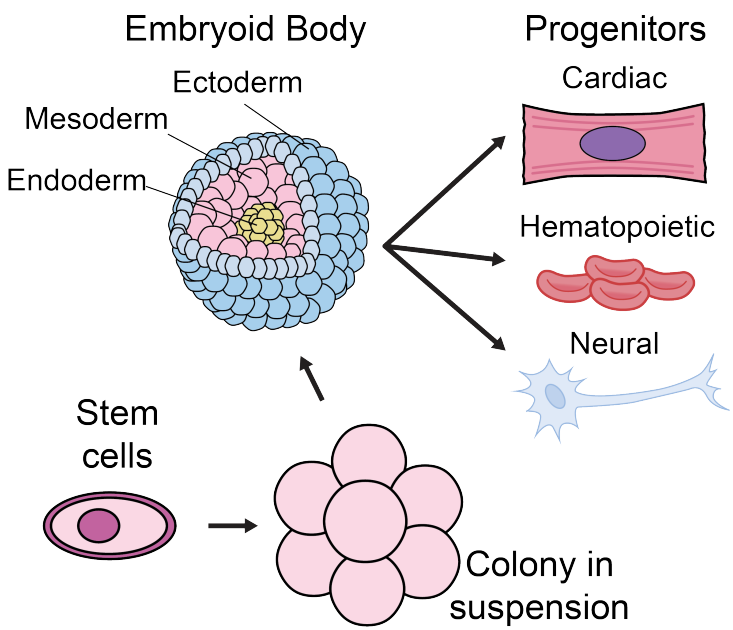
PHATE:



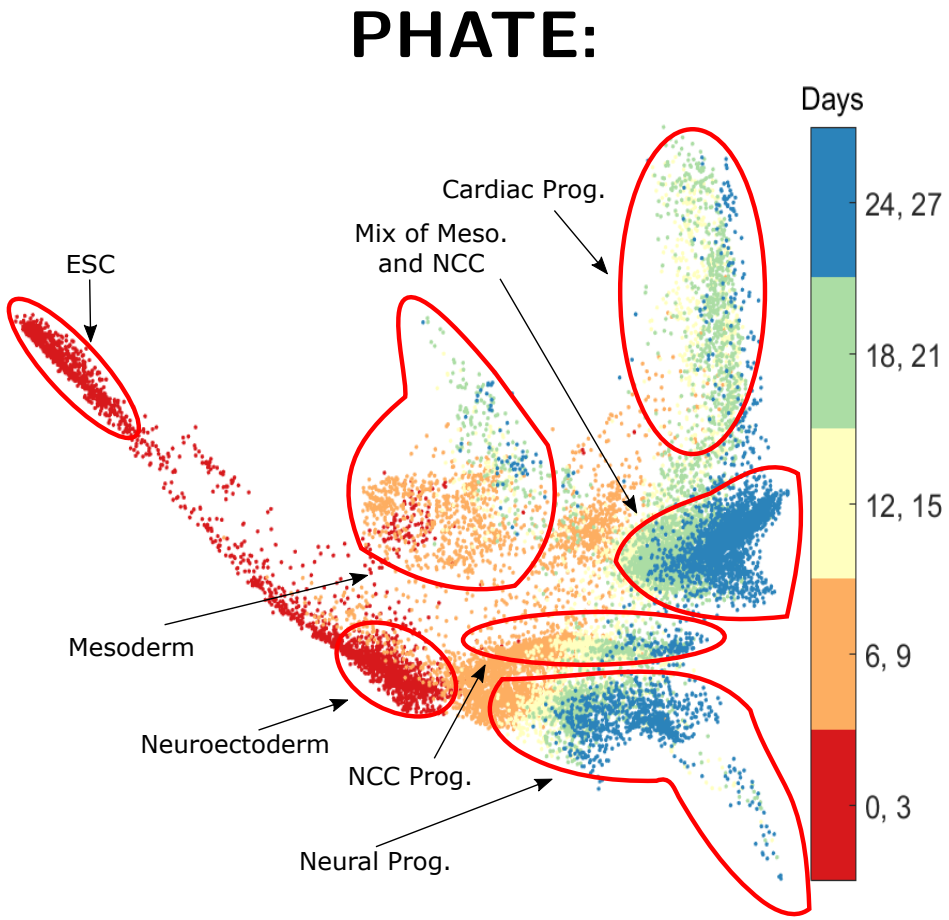
Data visualization

PHATE (Moon et al., Nat. Biotech. 2019)

Example #2: exploring differentiation trajectories in Embryoid Bodies



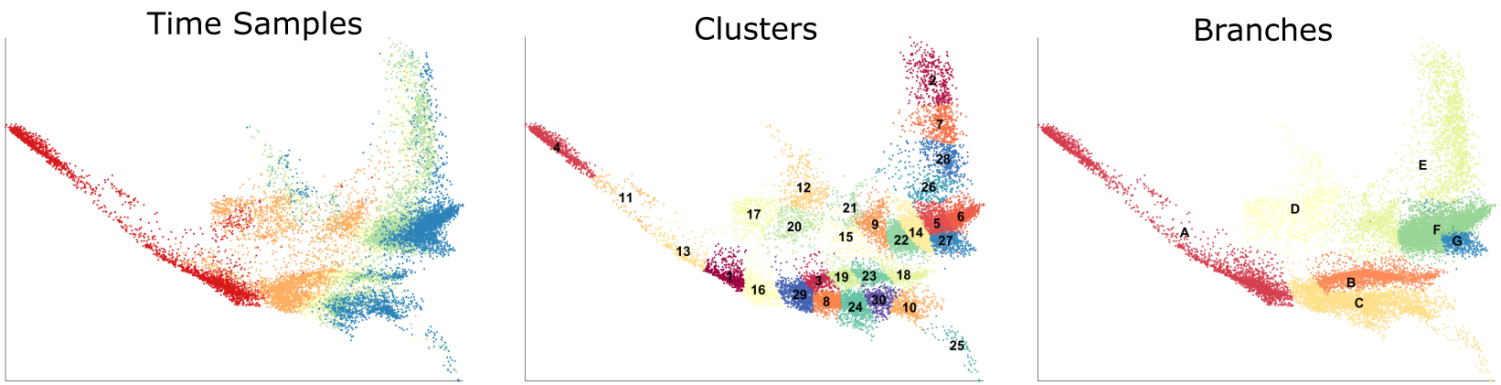
- New single-cell RNA-sequencing measured over 27-day timecourse

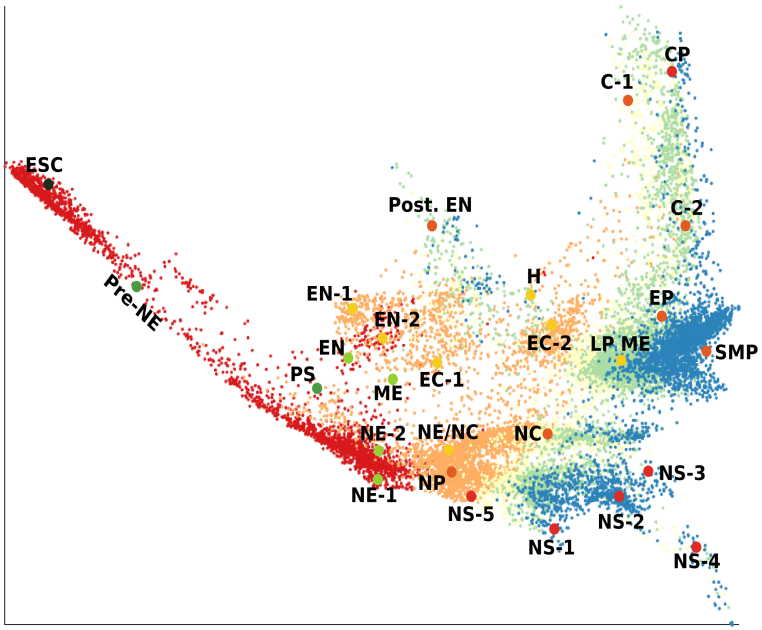


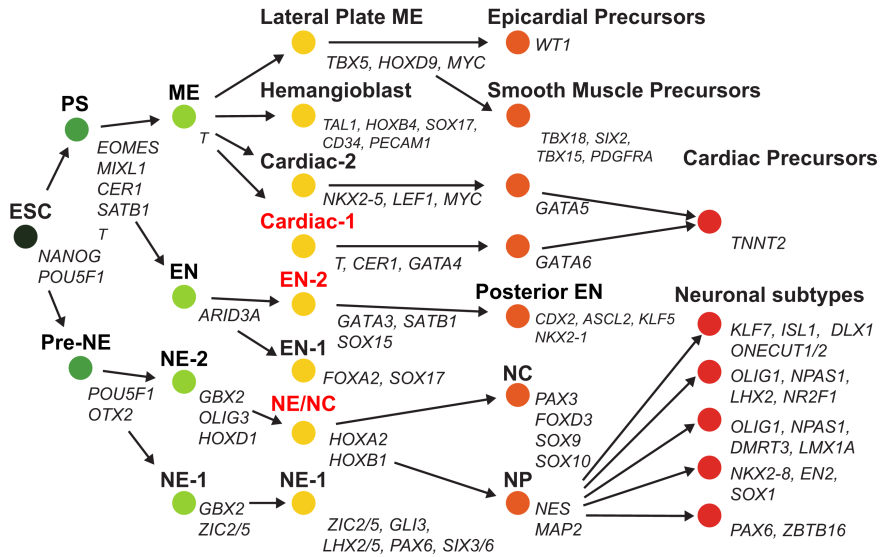
Data visualization

PHATE (Moon et al., Nat. Biotech. 2019)

Example #2: exploring differentiation trajectories in Embryoid Bodies



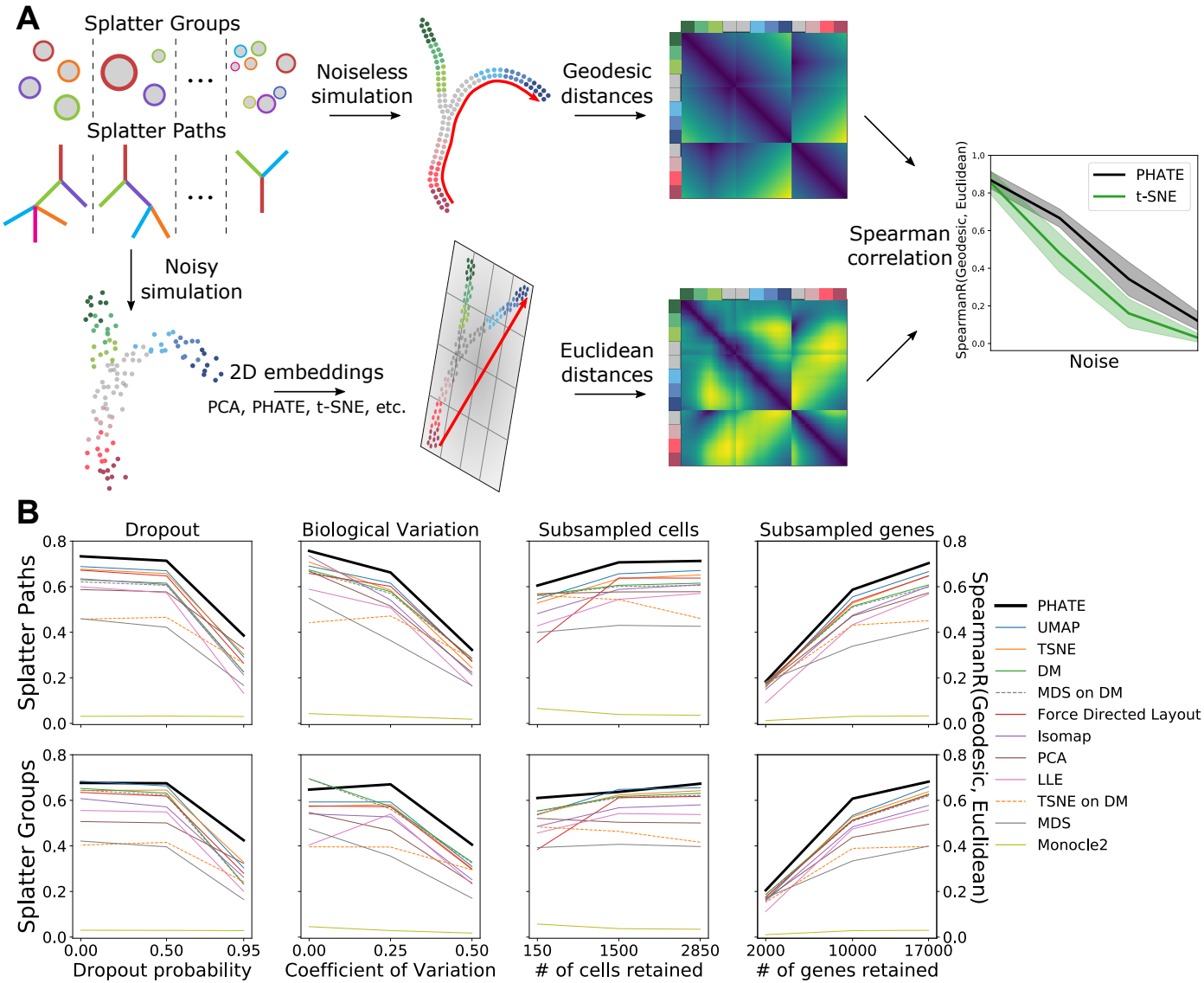




Data visualization

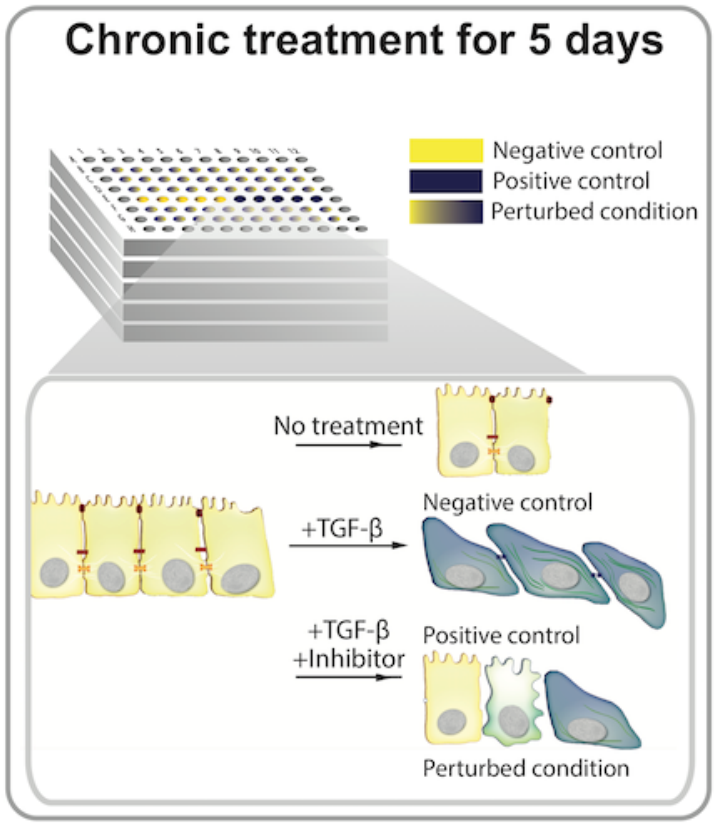
PHATE (Moon et al., Nat. Biotech. 2019)

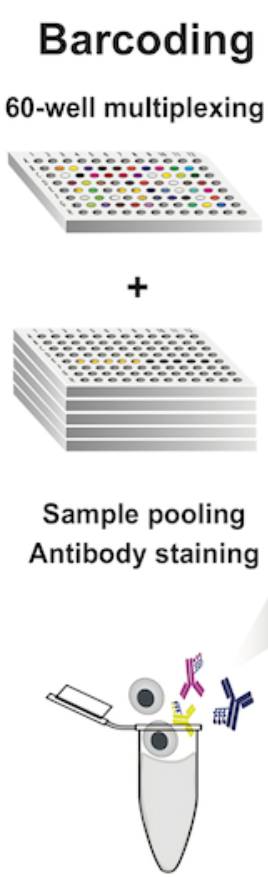
Evaluation & comparisons



Drug data exploration PhEMD (Chen et al., Nat. Meth. 2020)

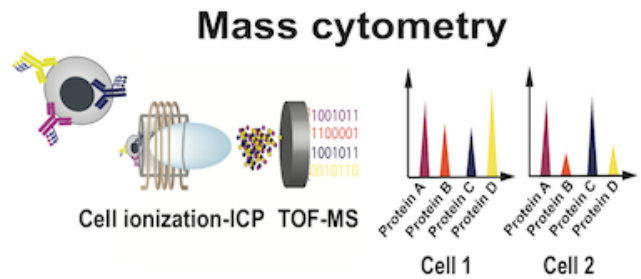
Embedding cell distributions with EMD-based diffusion maps (Coifman & Lafon, 2006)





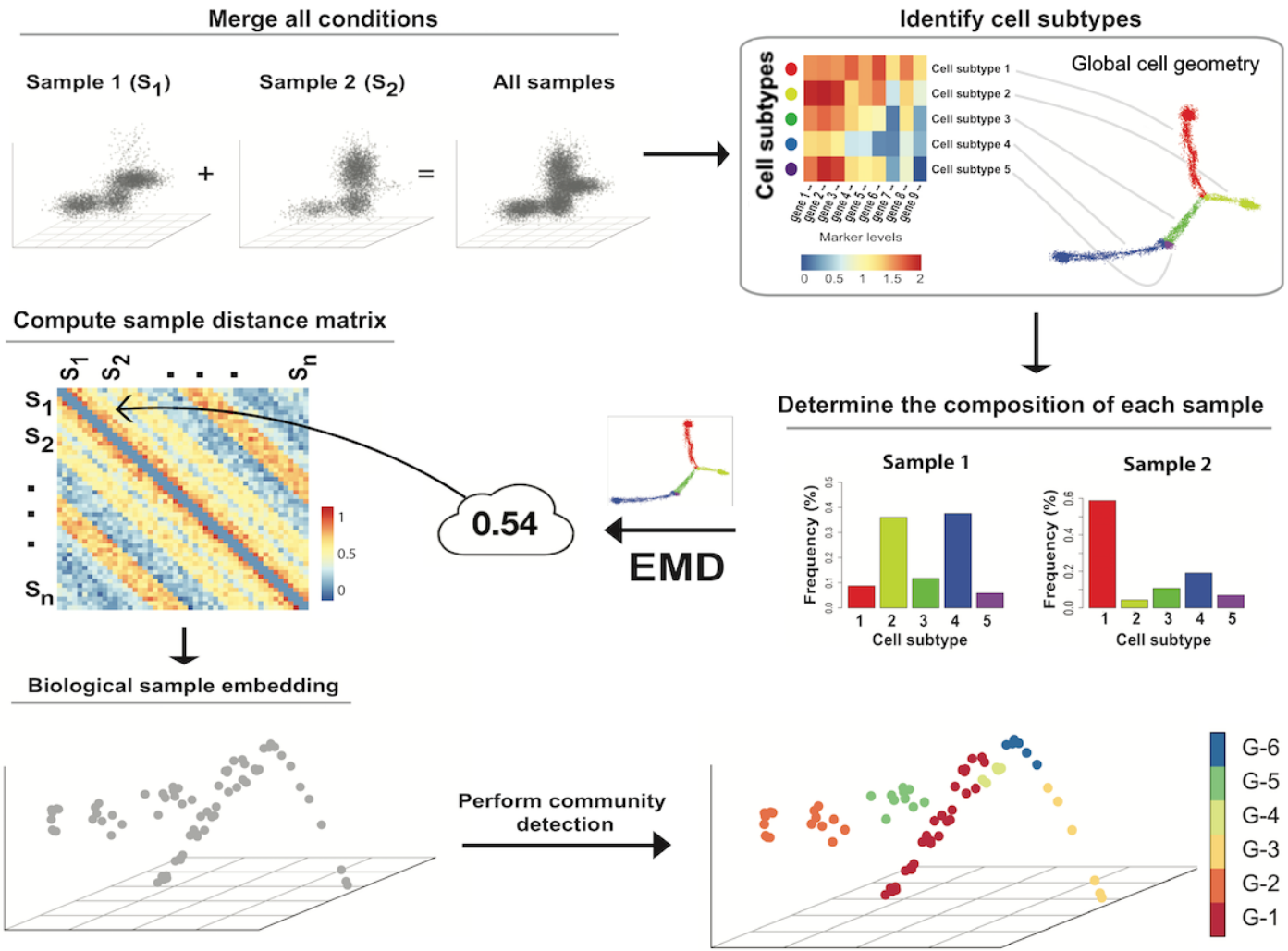
Antibody panel

Phenotypic Markers	Cell Metabolism	Survival
E-cadherin	p-AMPK	Cleaved Caspase 3
β-catenin	**TGF-β/BMP pathway**	c-myc
CD24	p-Smad2/p-Smad3	Survivin
Vimentin	p-Smad1/5	**MAPK pathway**
CD49e (ITGA5)	**EMT TF**	p-MEK1/2
CD44	Twist	p-ERK1/2
Cell Cycle	Snail	p-Creb
Cyclin B1	Nanog	p-P38
p-Rb	**AKT pathway**	**RTK**
p-H3	p-AKT	p-PLCγ2
p-MARCKS	p-S6	p-SHP2
JAK/STAT pathway	p-GSK3β	p-Tie2
p-STAT3	**NFκB signaling**	**Integrin associated**
p-STAT5	p-NFκB p65	p-FAK



Drug data exploration PhEMD (Chen et al., Nat. Meth. 2020)

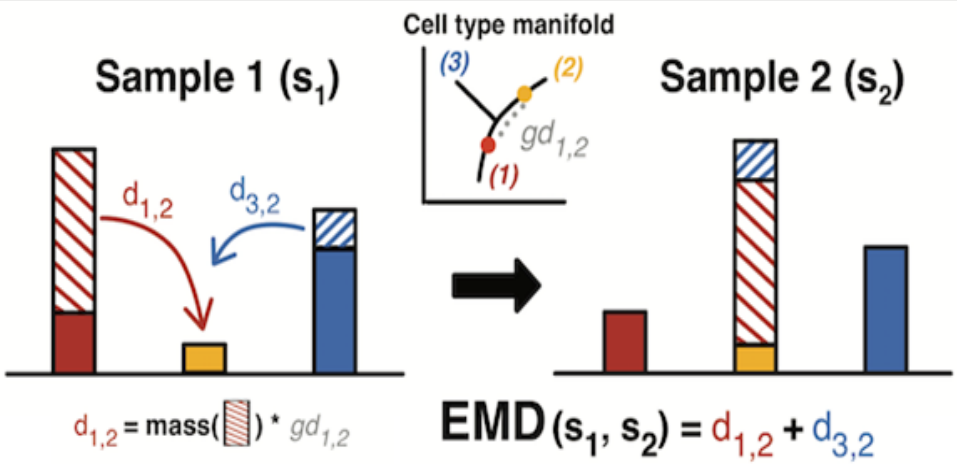
Embedding cell distributions with EMD-based diffusion maps (Coifman & Lafon, 2006)

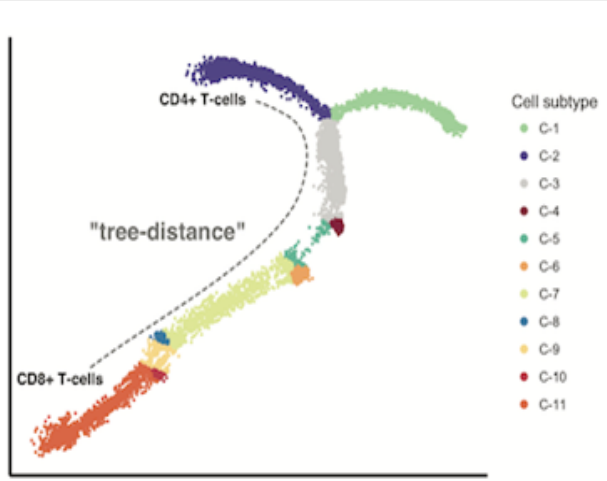


Drug data exploration PhEMD (Chen et al., Nat. Meth. 2020)

Embedding cell distributions with EMD-based diffusion maps (Coifman & Lafon, 2006)

Earth-Mover's Distances (EMD) between samples quantify the intrinsic difference in cell distribution over the data manifold.





Diffusion maps embedding of samples:

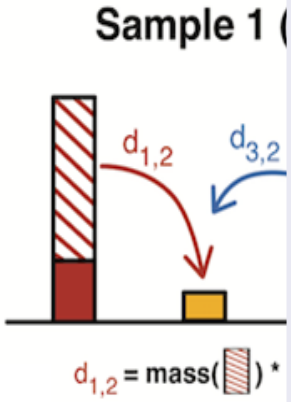
1. Pairwise EMD → sample neighborhoods → sample-wise diffusion
2. Eigendecomposition of P^t → diffusion coordinates of samples

Drug data exploration PhEMD (Chen et al., Nat. Meth. 2020)

Embedding cell distributions with EMD-based diffusion maps (Coifman & Lafon, 2006)

Earth-Move

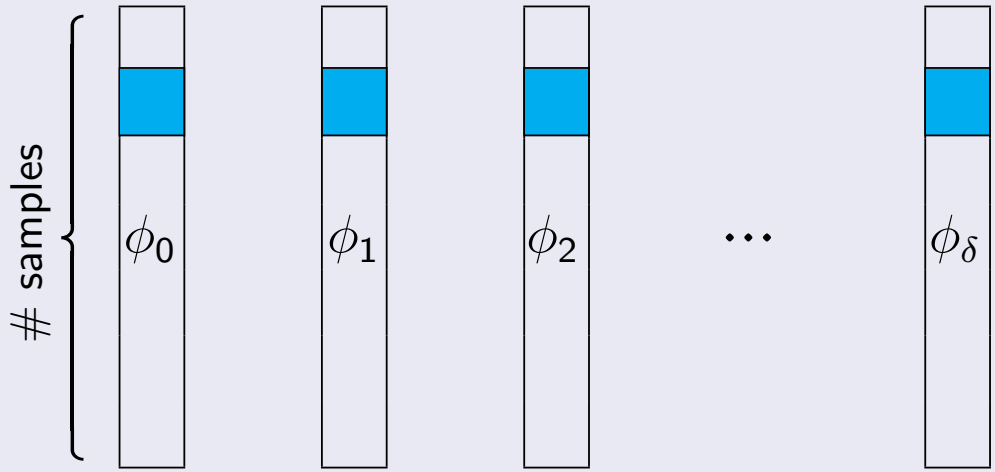
intrinsic dif



Diffusion coordinates

Eigenvals of P: $1 = \boxed{\lambda_0} \geq \boxed{\lambda_1} \geq \boxed{\lambda_2} \geq \cdots \geq \boxed{\lambda_\delta} > 0$

Eigenvecs of P:



Diffusion Map at time t:

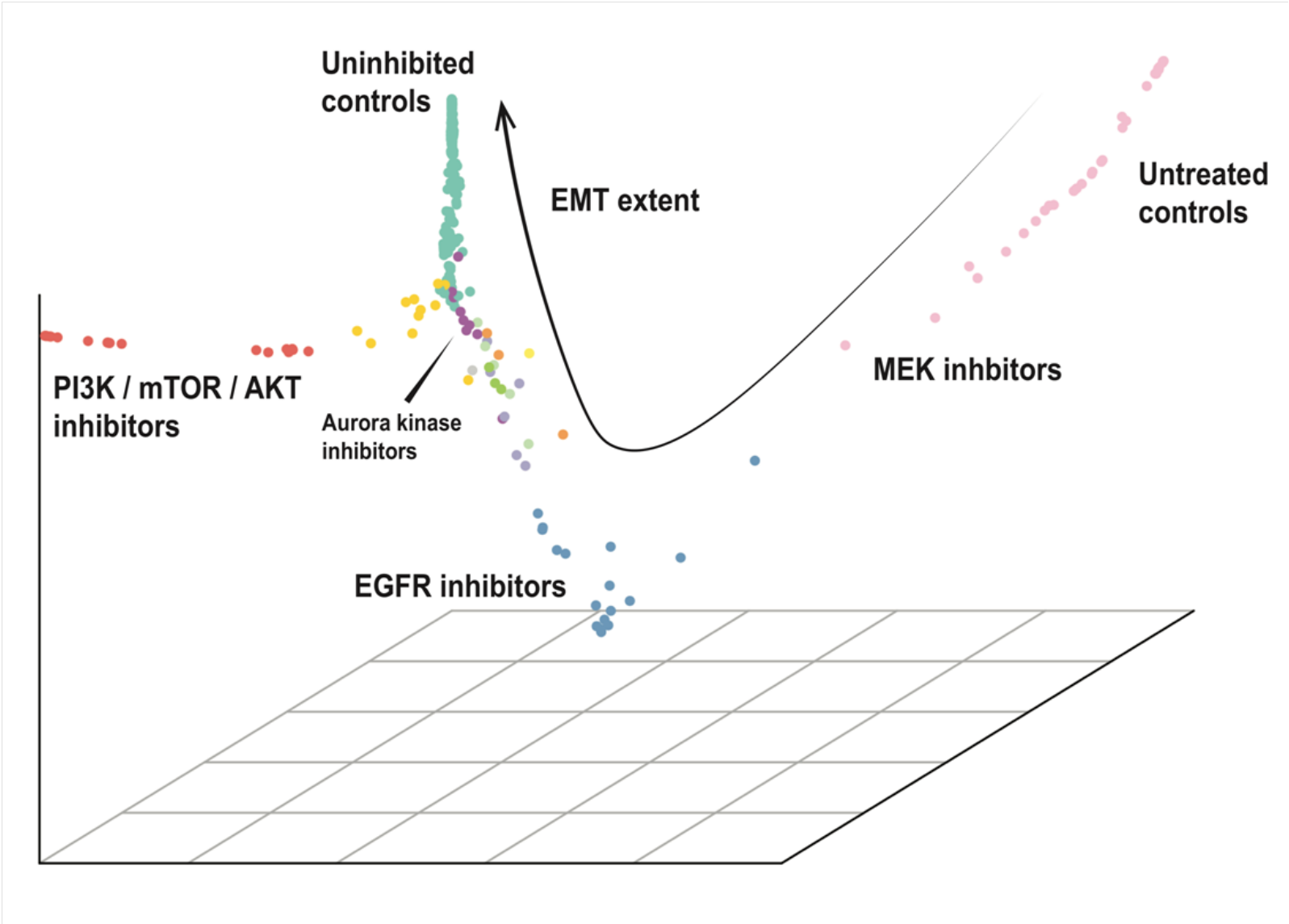
$$S \mapsto \Phi^t(S) \triangleq [\lambda_1^t \phi_1(S), \lambda_2^t \phi_2(S), \ldots, \lambda_\delta^t \phi_\delta(x)]^T$$

Diffusion m

1. Pairwi
2. Eigendecomposition of $P^t \rightarrow$ diffusion coordinates of samples

Drug data exploration PhEMD (Chen et al., Nat. Meth. 2020)

Embedding cell distributions with EMD-based diffusion maps (Coifman & Lafon, 2006)

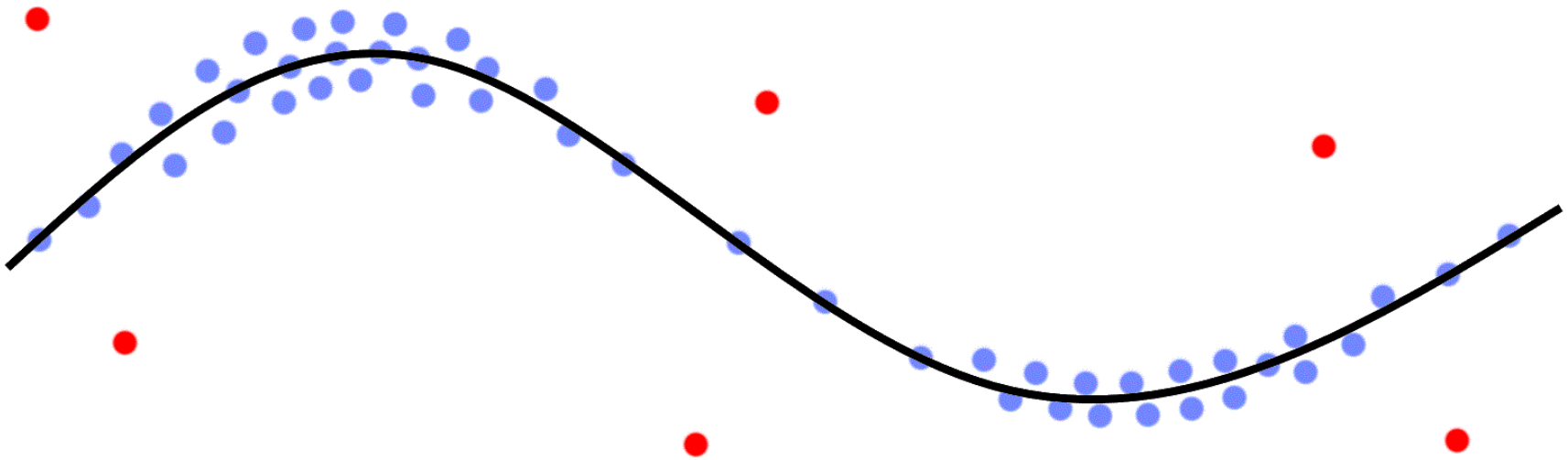


Data generation

SUGAR (Lindenbaum et al., NeurIPS 2018)

Walk toward the data manifold from randomly generated points

Generate random points:

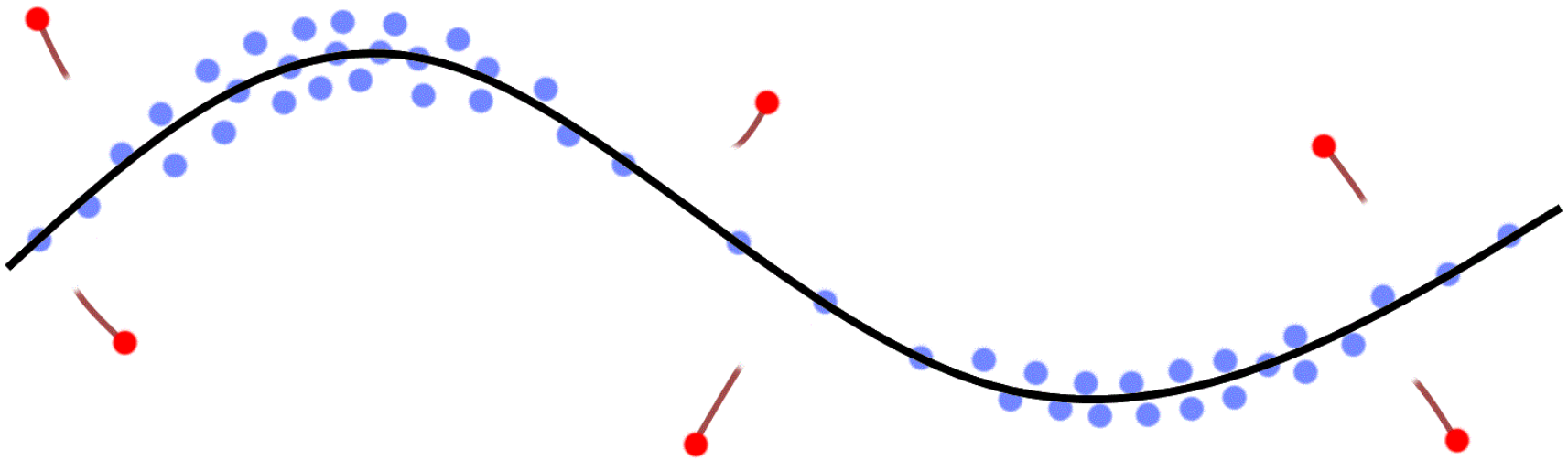


Data generation

SUGAR (Lindenbaum et al., NeurIPS 2018)

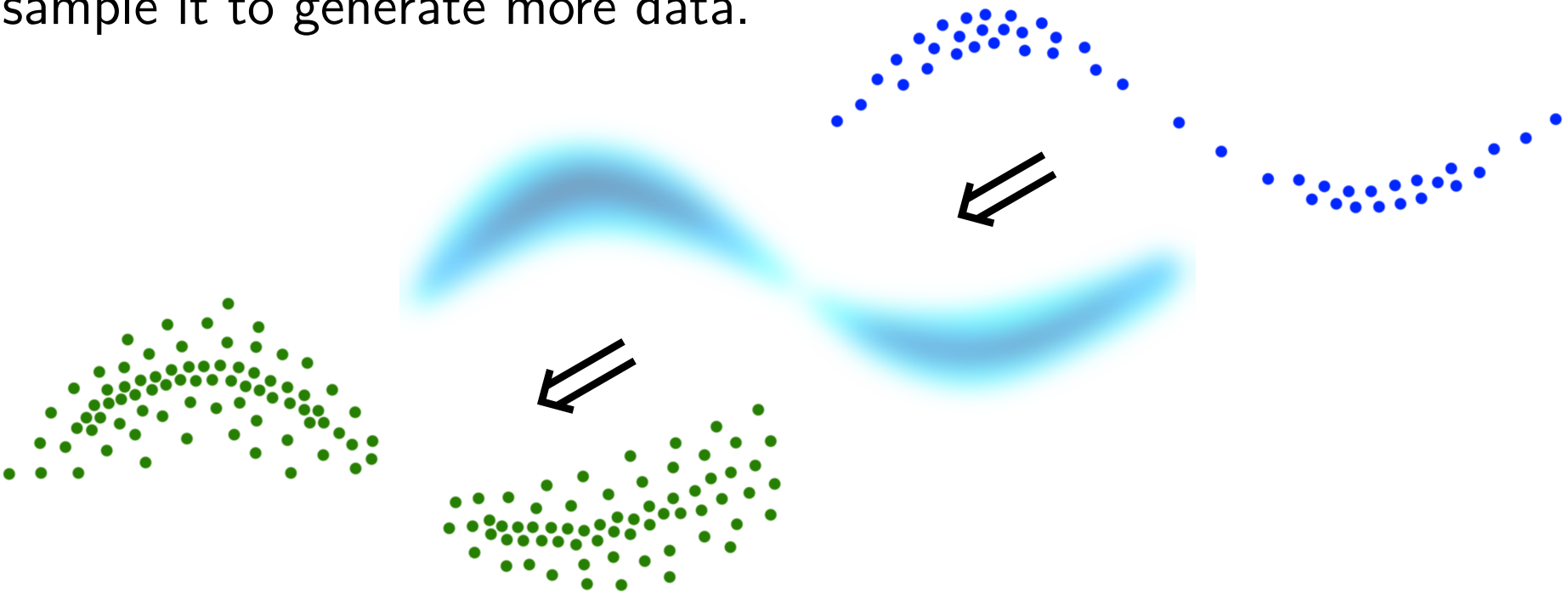
Walk toward the data manifold from randomly generated points

Generate random points:



Walk towards the data manifold with diffusion: $x \mapsto \sum_{y \in \text{data}} y \cdot p^t(x, y)$

Generative models typically infer distribution from collected data, and sample it to generate more data.

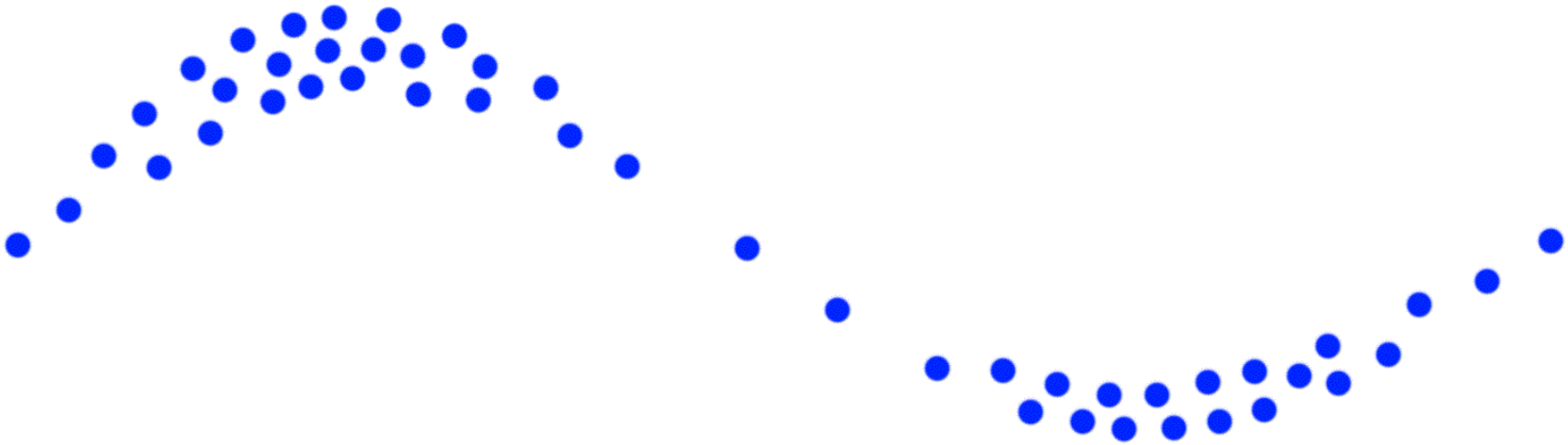


- Biased by sampling density
- May miss rare populations
- Does not preserve the geometry

Data generation

SUGAR (Lindenbaum et al., NeurIPS 2018)

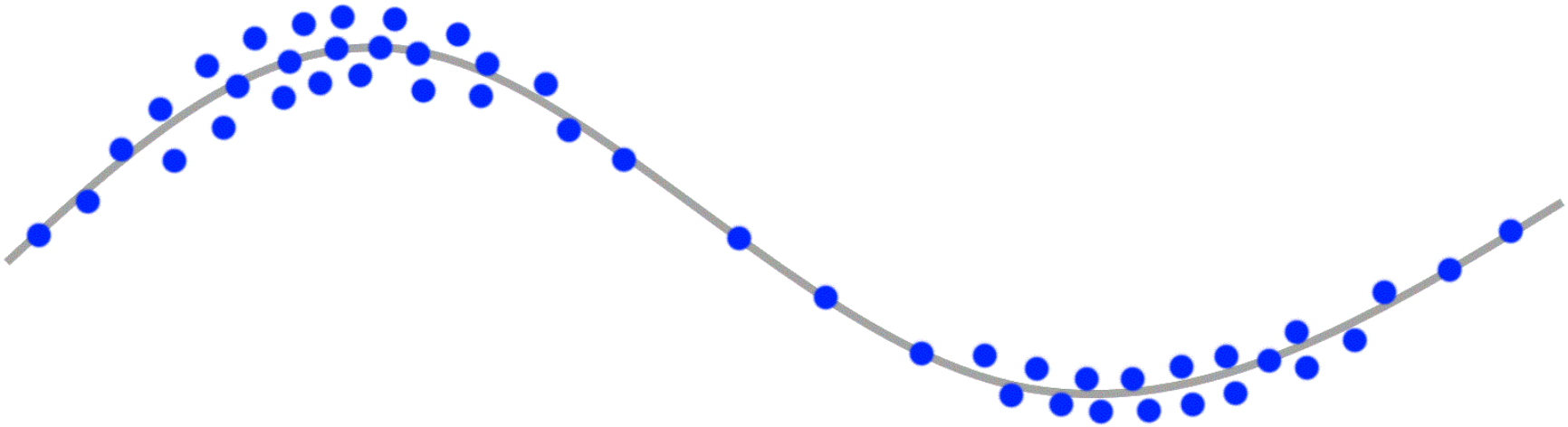
New approach: geometry based data generation



Data generation

SUGAR (Lindenbaum et al., NeurIPS 2018)

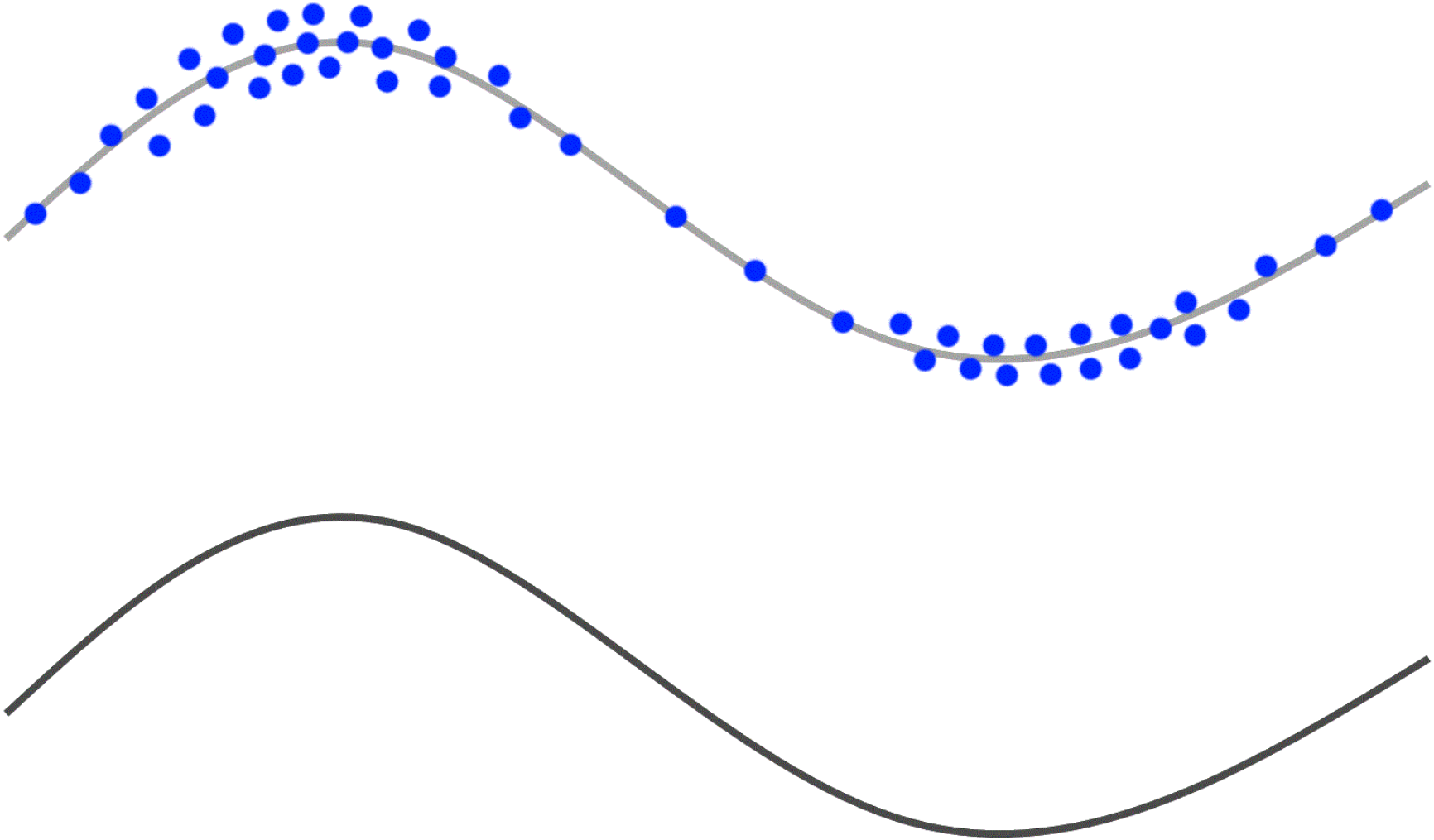
New approach: geometry based data generation



Data generation

SUGAR (Lindenbaum et al., NeurIPS 2018)

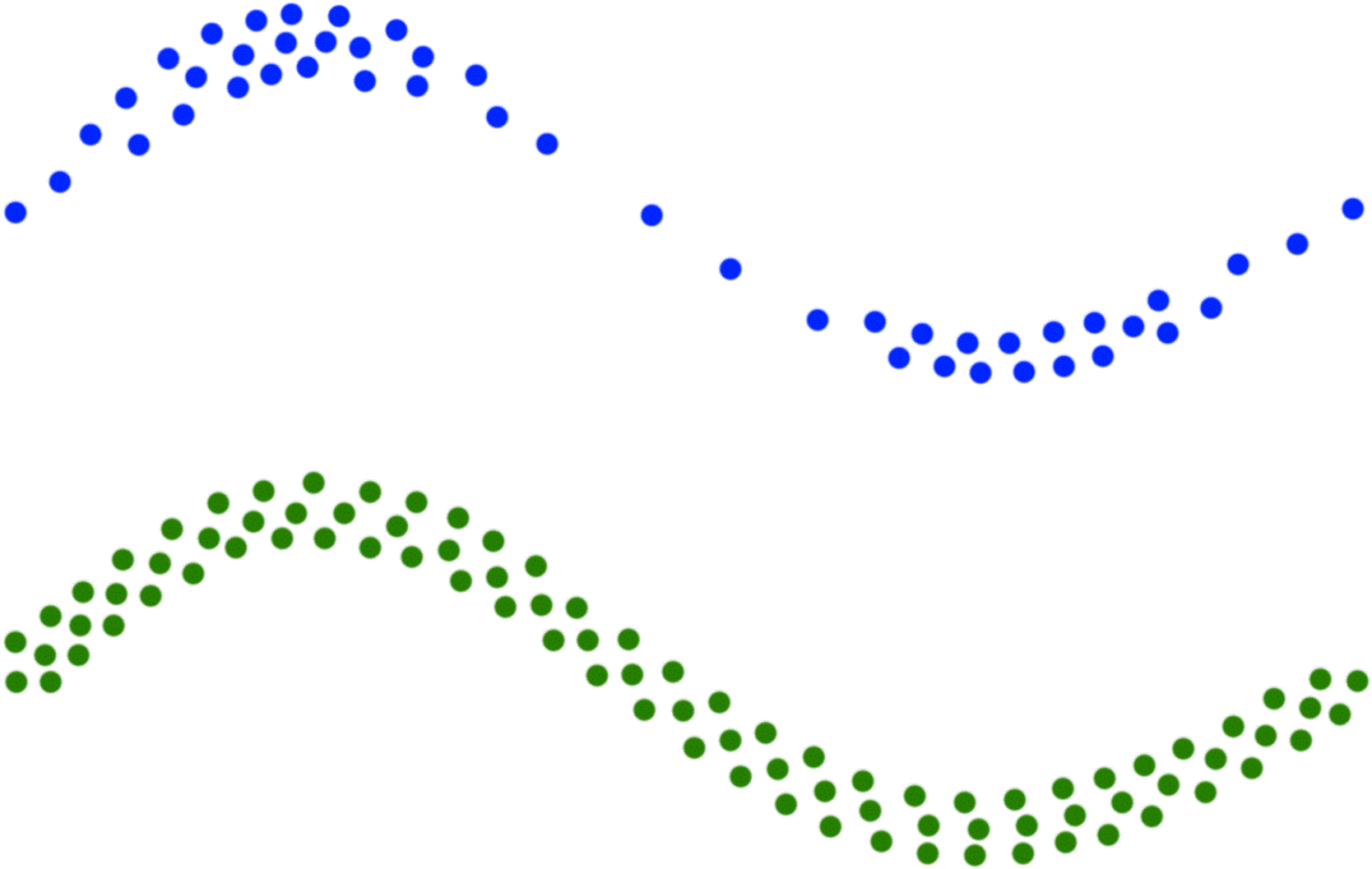
New approach: geometry based data generation



Data generation

SUGAR (Lindenbaum et al., NeurIPS 2018)

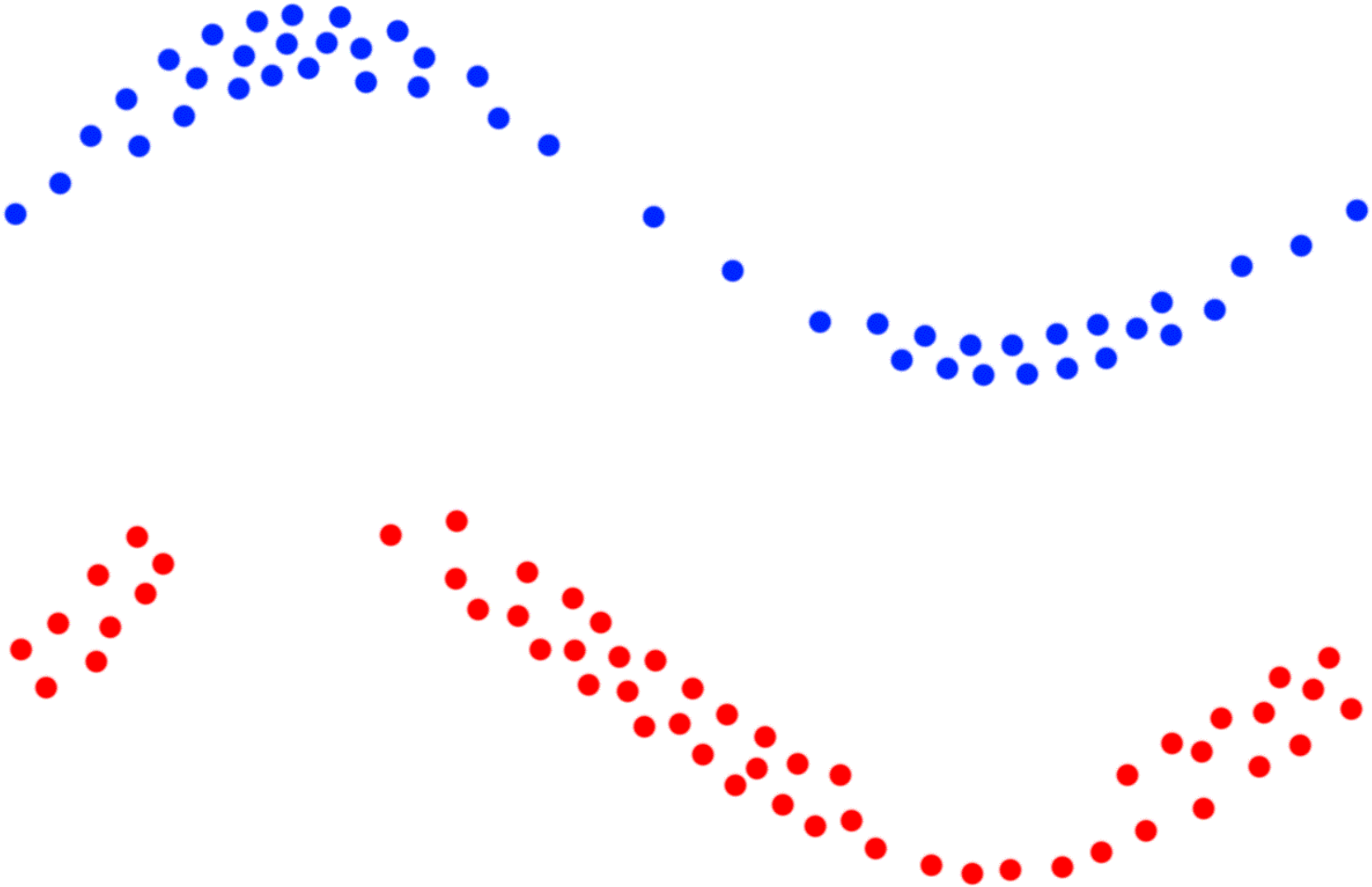
New approach: geometry based data generation



Data generation

SUGAR (Lindenbaum et al., NeurIPS 2018)

New approach: geometry based data generation

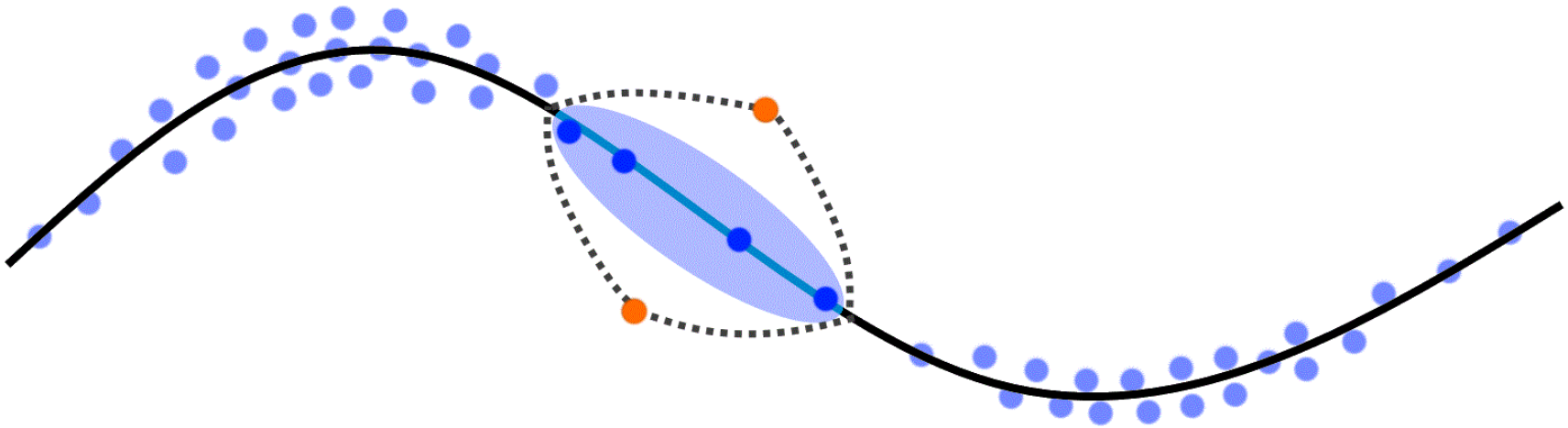


Data generation

SUGAR (Lindenbaum et al., NeurIPS 2018)

Correct density with MGC kernel (Bermanis et al., ACHA 2016)

Separate density/geometry with new kernel: $k(x,y) = \sum_{r \in \text{data}} \frac{g(x,r)\, g(y,r)}{\|g(r,\cdot)\|_1}$



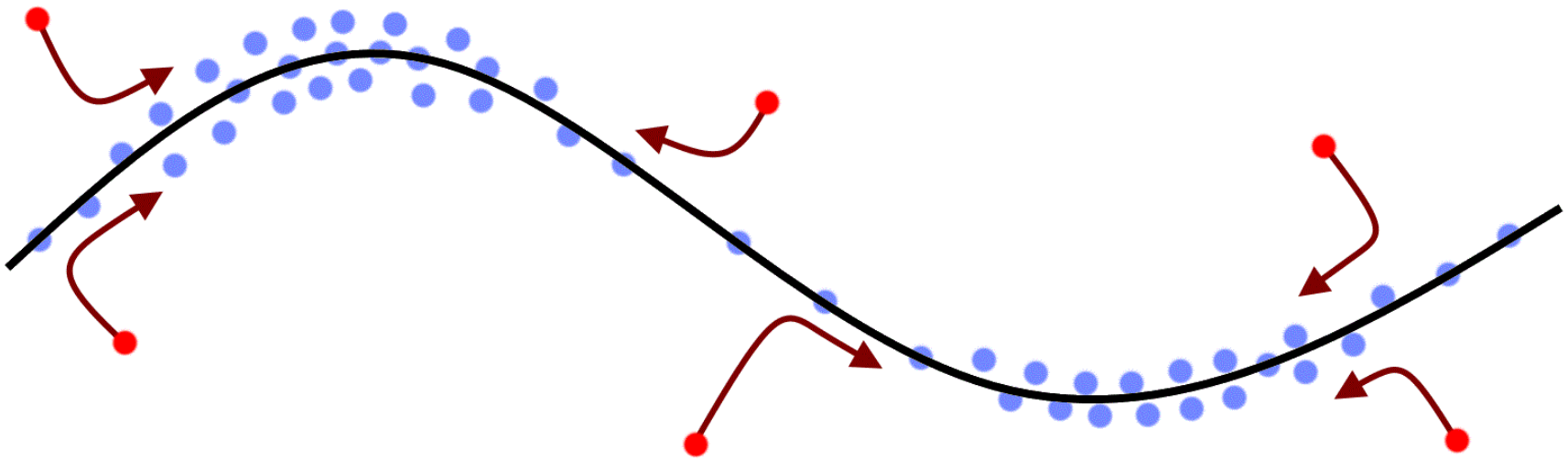
Use new diffusion process $p(x,y) = \frac{k(x,y)}{\|k(x,\cdot)\|_1}$ to walk to the manifold

Data generation

SUGAR (Lindenbaum et al., NeurIPS 2018)

Correct density with MGC kernel (Bermanis et al., ACHA 2016)

Separate density/geometry with new kernel: $k(x,y)=\sum_{r\in\text{data}} \frac{g(x,r)\,g(y,r)}{\|g(r,\cdot)\|_1}$



Use new diffusion process $p(x,y)=\frac{k(x,y)}{\|k(x,\cdot)\|_1}$ to walk to the manifold

Data generation

SUGAR (Lindenbaum et al., NeurIPS 2018)

Fill sparse areas to create uniform distribution

Question: How should we initialize new points to end up with uniform sampling from the data manifold?

Answer: For each $x \in$ data, initialize $\hat{\ell}(x)$ points sampled from $\mathcal{N}(x, \Sigma_x)$; set $\hat{\ell}$ as the mid-point between the upper & lower bounds in the following proposition.

Proposition

The generation level $\hat{\ell}(x)$ required to equalize density is bounded by

$$\det\left(I + \frac{\Sigma_x}{2\sigma^2}\right)^{\frac{1}{2}} \frac{\max(\hat{d}(\cdot)) - \hat{d}(x)}{\hat{d}(x)+1} - 1 \leq \hat{\ell}(x) \leq \det\left(I + \frac{\Sigma_x}{2\sigma^2}\right)^{\frac{1}{2}} \left[\max(\hat{d}(\cdot)) - \hat{d}(x)\right],$$

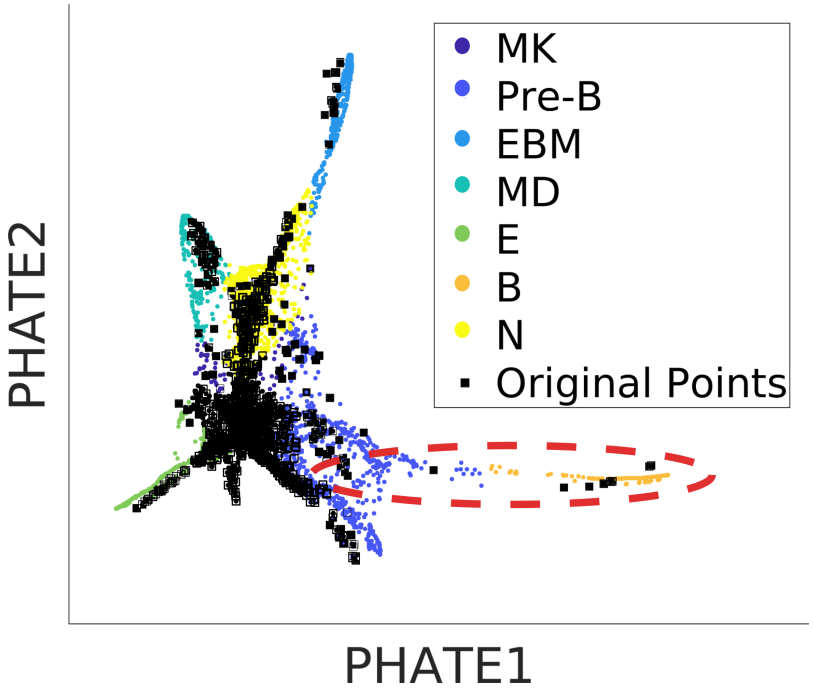
where σ is a scale used when defining Gaussian neighborhoods $g(x, y)$ for the diffusion geometry, and $\hat{d}(x) = \|g(x, \cdot)\|_1$ estimates local density.

Data generation

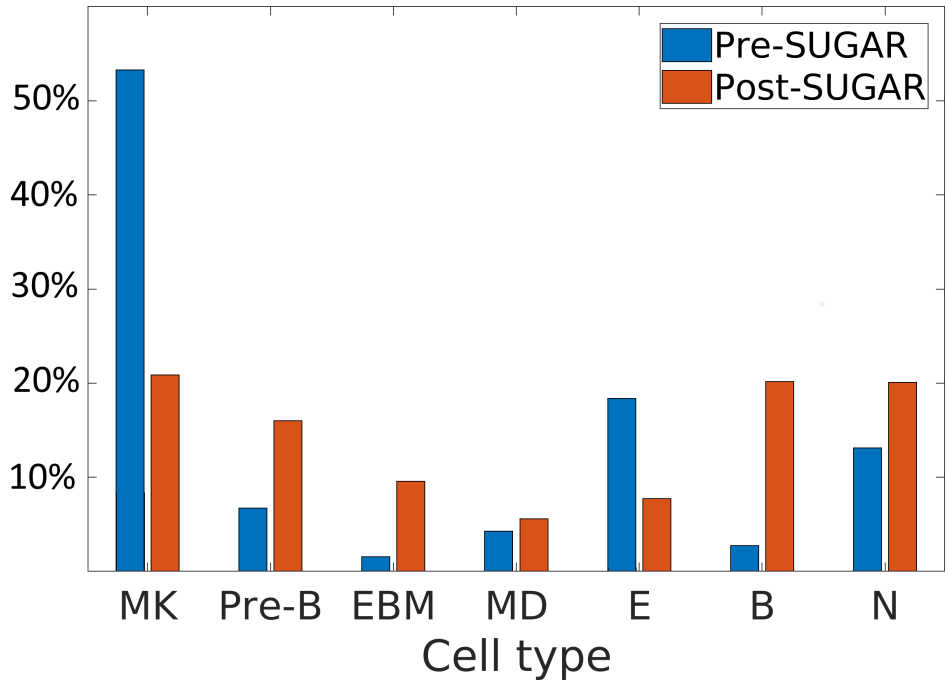
SUGAR (Lindenbaum et al., NeurIPS 2018)

Illuminate hypothetical cell types in single-cell data from Velten et al. (2017)

Recovering originally-undersampled lineage in early hematopoeisis:



B-cell maturation trajectory enhanced by SUGAR



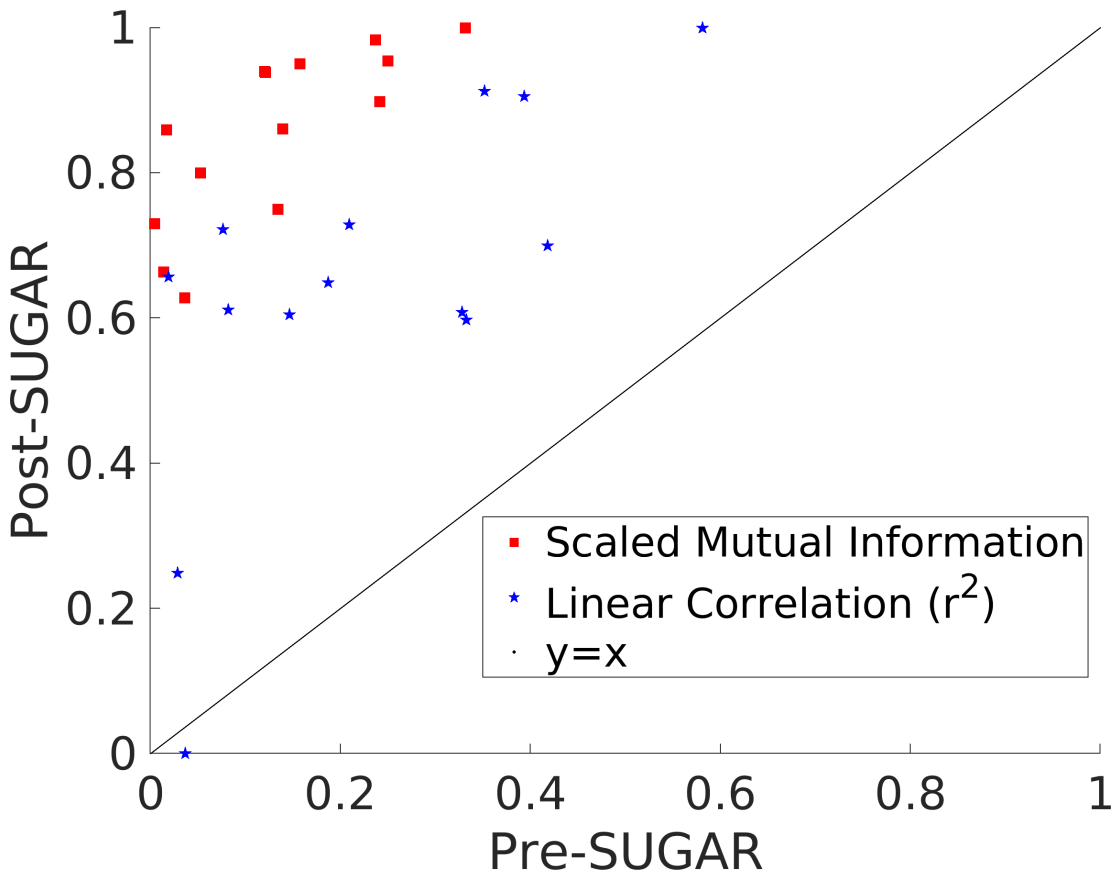
SUGAR equalizes the total cell distribution

Data generation

SUGAR (Lindenbaum et al., NeurIPS 2018)

Recover gene-gene relationships in single-cell data from Velten et al. (2017)

SUGAR improves module correlation and MI identified by Velten et al.



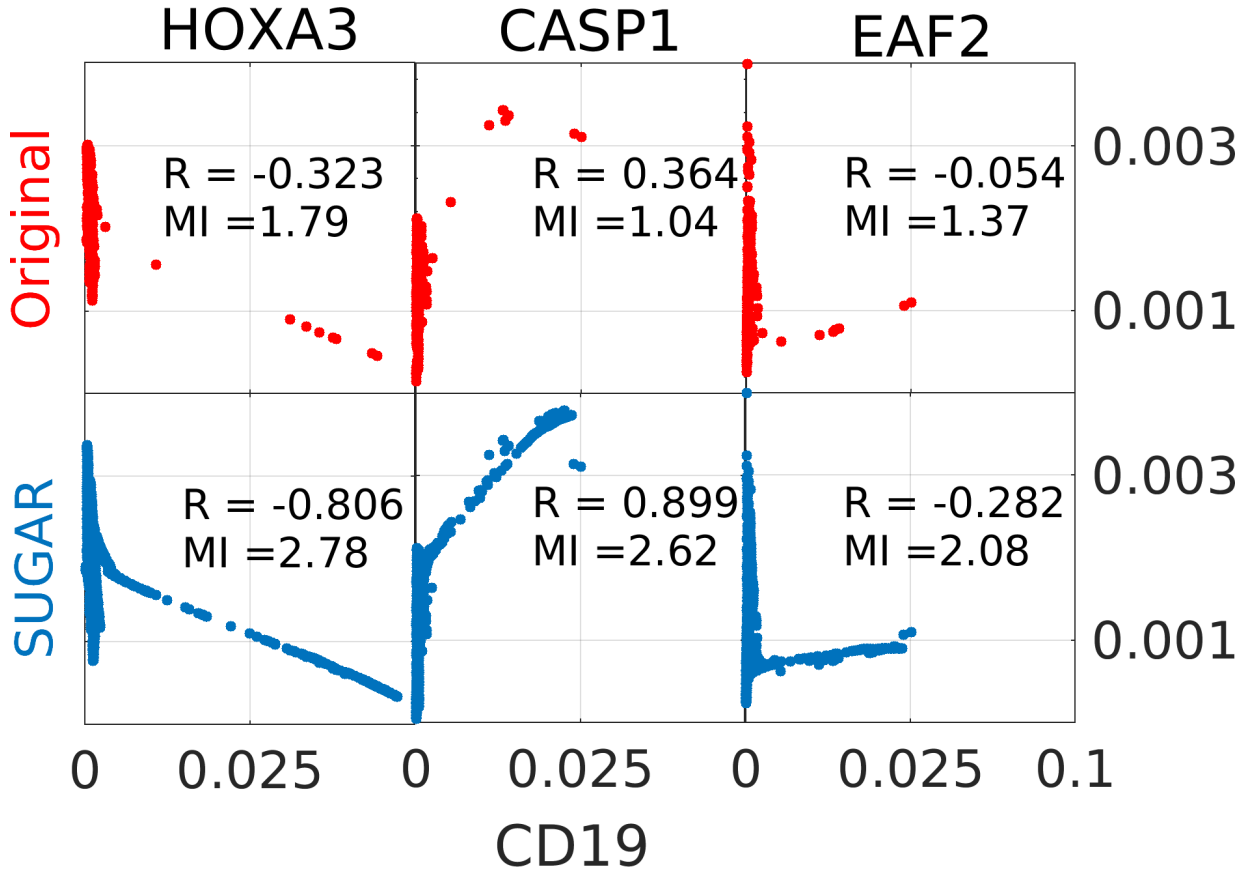
Velten et al., Nature Cell Biology 19 (2017)

Data generation

SUGAR (Lindenbaum et al., NeurIPS 2018)

Recover gene-gene relationships in single-cell data from Velten et al. (2017)

Generated cells also follow canonical marker correlations

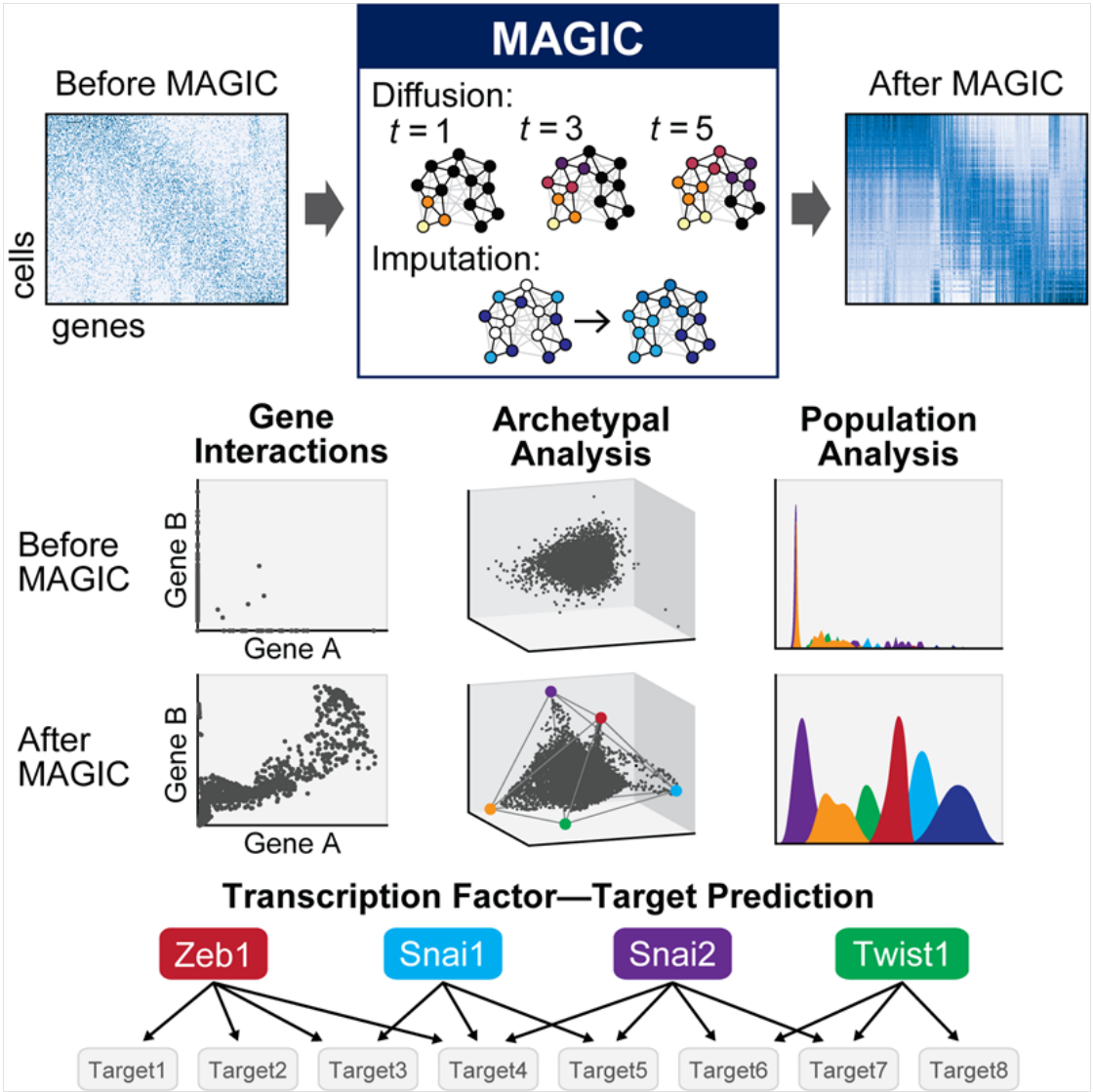


Li et al., Nature communications 7 (2016)

Imputation & denoising

MAGIC (van Dijk et al., Cell 2018)

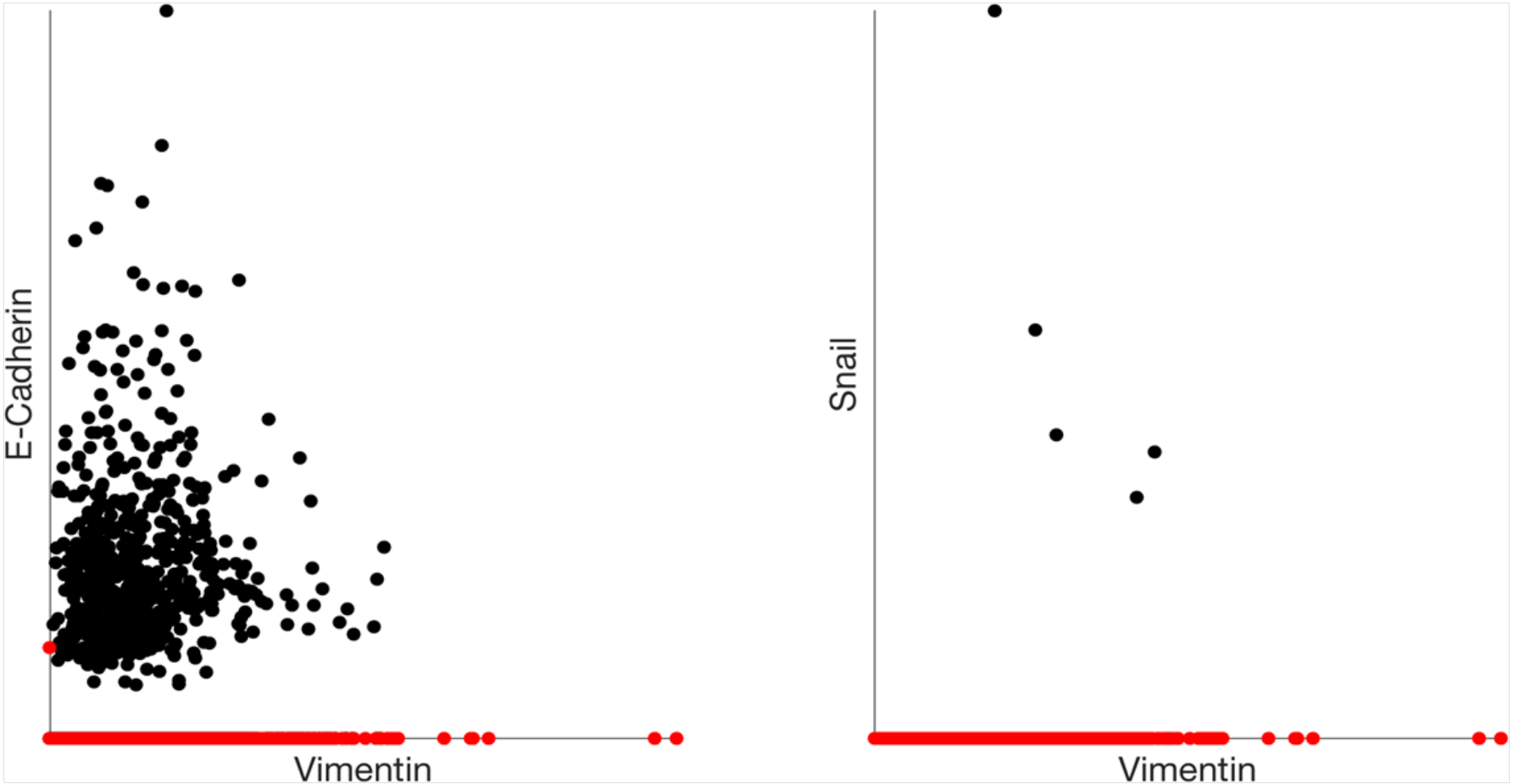
Overview



Imputation & denoising

MAGIC (van Dijk et al., Cell 2018)

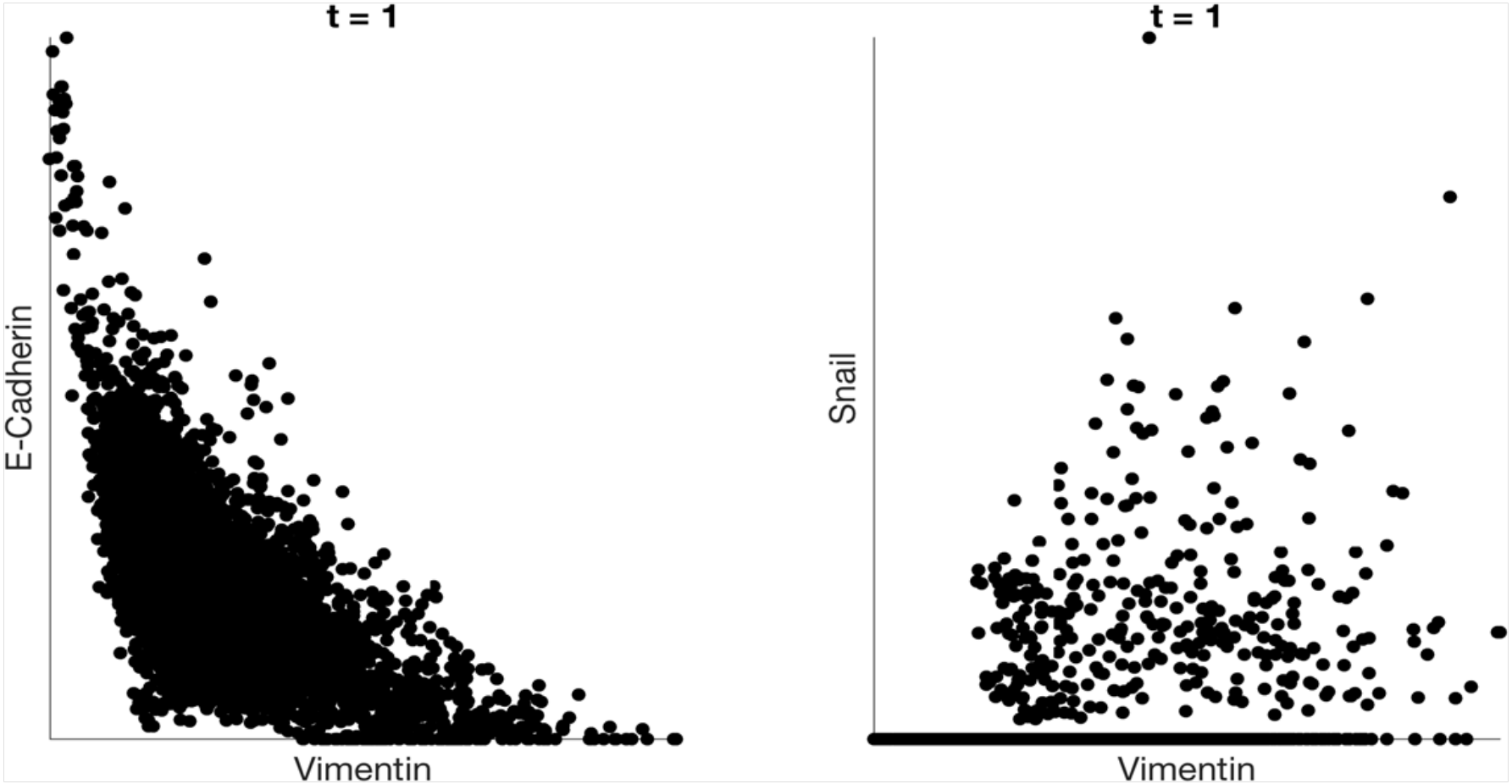
Recovering gene interactions in EMT data



Imputation & denoising

MAGIC (van Dijk et al., Cell 2018)

Recovering gene interactions in EMT data



Understanding diffusion geometry

Harmonic analysis on data manifold / foundations of graph signal processing

The diffusion operator $P^t \longrightarrow$ **heat kernel** $e^{t\Delta}$ when # data points $\to \infty$, neighborhood radius $\to 0$, up to density normalization.

- The eigenvectors / eigenfunctions of P^t form **generalized Fourier harmonics** over the data geometry
- The eigenvalues of P^t take the form of $\boldsymbol{e}^{-\boldsymbol{t}\cdot(\text{frequency})^2}$
- $f(x) \mapsto P^t f(x)$ acts as a lowpass filter

Harmonic analysis interpretation of presented methods:

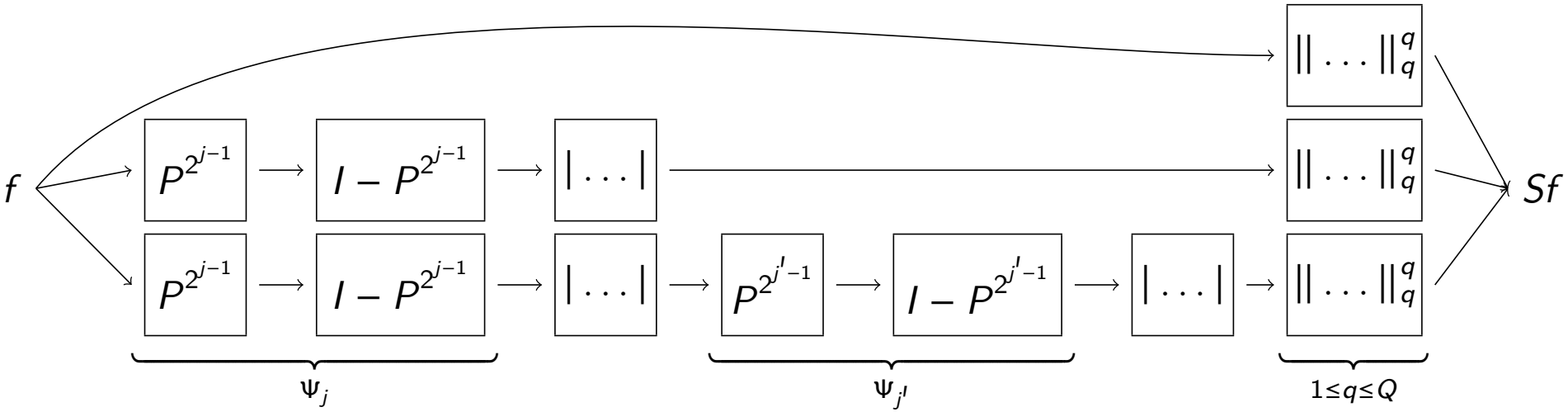
- SUGAR / MAGIC – based on **lowpass filtering** of data features
- PHATE / DM – based on **impulse responses** of lowpass filters

Beyond lowpass - diffusion filters over intrinsic data geometry:

- $f(x) \mapsto (I - P^t)f(x)$ – highpass filter
- $f(x) \mapsto P^t(I - P^t)f(x)$ – bandpass filter / diffusion wavelet

Geometric scattering (Gao et al., ICML 2019)

Deep diffusion-based filter bank for graph / manifold representation



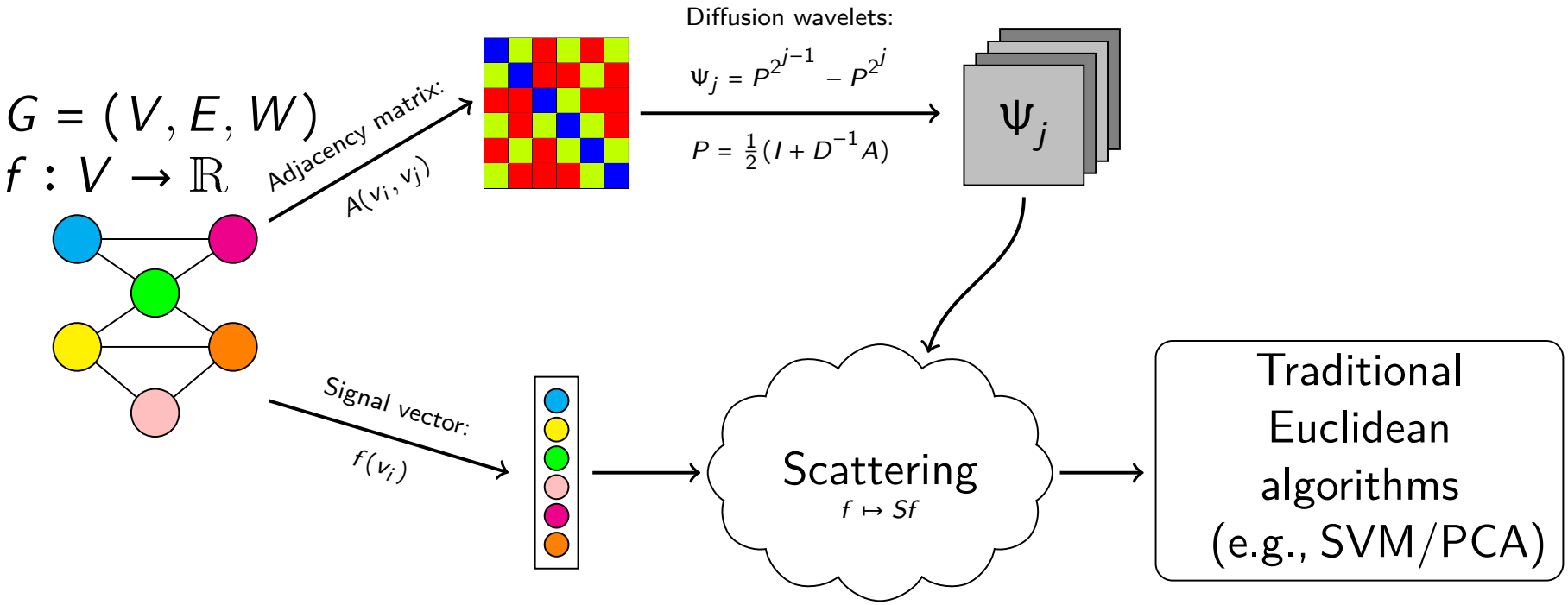
- Provides whole-graph representation for graph data analysis
- Mathematical framework for geometric deep learning
  - Analogous to Euclidean scattering by Mallat (CPAM, 2012)
- New notion of deformation stability using rigid motions & distribution variations on manifolds (Perlmutter et al., NeurIPS DLT workshop 2018)

Geometric scattering

(Gao et al., ICML 2019)

Feature extraction for graph data analysis

Scattering features embed graphs with signals over their vertices to a Euclidean feature space indexed by scattering paths (i.e., j, j', q)

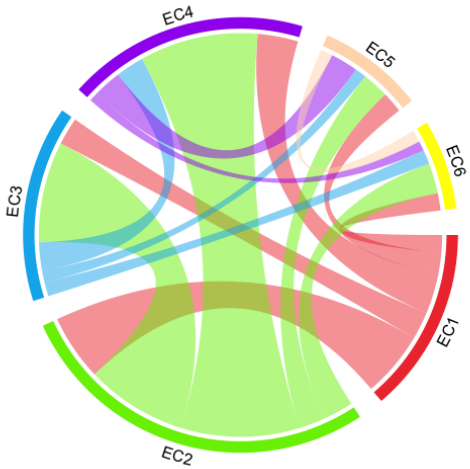


- Multiple signals handled by concatenation of scattering features

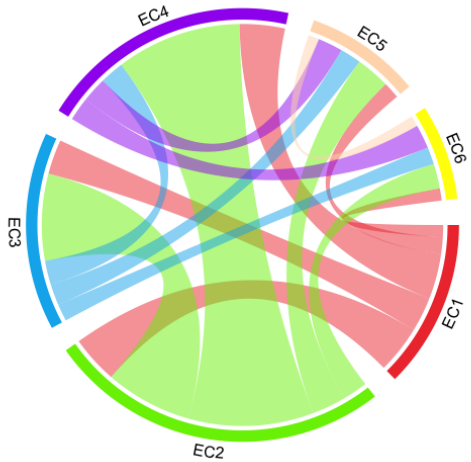
Geometric scattering (Gao et al., ICML 2019)

Example: exploring enzyme class exchange preferences

Inferring EC exchange preferences in enzyme evolution:



Observed by Cuesta et al. (Biophysical Journal, 2015)



Inferred via geometric scattering features

Exchange pref. inference

Compute pref(EC-i, EC-j) :=

$$w_j \cdot \left[\min \left\{ \frac{D(i,j)}{D(i,i)}, \frac{D(j,i)}{D(j,j)} \right\} \right]^{-1}$$

w_j = portion of enzymes in EC-j that choose another EC as their nearest subspace; $D(i,j)$ = mean dist. of enzymes in EC-i from PCA (90% exp. var.) subspace of EC-j.

- Geometric scattering features extracted from ENZYMES (Borgwardt et al., Bioinformatics 2005) containing 100 enzyme graphs from each EC.
- PCA over scattering: EC subspaces of 5–7 dims. ; full space of 16 dims.

Conclusion

Exploratory data analysis, especially in genonmics/proteomics, often requires to separate data geometry from distribution.

Diffusion geometry enables a multitude of tools highly suitable for geometry-based analysis:

- PHATE - data visualization with diffusion geometry
- PhEMD - learning drug perturbation manifold
- SUGAR - geometry-based data generation
- Geometric scattering - graph/manifold-level representations

Additional work includes, for example:

- MAGIC - data imputation & denoising (van Dijk et al., 2018)
- Data fusion with harmonic alignment (Stanley et al., 2019)

Acknowledgements

Yale Applied Math Program:
Ofir Lindenbaum
Jay Stanley
Raphy Coifman

Yale School of Medicine:
Smita Krishnaswamy
Scott Gigante
Will Chen
Dan Burkhardt
David van Dijk
Zheng Wang
Natalia Ivanova

Utah State University:
Kevin Moon

Michigan State University:
Matt Hirn
Fong Gao
Mike Perlmutter

University of Zurich:
Nevena Zivanovic
Bernd Bodenmiller

Memorial Sloan Kettering Cancer Center:
Dana Pe'er
Roshan Sharma

Funding: