Generative modeling on the space of empirical measures

Nic Fishman and Gokul Gowri

2025.10.17

roadmap

1. ***distribution encoders:*** achieving a CLT for encoders

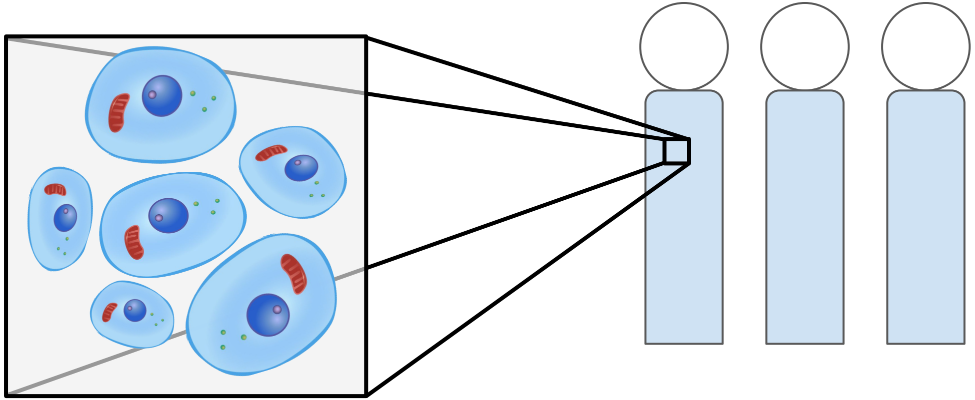
2. ***unsupervised distribution embeddings:*** distribution representations using conditional generators

3. ***any-to-any transport:*** learning transport maps between any pair of distributions, using dist. encoders

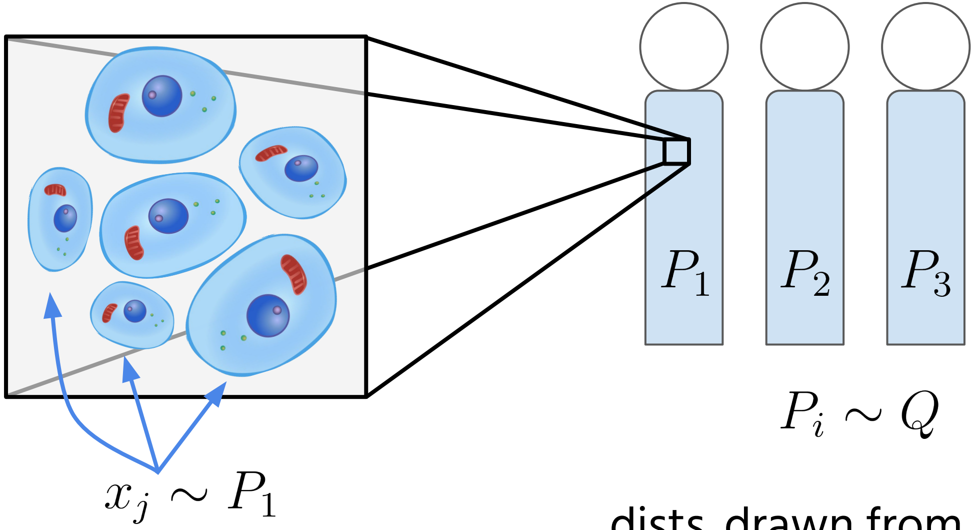
4. ***outlook:*** what's next?

distribution embeddings

many complex systems are *multiscale*



many complex systems are *multiscale*

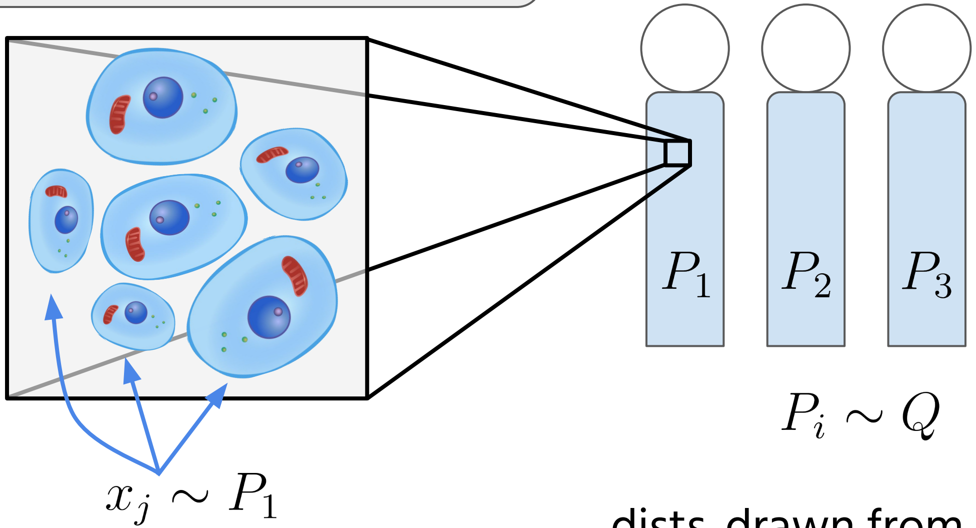


sets of samples
drawn from a dist.

dists. drawn from a metadist.

many complex systems are *multiscale*

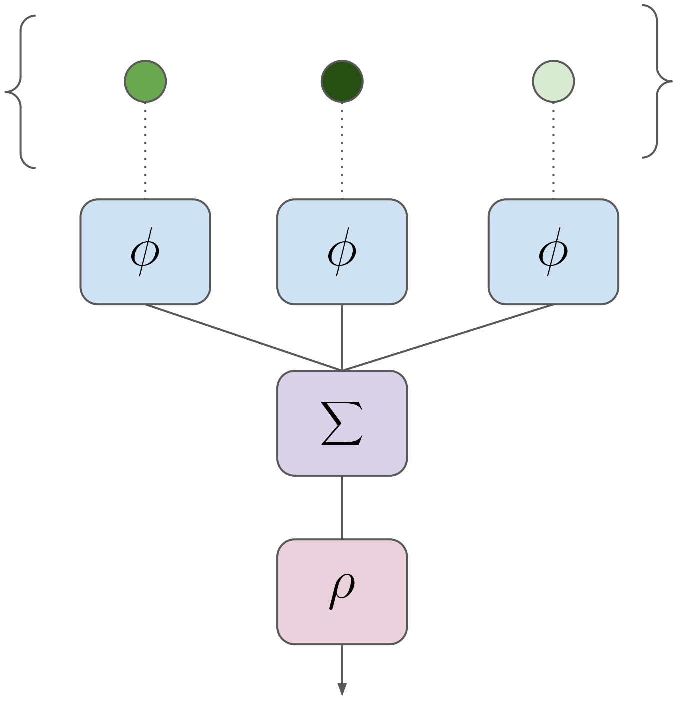
how do we get an embedding of $\boldsymbol{P}$ from samples $\boldsymbol{x} \sim \boldsymbol{P}$?



sets of samples
drawn from a dist.

dists. drawn from a metadist.

deep sets (and related) offer tools for representing sets



Zaheer et al., 2017

what properties do we want for a distribution encoder?

ideally, we would like a central limit theorem for an encoder

what properties do we want for a distribution encoder?

ideally, we would like a central limit theorem for an encoder

for a sample set $S = \{x_i\}_{i=1}^m$ where $x_i \sim P$

we want the encoding to concentrate around its population value:

$$\sqrt{m}(\mathcal{E}(S) - \phi(P)) \xrightarrow{d} \mathcal{N}(0, \Sigma)$$

what properties do we want for a distribution encoder?

ideally, we would like a central limit theorem for an encoder

for a sample set $S = \{x_i\}_{i=1}^{m}$ where $x_i \sim P$

we want the encoding to concentrate around its population value:

$$\sqrt{m}(\mathcal{E}(S) - \phi(P)) \xrightarrow{d} \mathcal{N}(0, \Sigma)$$

so that any plug-in loss becomes an unbiased, asymptotically normal estimator

$$\sqrt{m}\,\left(\ell\left(\mathcal{E}(S)\right) - \ell\left(\phi(P)\right)\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

distributional invariance enables a CLT for encoders

we call an encoder satisfying the following properties *distributionally invariant*

1. ***permutation invariance:*** reordering samples does not change the embedding

$$\mathcal{E}(S) = \mathcal{E}(\pi(S))$$

2. ***proportional invariance:*** duplicating every sample *K* times does not change the embedding

$$\mathcal{E}(S) = \mathcal{E}(\cup_{k=1}^{K} S)$$

3. ***smooth pooling:*** Hadamard differentiability of the pooling operator

embeddings are normally distributed

we run the following experiment:

with a mean-pooled GNN

fix a distribution P_i

for different set sizes *m*

1. sample many size *m* sets
2. visualize embeddings

we see normality as m becomes large

embeddings are normally distributed

we run the following experiment:

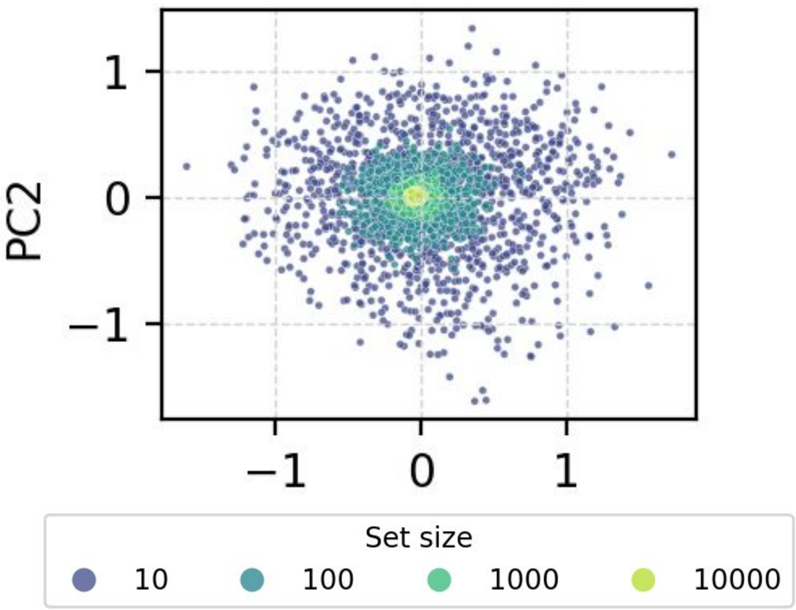
with a mean-pooled GNN

fix a distribution P_i

for different set sizes *m*

1. sample many size *m* sets
2. visualize embeddings

we see normality as m becomes large



what about when we have additional, unlabeled data?

we have shown how to build distribution encoders

but often we have a small amount of labeled and a large amount of unlabeled data.

how can we build unsupervised distribution embeddings?

unsupervised distribution embeddings

generative distribution embeddings

intuition: the ideal distribution embedding should provide enough information to generate samples from the corresponding distribution

$$\mathcal{G}(\mathcal{E}(S_{i,m})) \xrightarrow{d} P_i \quad \text{as} \quad m \to \infty$$

implementing generative distribution embeddings

to build a GDE, all you need to do is combine

1. a distributionally invariant encoder

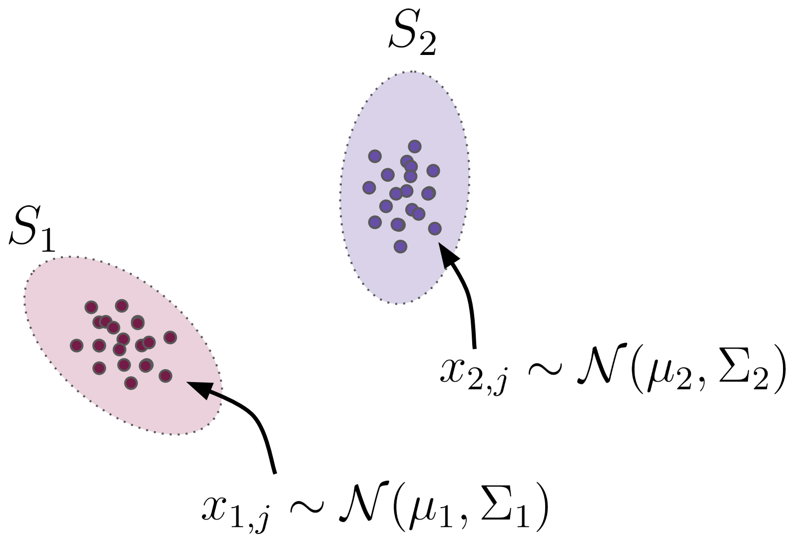
    e.g., a mean-pooled GNN, a mean-pooled self-attention

2. a conditional generator

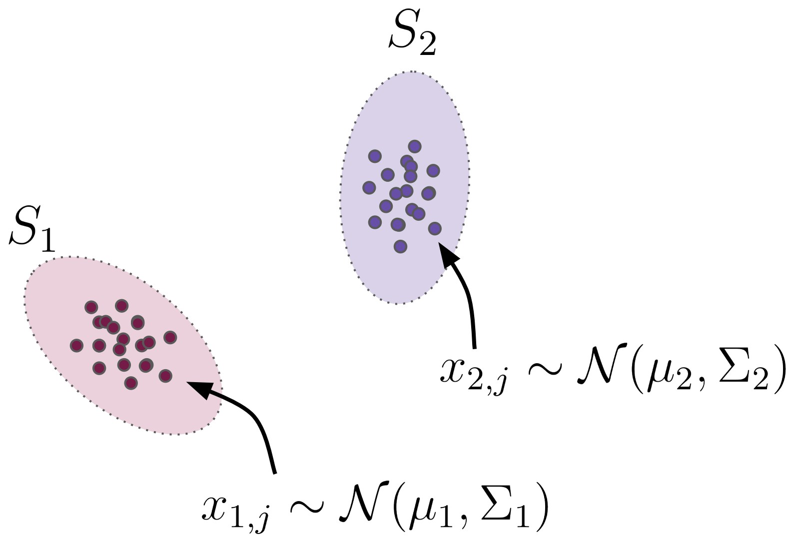
    e.g., a diffusion model, an autoregressive LLM

and train end-to-end using the conditional generator loss

a toy example with Gaussians

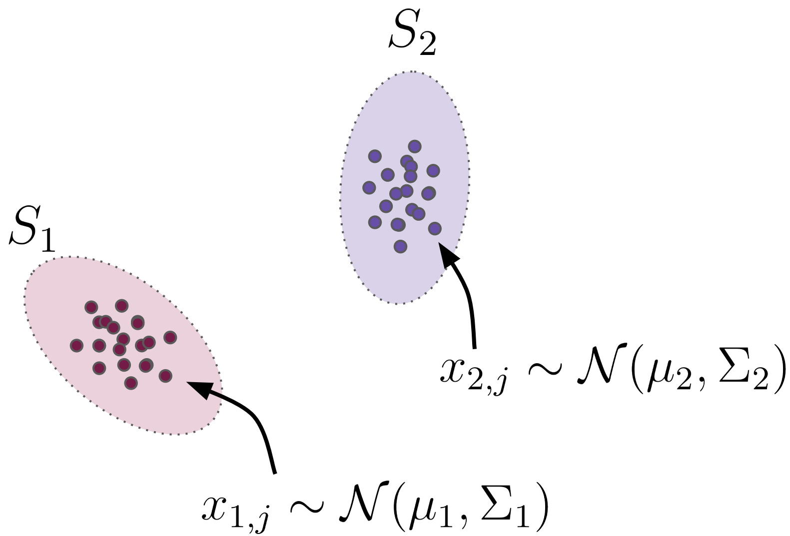


a toy example with Gaussians



$$\mu_i \sim \mathrm{Unif}([0,5]^d)$$
$$\Sigma_i \sim \mathcal{W}^{-1}(\Phi, \nu)$$
$$x_{ij} \sim \mathcal{N}(\mu_i, \Sigma_i)$$
$$\mathcal{D} = \left\{ S_i = \{x_{ij}\}_{j=1}^{m} \right\}_{i=1}^{n}$$

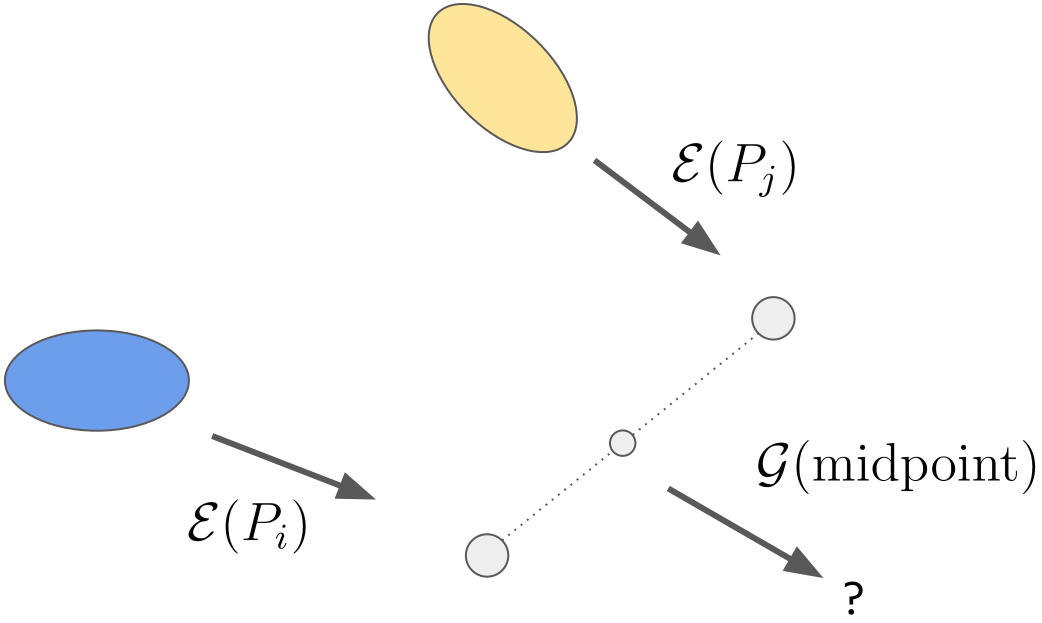
a toy example with Gaussians



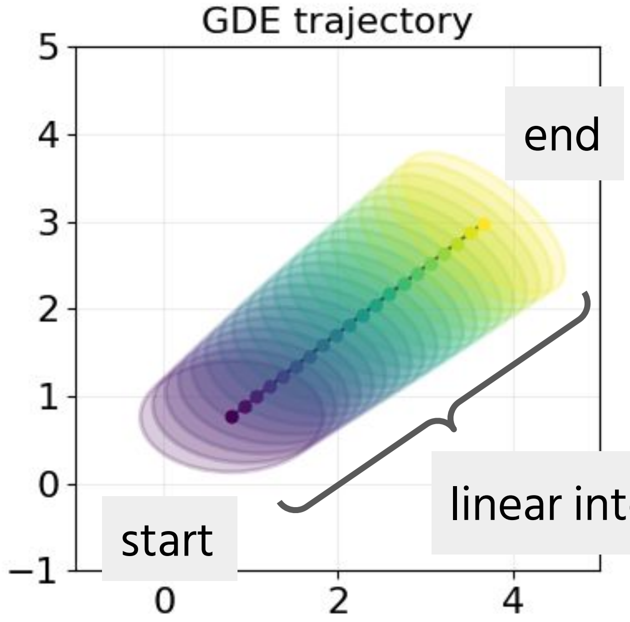
$$\mu_i \sim \mathrm{Unif}([0,5]^d)$$
$$\Sigma_i \sim \mathcal{W}^{-1}(\Phi, \nu)$$
$$x_{ij} \sim \mathcal{N}(\mu_i, \Sigma_i)$$
$$\mathcal{D} = \left\{ S_i = \{x_{ij}\}_{j=1}^{m} \right\}_{i=1}^{n}$$

we will learn representations of distributions using
encoder: a mean-pooled deep sets
generator: conditional diffusion

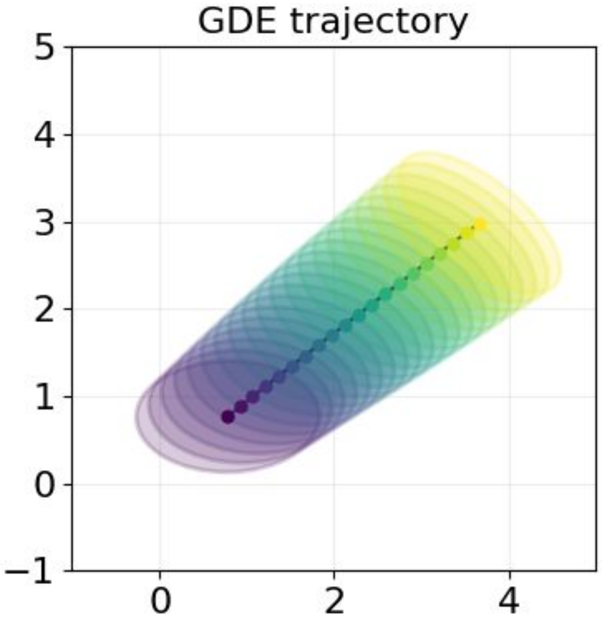
probing the geometry of the latent space



probing the geometry of the latent space

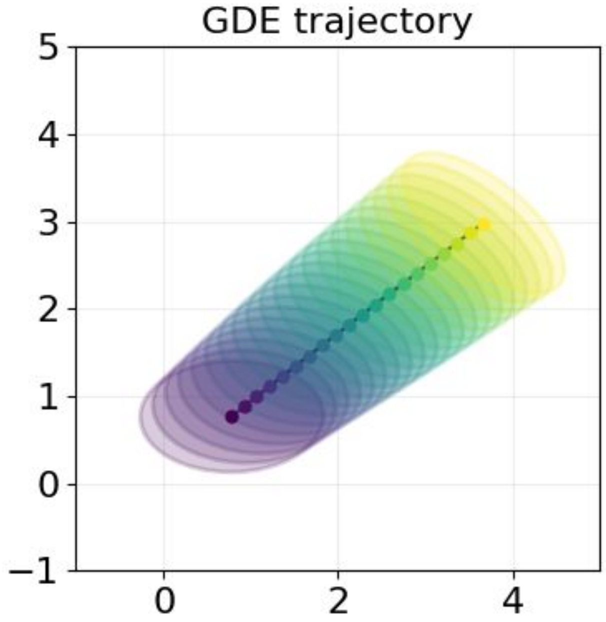


probing the geometry of the latent space



this is so pretty

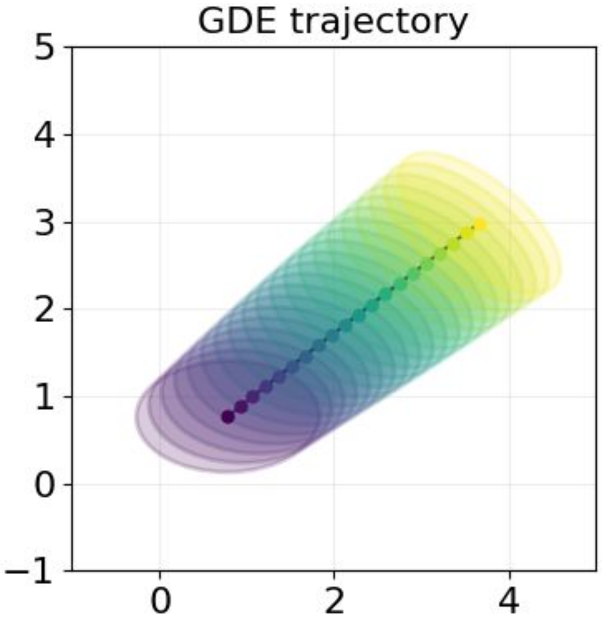
probing the geometry of the latent space



this is so pretty

maybe it's like, math, somehow

probing the geometry of the latent space

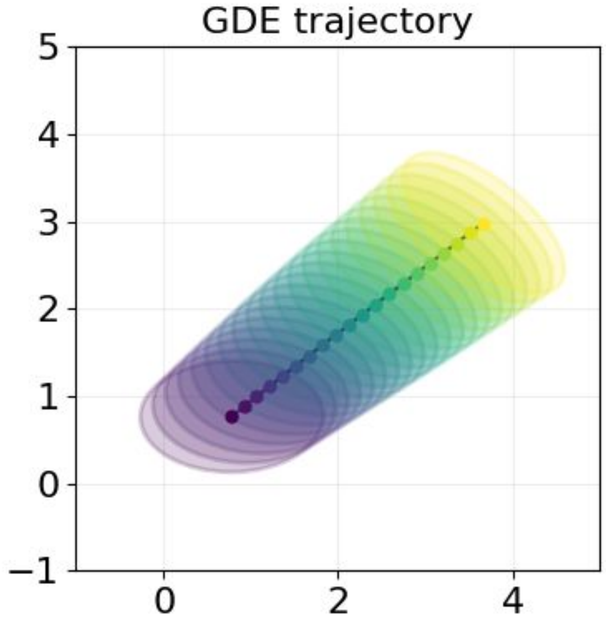


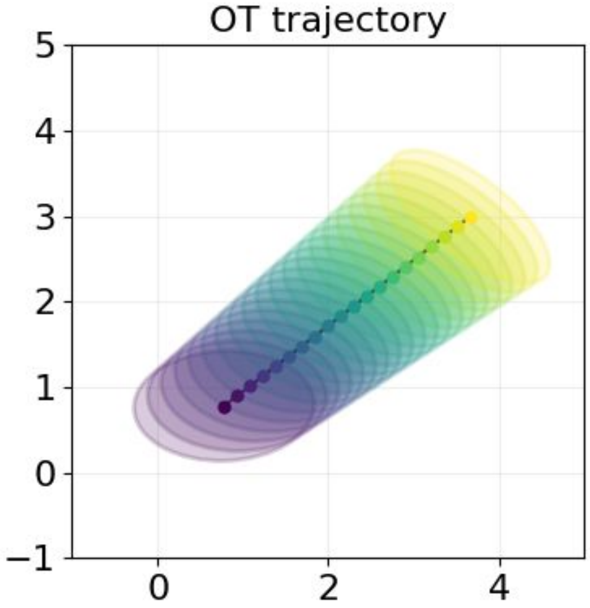
this is so pretty

maybe it's like, math, somehow

it is! this is almost exactly the optimal transport under the wasserstein 2 distance

probing the geometry of the latent space





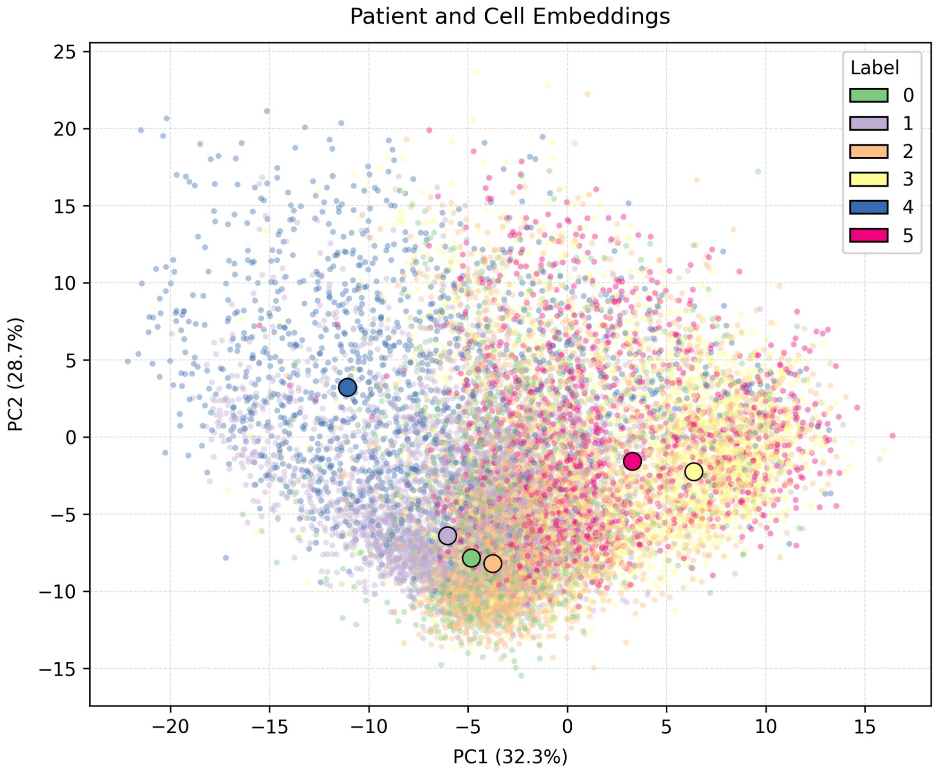
notes on the geometry of GDE latent spaces

1. GDE interpolants are not exactly optimal transport interpolants in general, in practice they are OT-like but restricted to a family of distributions

2. GDE latent spaces can be warped by metadistribution-induced weighting on the statistical manifold

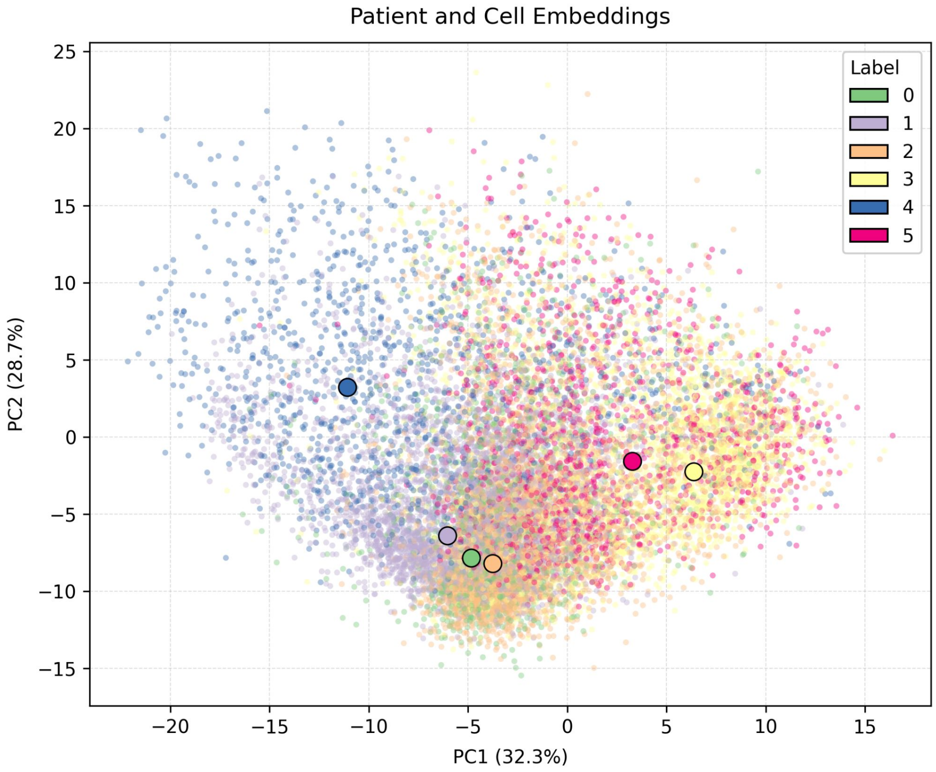
3. while GDE interpolants resemble OT interpolants at the population level, they don't directly provide information about unit level trajectories

example: donor-level representations from cell-level data



data from Fullard et al., 2025

example: donor-level representations from cell-level data

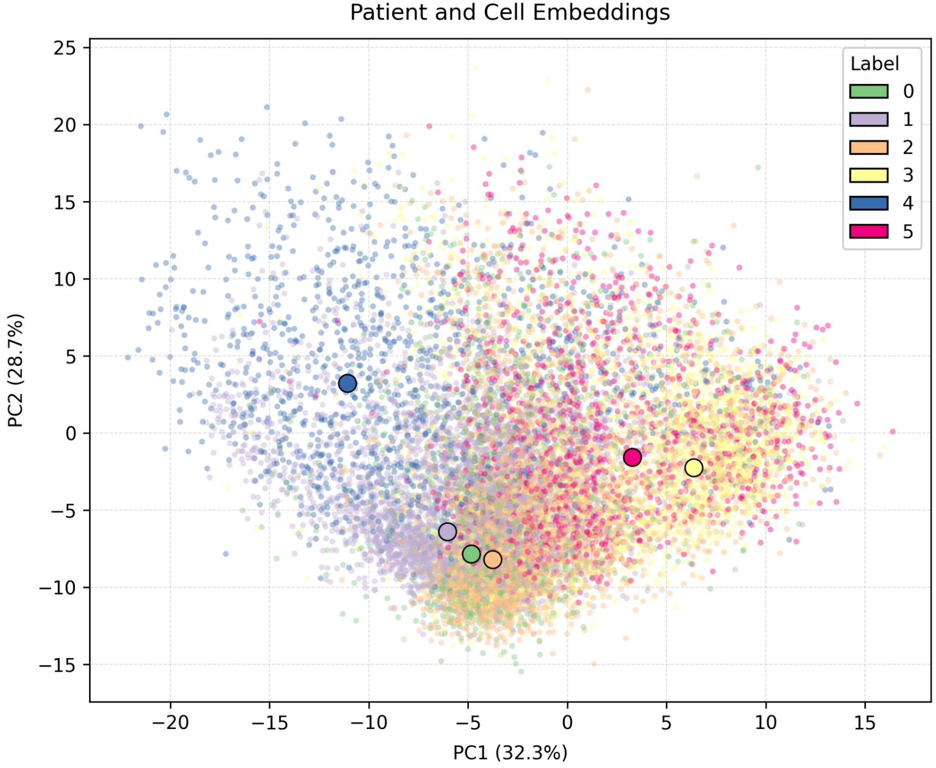


We construct a model to predict donor attributes based on single cells.

For the supervised task we use 10% of the donors, for the semisupervised we use that 10% as labeled data and the remaining 90% unlabeled.

data from Fullard et al., 2025

example: donor-level representations from cell-level data



We construct a model to predict donor attributes based on single cells.

For the supervised task we use 10% of the donors, for the semisupervised we use that 10% as labeled data and the remaining 90% unlabeled.

Metric	Semi-Supervised	Supervised
Accuracy	**0.8887**	0.8791
Balanced Accuracy	**0.5383**	0.5291
Roc Auc	**0.5131**	0.4872
F1 Score	**0.1479**	0.1293

data from Fullard et al., 2025

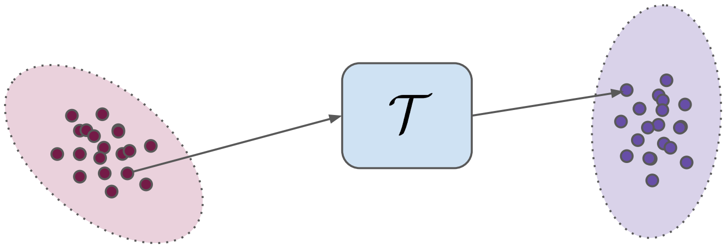
GDEs are applicable in many real world settings

GDEs can also be useful for modelling...

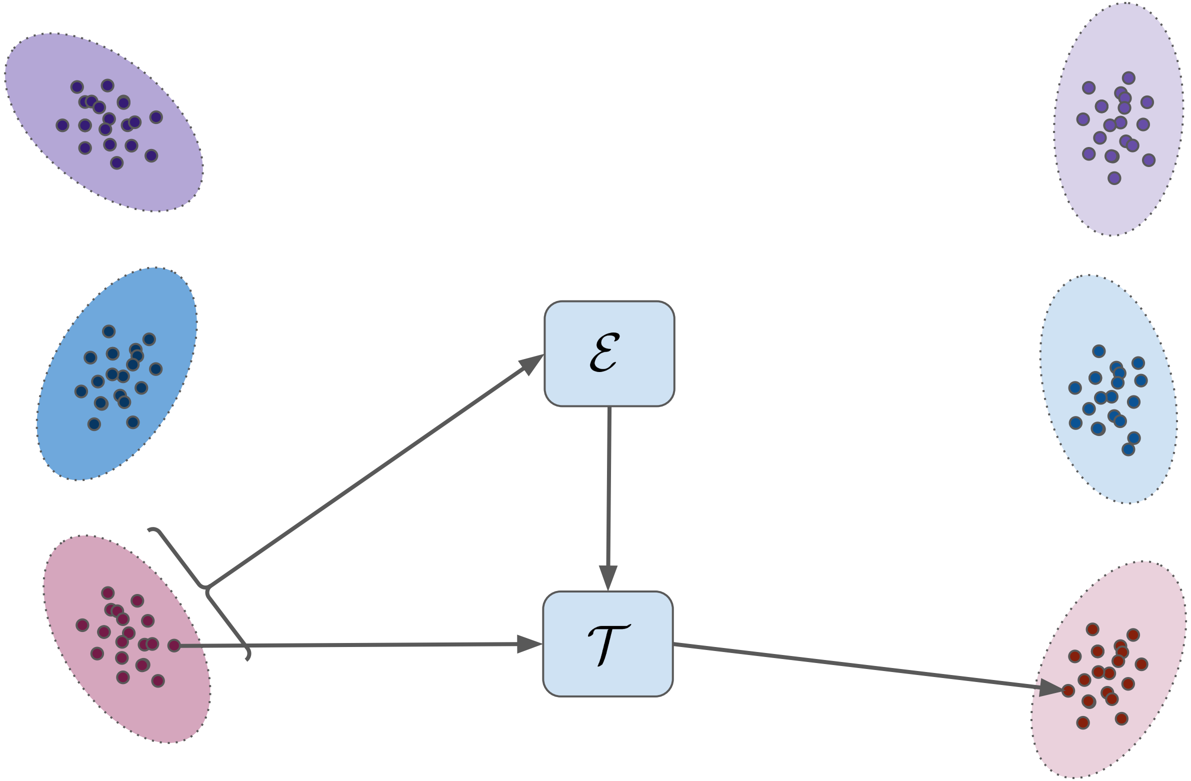
- high throughput genetic perturbation screens
- clonal properties in lineage traced scRNA-seq data
- promoter expression screens
- viral sequence evolution

see Generative Distribution Embeddings, arXiv; NeurIPS 2025 transport

many models transport from one distribution to another



Goodfellow et al., 2014; Genevay et al., 2018; Lipman et al., 2022; Albergo et al., 2022 source conditioning provides tools for simultaneously solving many transport problems



Atanackovic et al., 2024

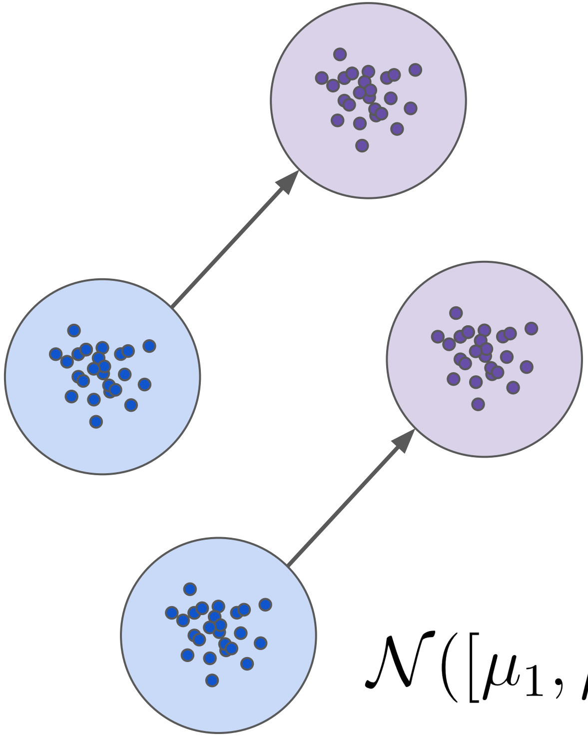
what if we want to make use of unpaired data?

source-conditioning is applicable only when we have explicit source-target pairs

what if we have access to many "orphan" marginals without source/target labels?

this happens in real data (we'll show some examples)

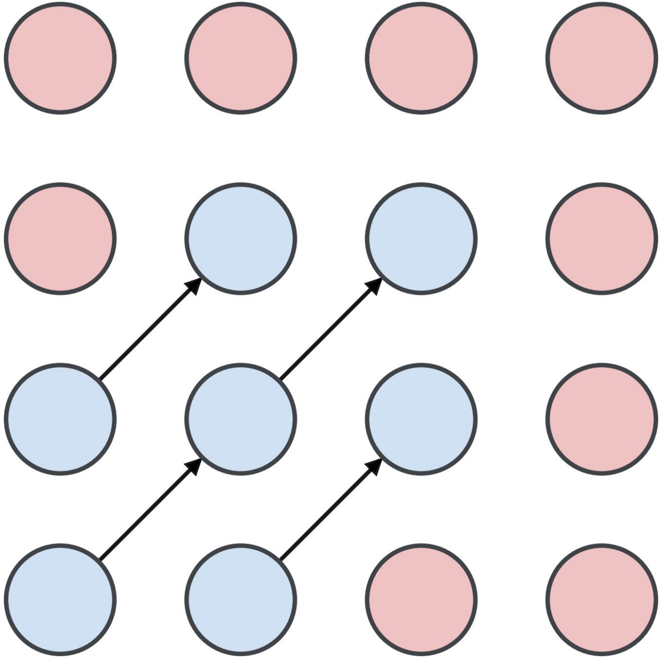
a toy example with Gaussians



$$\mathcal{N}([\mu_1 + 3, \mu_2 + 3], I_2)$$

$$\mathcal{N}([\mu_1, \mu_2], I_2)$$

a toy example with gaussians



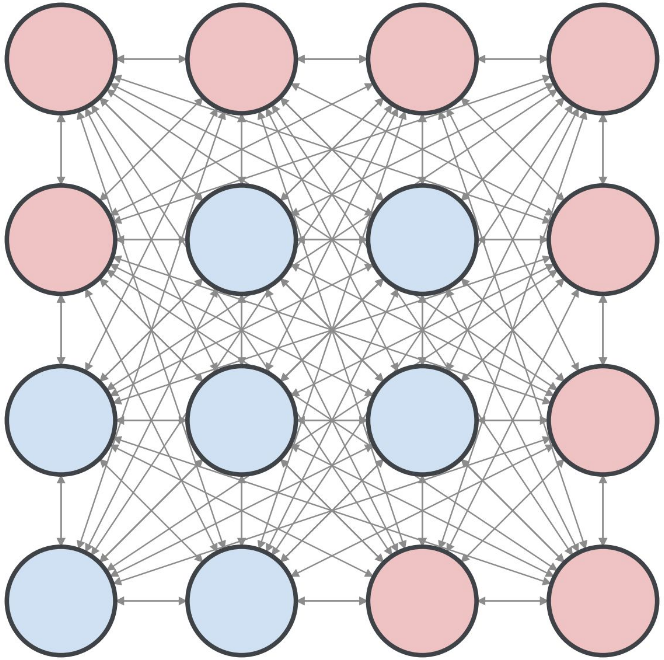
source-target pairing observed for

$$\mu_1, \mu_2 \in [0, 3]$$

orphan marginals observed for

$$\mu_1, \mu_2 \in [3, 5]$$

a toy example with gaussians



1. use all pairings to train an "unsupervised" any-to-any model

2. train a small "latent predictor" to predict the target latent given the source latent

transport with distribution embeddings

goal: learn a model which can transport between any pair of distributions by conditioning on source and target distribution embeddings

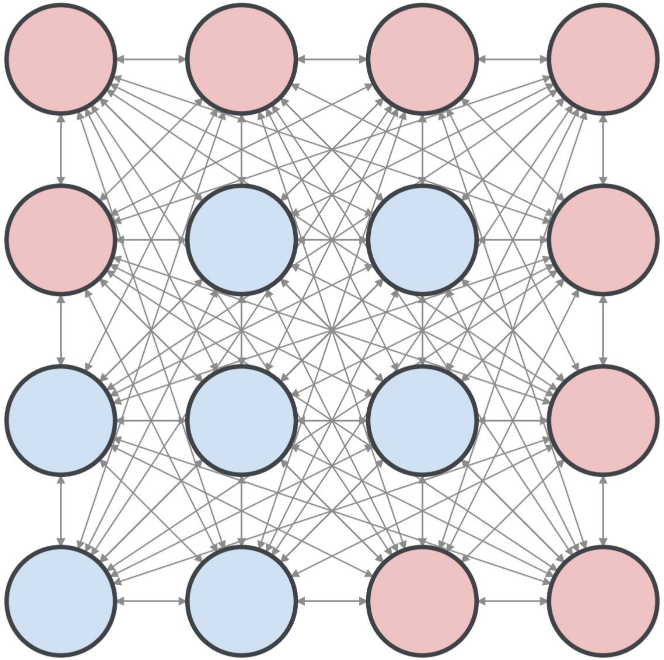
we sample two distributions and two sets of samples

$$P_i \sim Q \quad x_{jk} \sim P_j \quad S_{i,m} = \{x_{ik}\}_{k=1}^m$$
$$P_j \sim Q \quad x_{ik} \sim P_i \quad S_{j,m} = \{x_{jk}\}_{k=1}^m$$

our goal is to learn an any-to-any transport map

$$\mathcal{T}\big(S_{i,m},\, \mathcal{E}(S_{i,m}),\, \mathcal{E}(S_{j,m})\big) \xrightarrow{d} P_j \quad \text{as } m \to \infty.$$

a toy example with gaussians

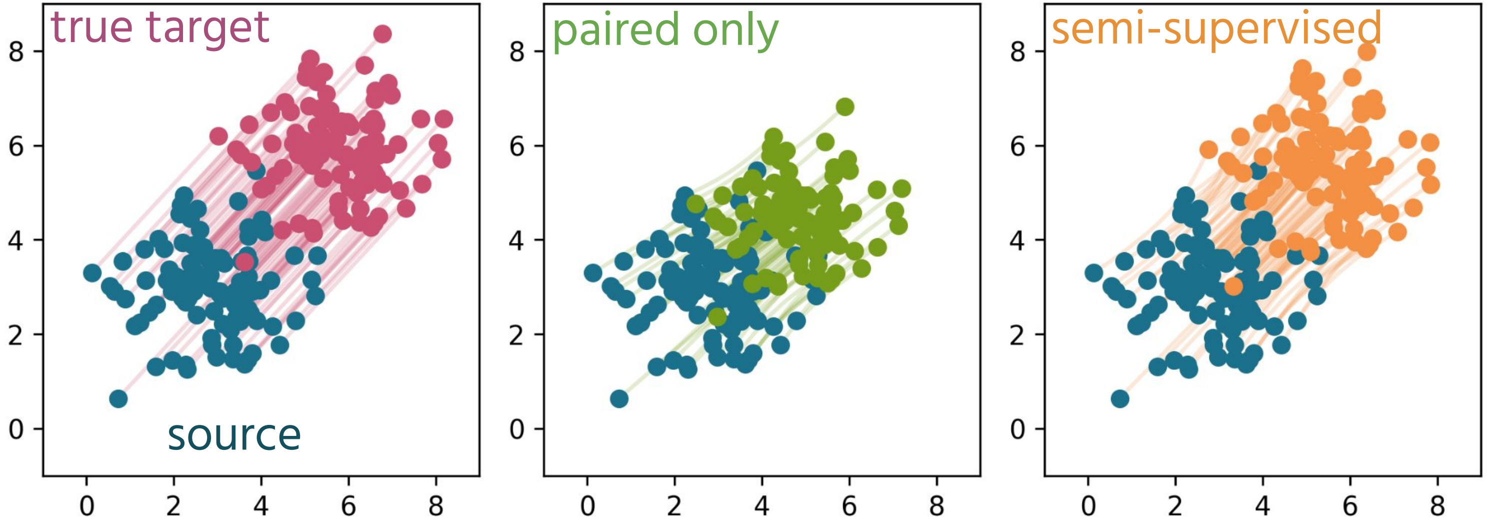


1. use all pairings to train an "unsupervised" any-to-any model

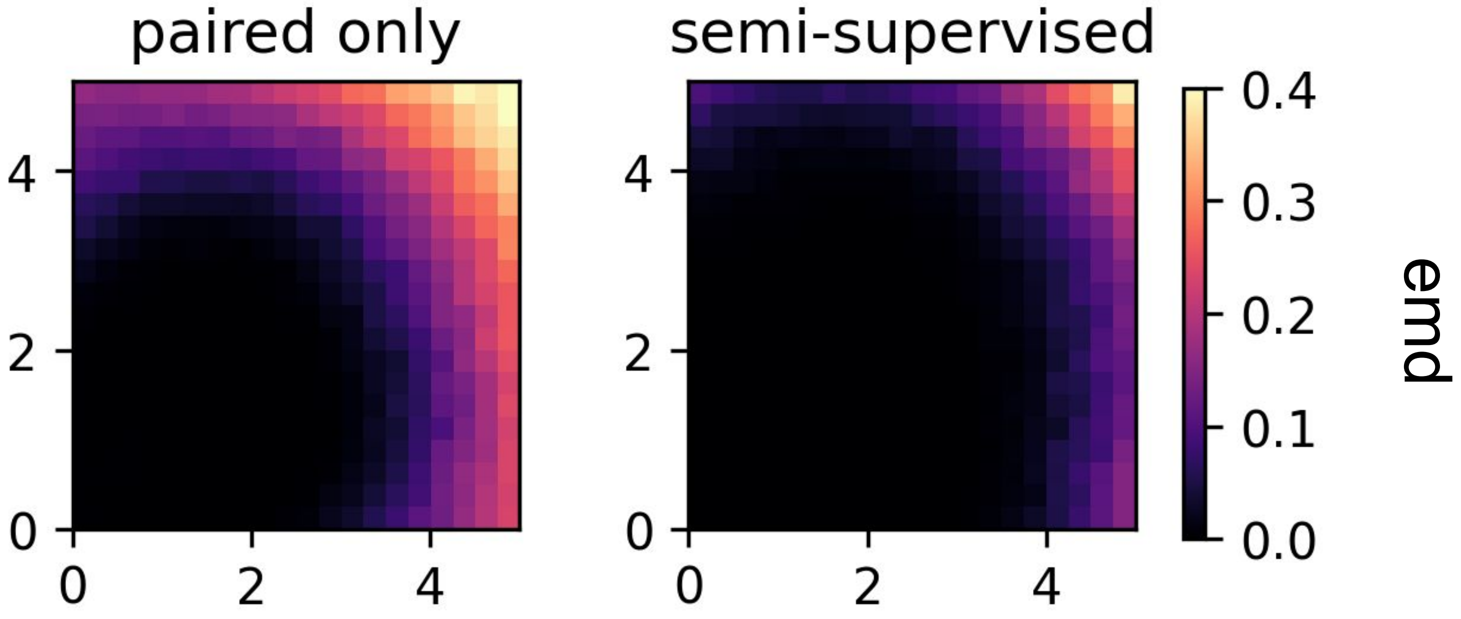
2. train a small "latent predictor" to predict the target latent given the source latent

we learn any-to-any transport using
encoder: mean pooled deep sets
transport model: flow matching

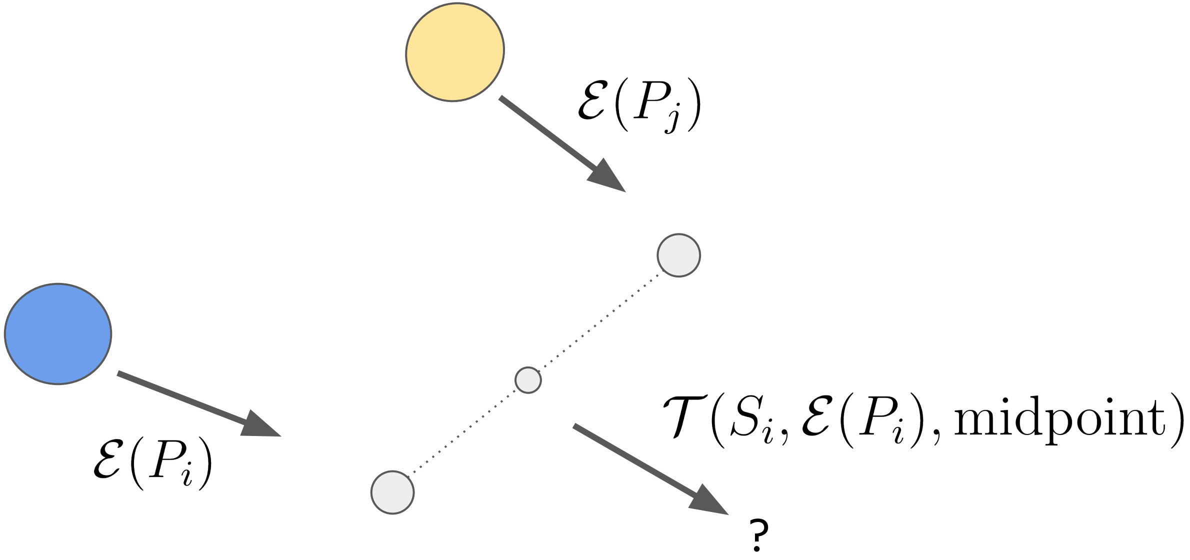
TDEs rescue performance outside of paired data distribution



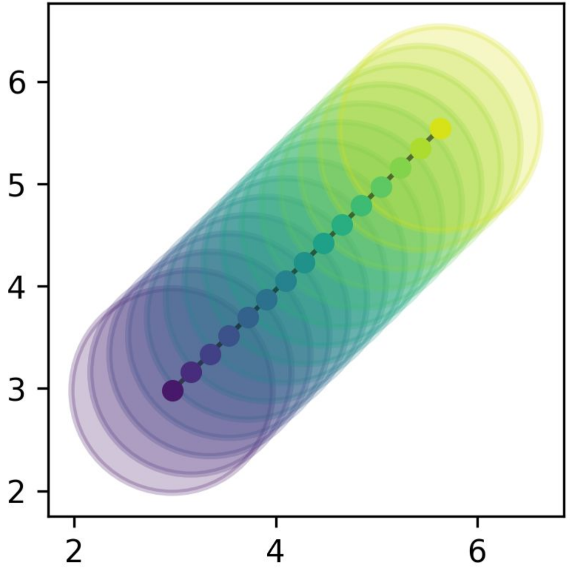
TDEs rescue performance outside of paired data distribution



probing the latent geometry (again)



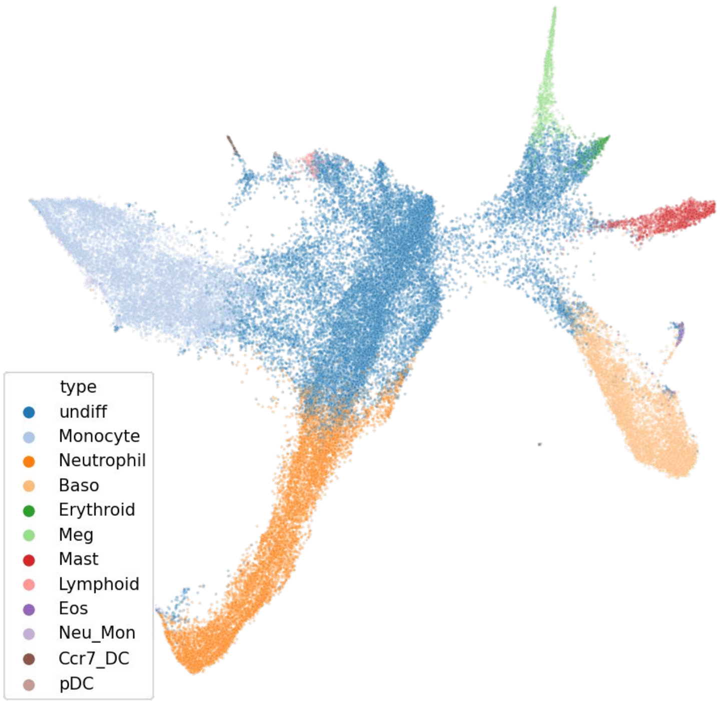
probing the latent geometry (again)



and again, we closely match the OT

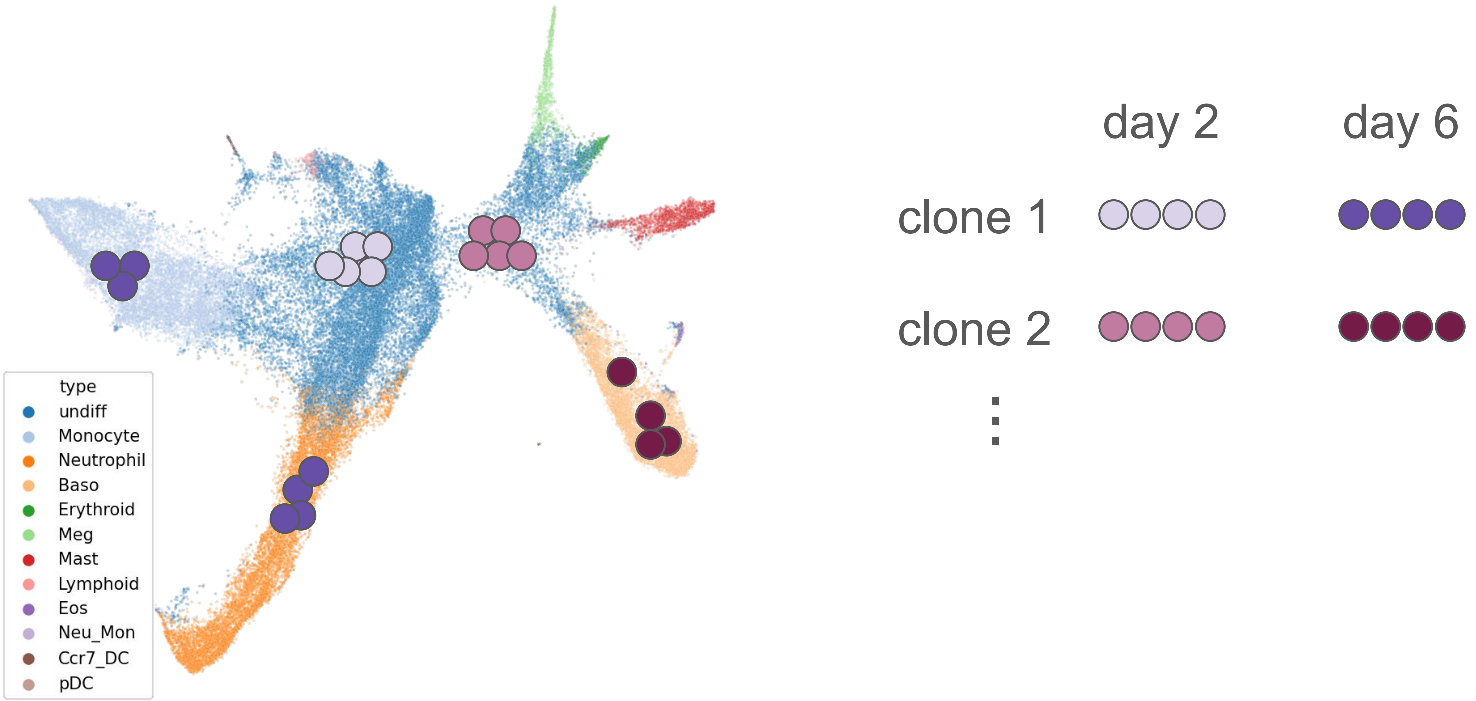
but this time with unit-level trajectories!

Lineage-traced scRNAseq measures dynamics of clones



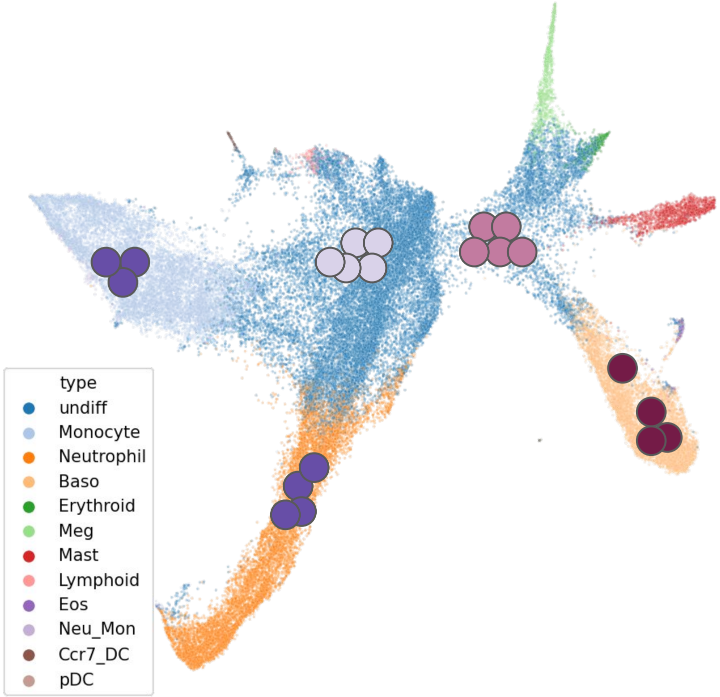
lineage traced scRNA-seq data from Weinreb et al., 2020

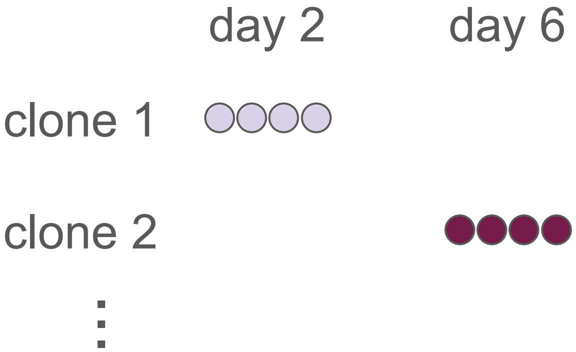
Lineage-traced scRNAseq measures dynamics of clones



lineage traced scRNA-seq data from Weinreb et al., 2020

Lineage-traced scRNAseq measures dynamics of clones

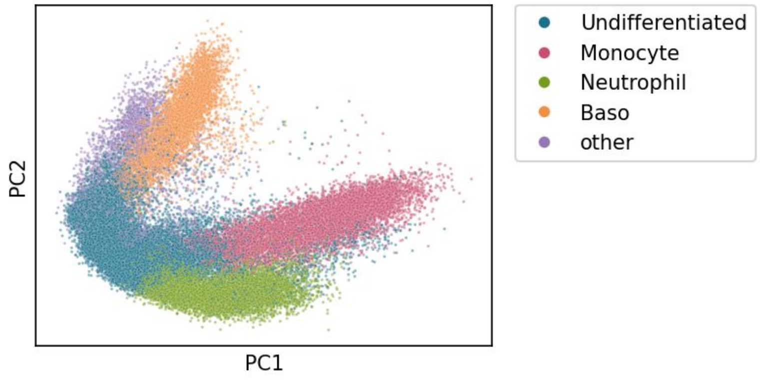


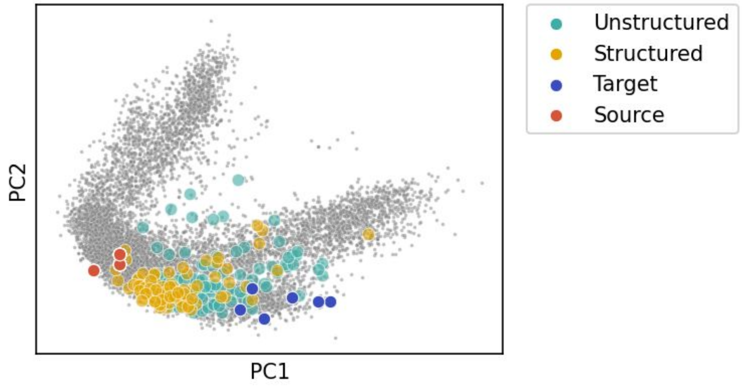


~90% of clones appear at only one timepoint!

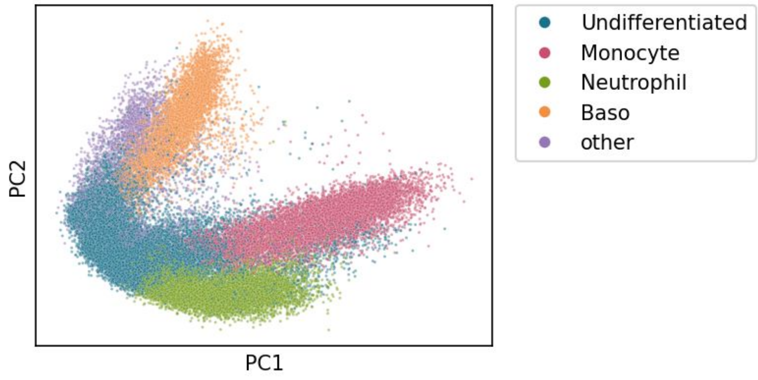
lineage traced scRNA-seq data from Weinreb et al., 2020

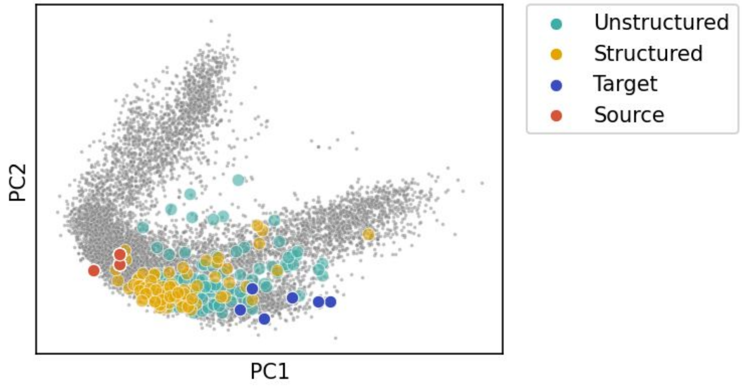
TDEs improve performance by using "orphan" clones

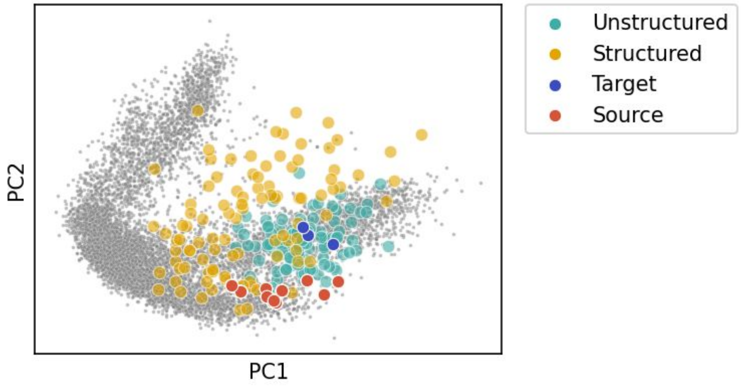




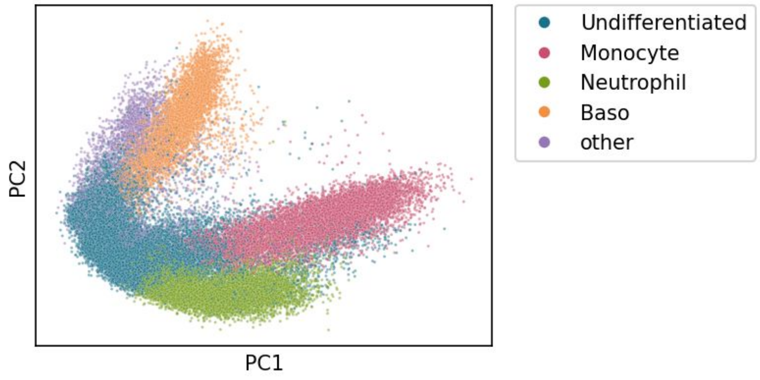
TDEs improve performance by using "orphan" clones

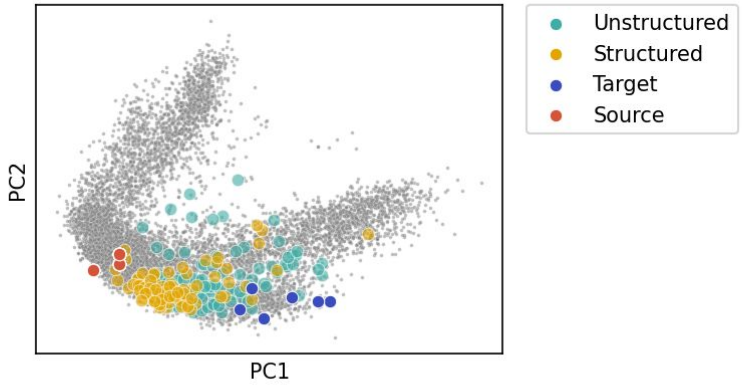




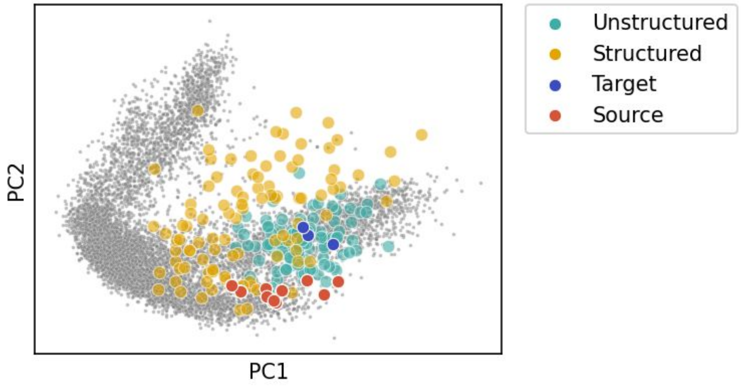


TDEs improve performance by using "orphan" clones





Method	Energy Distance
Source-conditioned	3.34
TDE	**3.20**



some things we're curious about

- applying these tools to understand cellular dynamics
  - in development and in response to perturbation

some things we're curious about

- applying these tools to understand cellular dynamics
  - in development and in response to perturbation
- can we theoretically understand the OT-like geometry of the latent spaces?

some things we're curious about

- applying these tools to understand cellular dynamics
  - in development and in response to perturbation
- can we theoretically understand the OT-like geometry of the latent spaces?
- what is the role of the choice of joint meta-distribution (i.e. distribution pairing)?

some things we're curious about

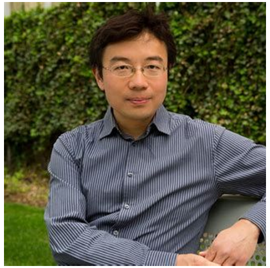
- applying these tools to understand cellular dynamics
    - in development and in response to perturbation
- can we theoretically understand the OT-like geometry of the latent spaces?
- what is the role of the choice of joint meta-distribution (i.e. distribution pairing)?
- what is the relationship between the complexity of a family of distributions and the resources required for a model to learn any-to-any transport?

some things we're curious about

- applying these tools to understand cellular dynamics
    - in development and in response to perturbation
- can we theoretically understand the OT-like geometry of the latent spaces?
- what is the role of the choice of joint meta-distribution (i.e. distribution pairing)?
- what is the relationship between the complexity of a family of distributions and the resources required for a model to learn any-to-any transport?
- can we build a foundation model for transport between any and all distributions? who wants to give us a trillion dollars to try

thank you!

Paolo
Fischer



Peng
Yin

Omar
Abudayyeh

Jonathan
Gootenberg

the workshop organizers
and
all of you :)