Connecting the dots:

A probabilistic model for biomolecular latent space trajectories

Mittelerde Meeting 2025

Andreas Kröpelin

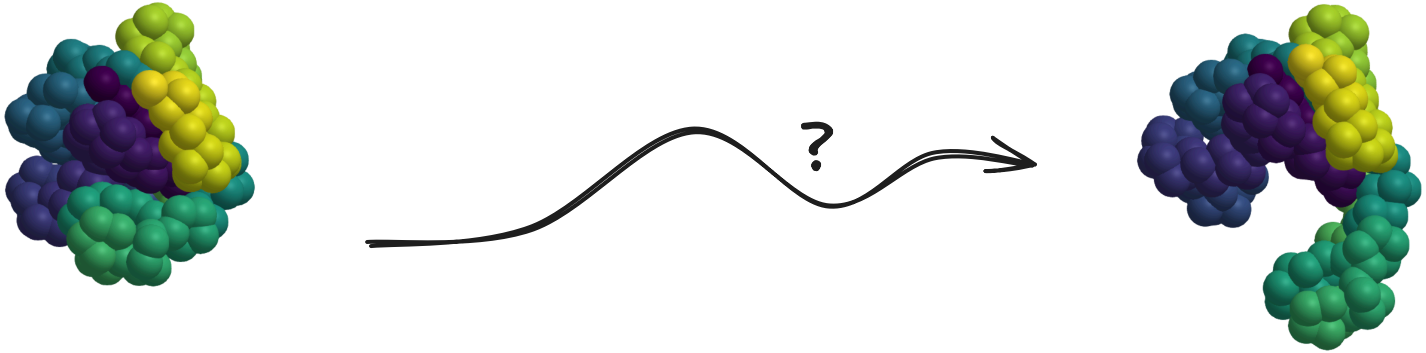
Microscopic Image Analysis Group

University Hospital Jena   Interactive Inference   CRC 1456

Conformational Dynamics of Biomolecules

- want to understand *molecular machines*

- can observe individual conformations



(Adenylate kinase, 1ake ↔ 4ake)

- want to find *continuous dynamic* of conformational change

  (in some *latent space*)

Mathematical abstraction

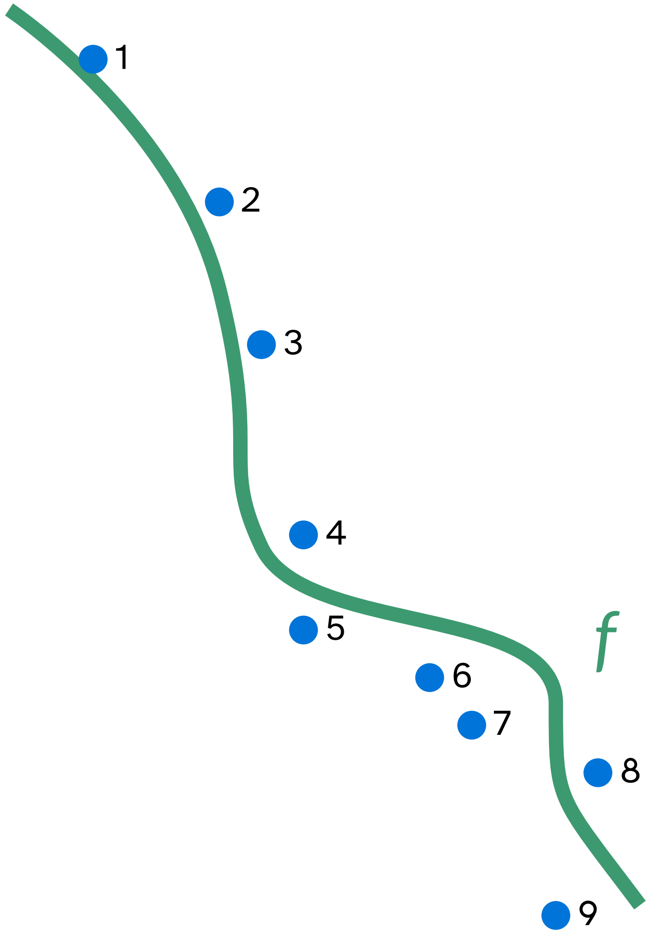
we have

observations $\mathcal{Y} = \{y_1, \ldots, y_m\} \subset \mathbb{R}^d$

and want to explain them with a

curve $f : [0, 1] \to \mathbb{R}^d$

by the magic of ✨ Bayes ✨



Mathematical abstraction

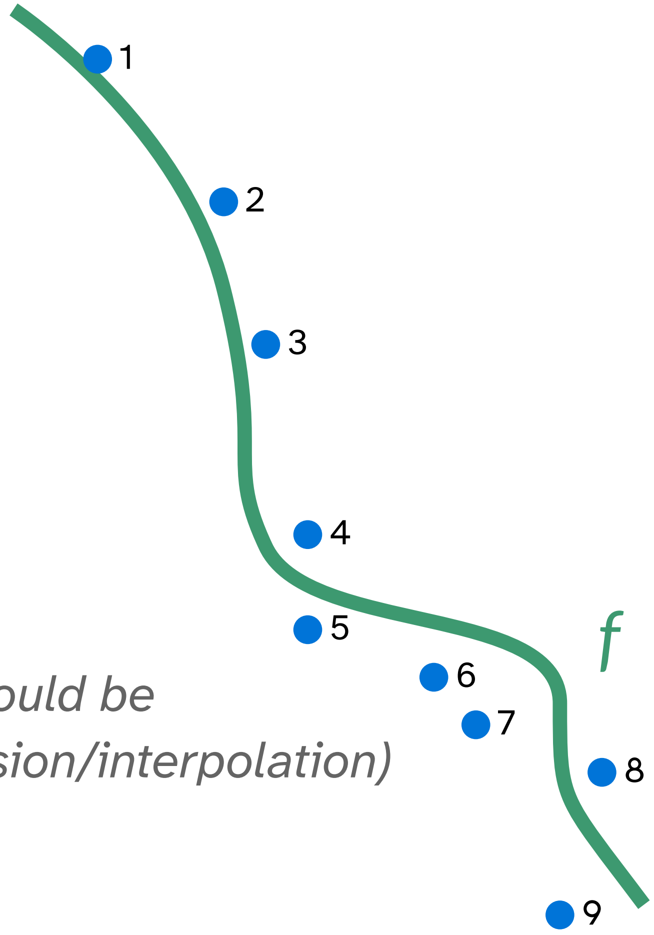
we have

observations $\mathcal{Y} = \{y_1, \ldots, y_m\} \subset \mathbb{R}^d$

and want to explain them with a

curve $f : [0, 1] \to \mathbb{R}^d$

by the magic of ✨ Bayes ✨



(this would be regression/interpolation)

Mathematical abstraction

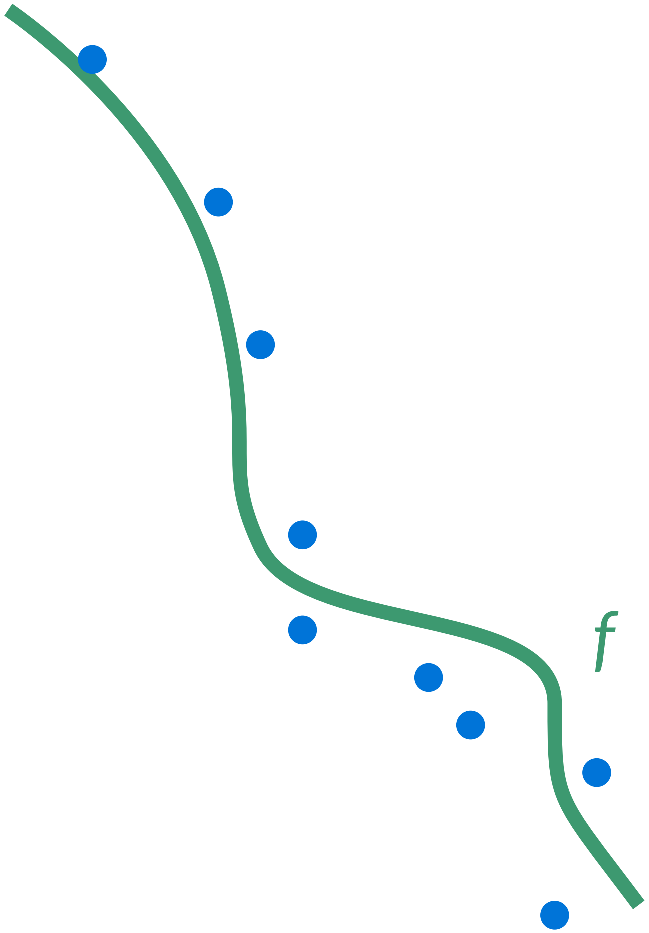
we have

unordered observations $\mathcal{Y} = \{y_1, \dots, y_m\} \subset \mathbb{R}^d$

and want to explain them with a

curve $f : [0, 1] \to \mathbb{R}^d$

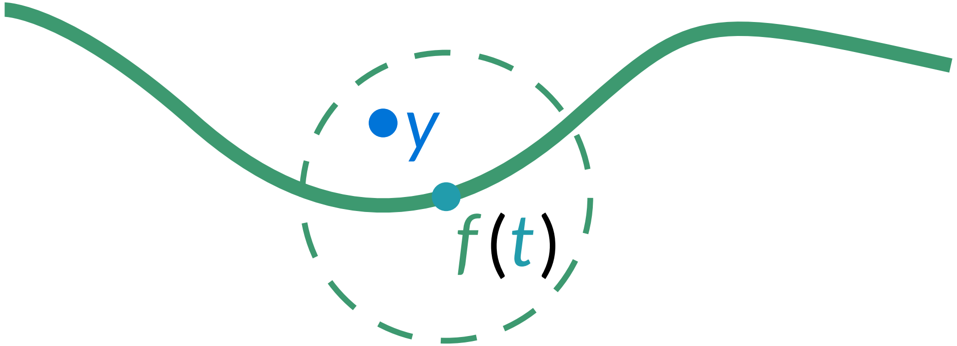
by the magic of ✨ Bayes ✨



Data generating process: Curve Mixture

Curve mixture $\mathcal{M}$ consists of curve f and observation noise σ.
Observation y is generated from $\mathcal{M}$ like this:

1. draw $t \sim \mathcal{U}[0, 1]$
2. draw $y \sim \mathcal{N}\left(f(t), \sigma^2 I\right)$



likelihood: $$p(y \mid \mathcal{M}) = \frac{1}{Z(\sigma)} \int_0^1 \exp\left(-\frac{\|y - f(t)\|^2}{2\sigma^2}\right) \mathrm{d}t$$

Smooth curves

We pefer over

… corresponding to preferring a low **bending energy**.

prior: $$p(\mathcal{M}) = \exp\left(-\tau \left(\int_0^1 \|f''(t)\|^2 \, dt\right) / \left(\int_0^1 \|f'(t)\| \, dt\right)^2\right)$$

Posterior: Likelihood and Prior combined

Bayes' theorem:

$$\underbrace{p(\mathcal{M} \mid \mathcal{Y})}_{\text{posterior}} = \prod_{y \in \mathcal{Y}} \underbrace{p(y \mid \mathcal{M})}_{\text{likelihood}} \cdot \underbrace{p(\mathcal{M})}_{\text{prior}}$$

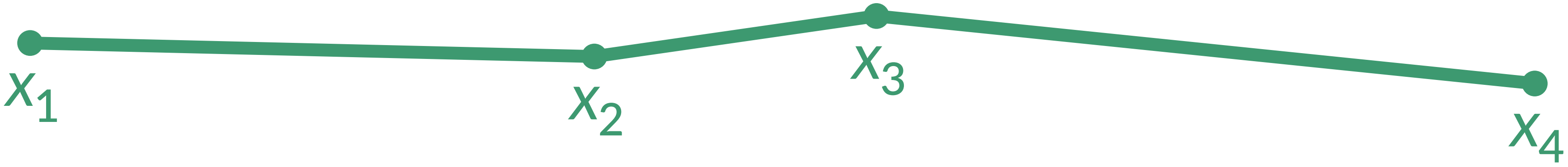
posterior: How much do we believe $\mathcal{M}$ given $\mathcal{Y}$?

likelihood: from the data generating process

prior: for the smoothness

Discrete approximation

Instead of arbitrary curves f, consider concatenations of **line segments**:

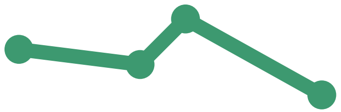
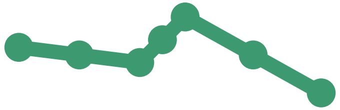


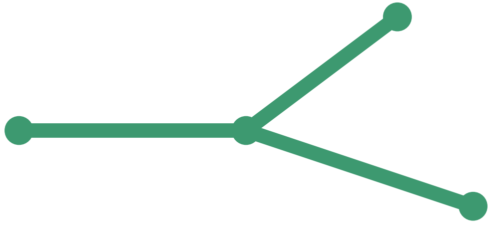
Model $\mathcal{M}$ now consists of

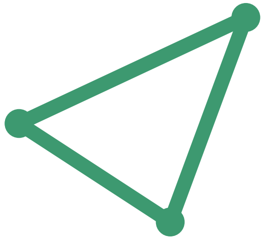
- observation noise $\sigma > 0$
- nodes $x_1, \ldots, x_n \in \mathbb{R}^d$
- connectivity information (which two nodes form a segment?)

Properties of discrete model

👍

- clear what the parameters are: $x_1, \dots, x_n \in \mathbb{R}^d$ and $\sigma \in \mathbb{R}$
- can compute integrals in likelihood and prior exactly
- $p($  $\mid \mathcal{Y}) = p($  $\mid \mathcal{Y})$
- easily generalizable to arbitrary topologies *(mixture of curve mixtures)*



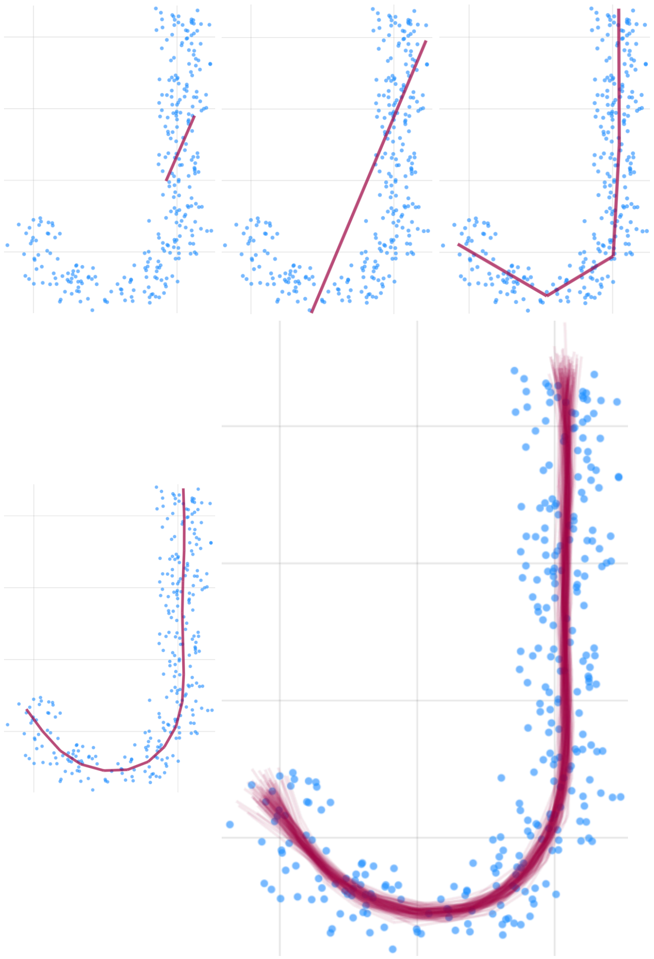


👎

- have to think about number of segments

Estimating a Curve Mixture

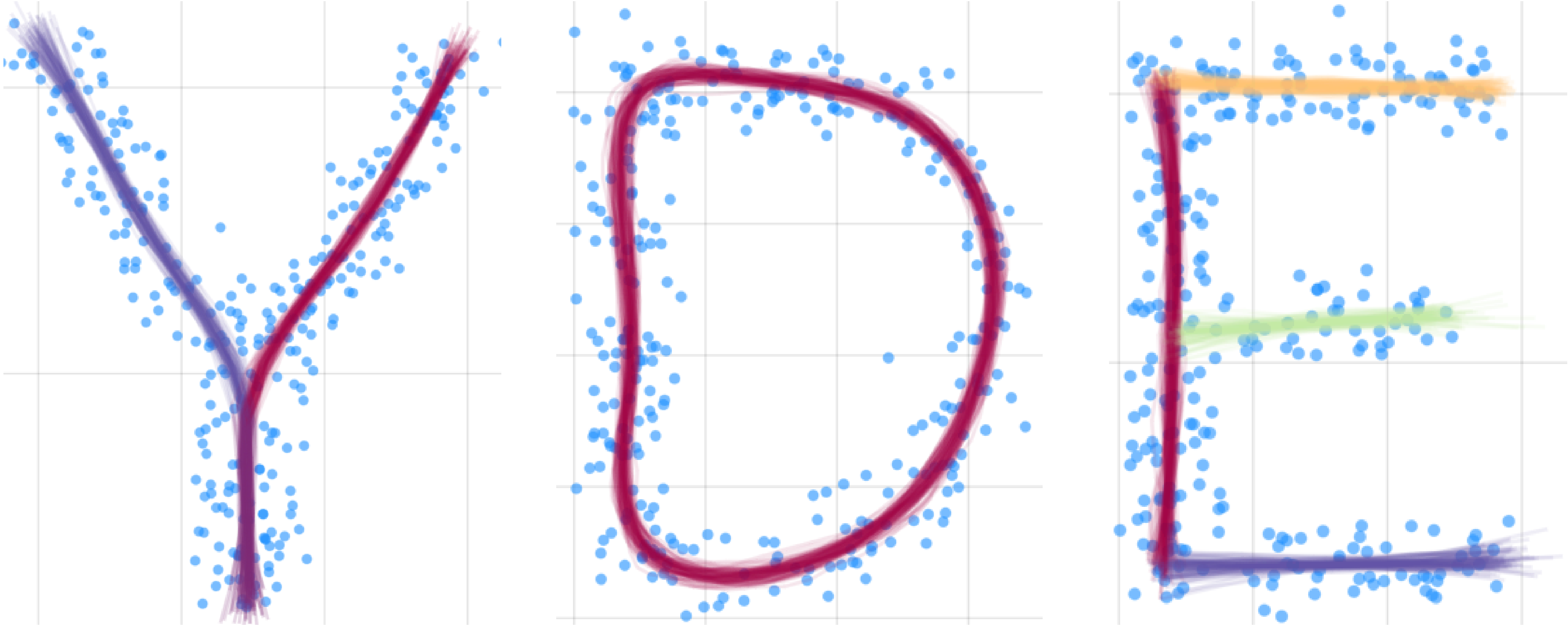
```
$\mathcal{M} \leftarrow$ single segment
**until satisfied**
    maximize posterior density (ADAM)
    finegrain $\mathcal{M}$, i.e. split every segment
sample $\mathcal{M}$ from posterior (NUTS)
```



(obtain gradient for ADAM and NUTS via automatic differentiation with `Mooncake.jl`)

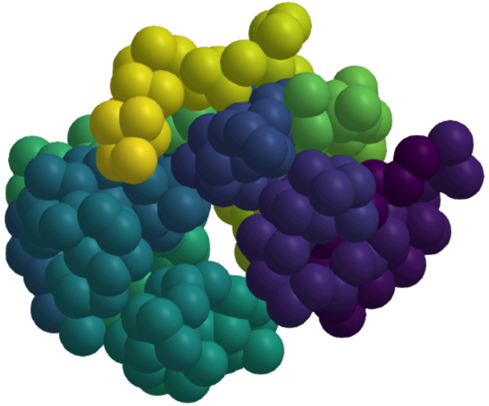
More letters!

(showcasing other topologies with a generalized model)

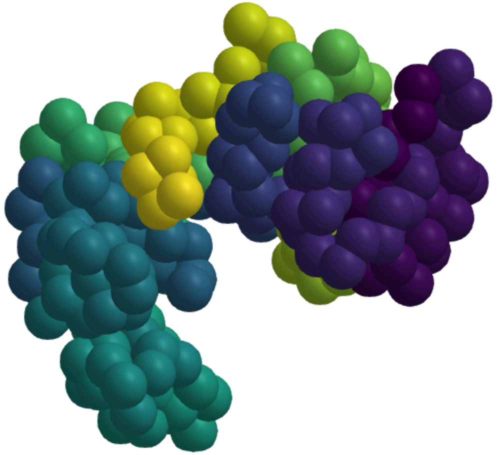


Demo: Human Serum Transferrin (1a8e)

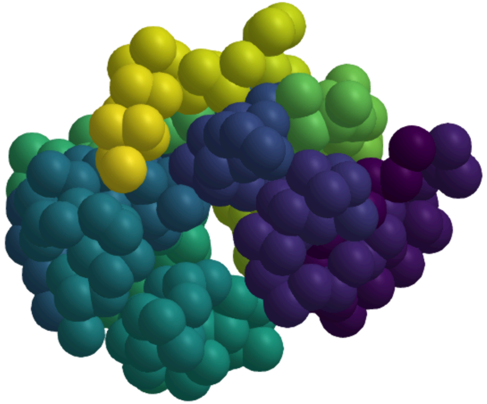
84 AA-chains from PDB with sequence similarity ≥ 90 % to 1a8e



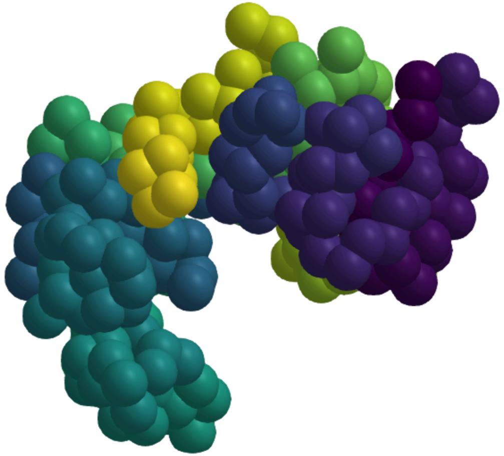
1a8e



2h52



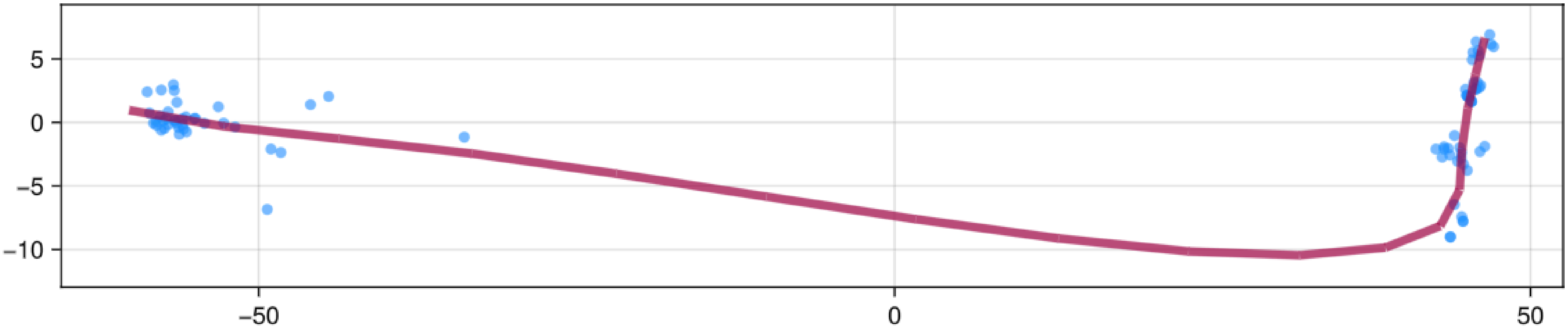
2o7u



5dhy

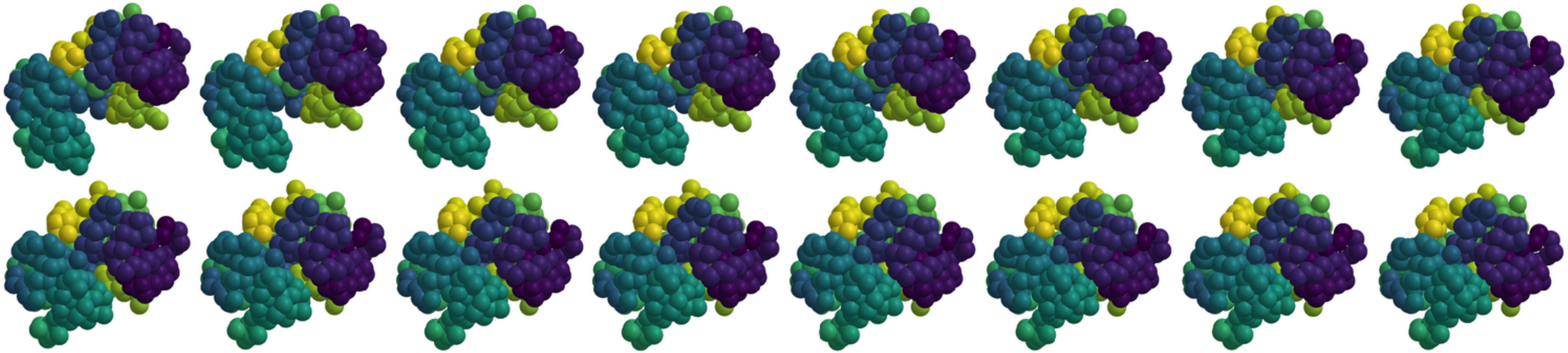
Demo: Human Serum Transferrin (1a8e)

Curve Mixture in 2D latent space (PCA)



Demo: Human Serum Transferrin (1a8e)

Traverse curve and project back into “structure space”



📺

Summary and next steps

- curve mixtures are a nice tool to find trajectories in data points
  - ⇝ predictions biologically meaningful?
  - ⇝ design user interface
- works reasonably well for PCA-embedded protein structures
  - ⇝ other latent embeddings?
  - ⇝ apply to RNA-seq data
- implementation in Julia let me focus on domain problem and performance (gradient, sampling, and optimization “for free”)

Let’s discuss!

andreas.kroepelin@uni-jena.de

a.ndreas.dev

@andreask@bayes.club

⇧ slides and code on GitHub