Interpretable embeddings from molecular simulations using Gaussian mixture variational autoencoders

Applied Machine Learning Days – EPFL 2022

Yasemin Bozkurt Varolgüneş

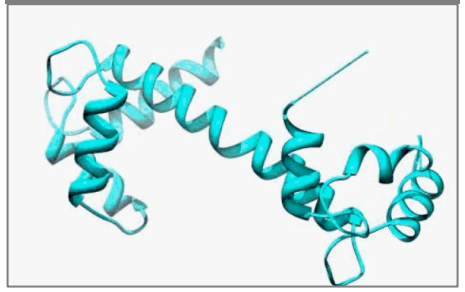
30-03-2022

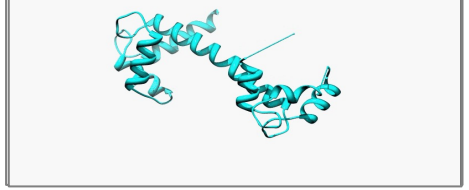
MAX PLANCK INSTITUTE
FOR POLYMER RESEARCH

In collaboration with
Dr. Tristan Bereau
Dr. Joseph F. Rudzinski

Structure + Dynamics = Function

Structure → Function ← Dynamics





Experiments

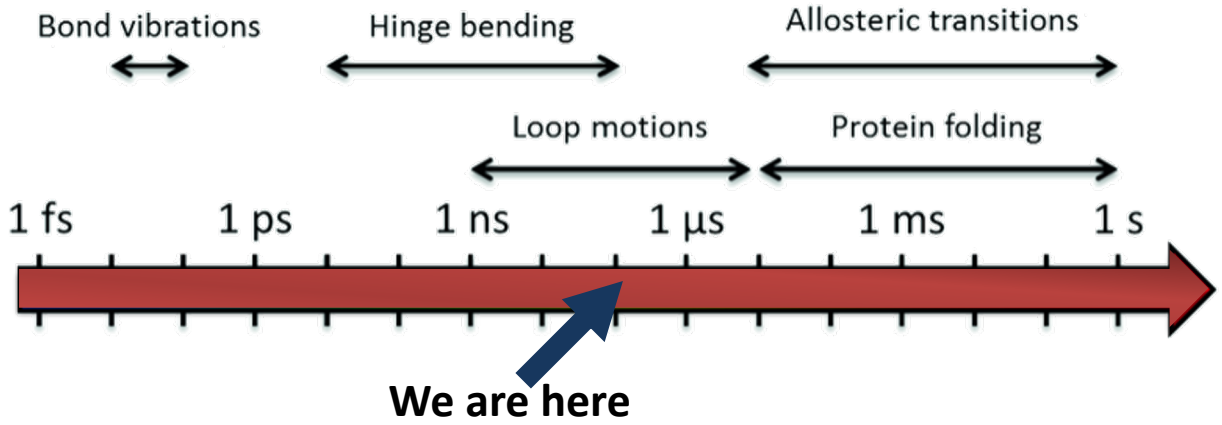
High spatial **or** temporal resolution

- High spatial but low temporal resolution
  Cryo-electron microscopy, X-ray diffraction
- High temporal but low spatial resolution
  Single molecule fluorescence resonance energy transfer

Molecular Dynamics (MD) Simulations

High spatial **and** temporal resolution

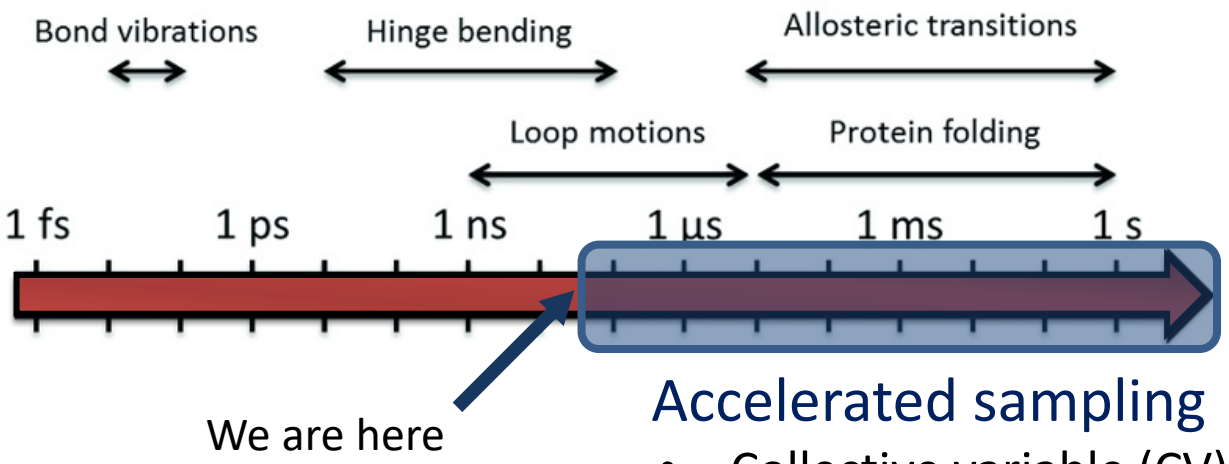
- By iteratively solving **equations of motion**
  Computationally expensive
  Timescale gap problem



Movie credit *Kozlowski et al. "A meta-server fpr prediction of IDPs". BMC Bioinformatics (2012)*
Nobel prize in Chemistry 2013 awarded to Karplus, Levitt and Warshel for development of MD

Figure credit *Adrien et al. "Raman and Infrared Spectra of Acoustical, Functional Modes of Proteins from All-Atom and Coarse-Grained Normal Mode Analysis". Springer (2019)*

Unsupervised ML in molecular simulations



Accelerated sampling

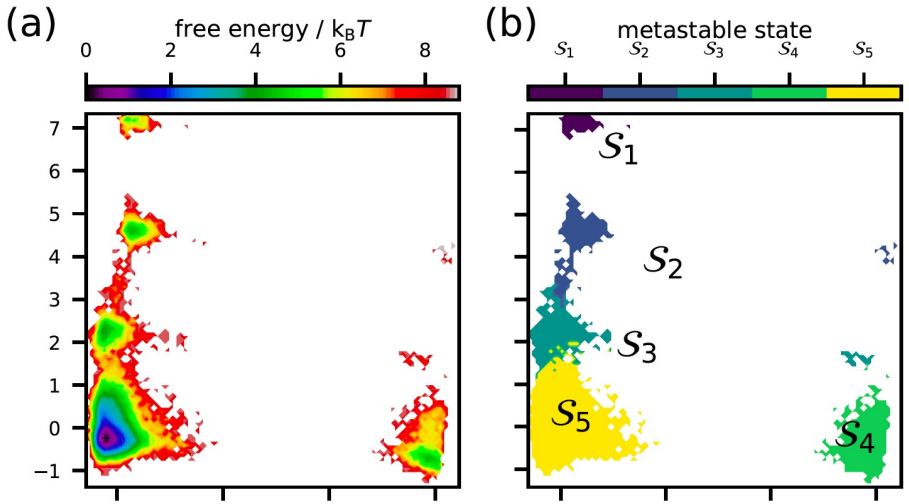
- Collective variable (CV) biasing

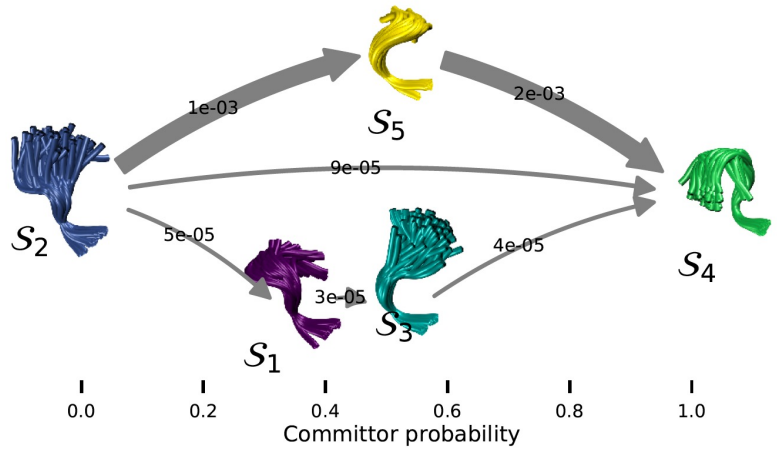
A good CV

- separates metastable states
- characterizes slow motions
- helps build kinetic models

CV examples

internal angles, pairwise distances, coordination numbers etc.





$$P(E) = \frac{1}{Z} \exp\left(-\frac{E}{k_B T}\right)$$

Data-driven route to discover good low-D representations <<<

Andrew Ferguson, "Machine learning-enabled enhanced sampling in biomolecular simulation and data-driven design", IPAM talk (2019)

Wehmeyer et al. "Introduction to Markov state modeling with the PyEMMA software", (2018)

(Vanilla) Autoencoder

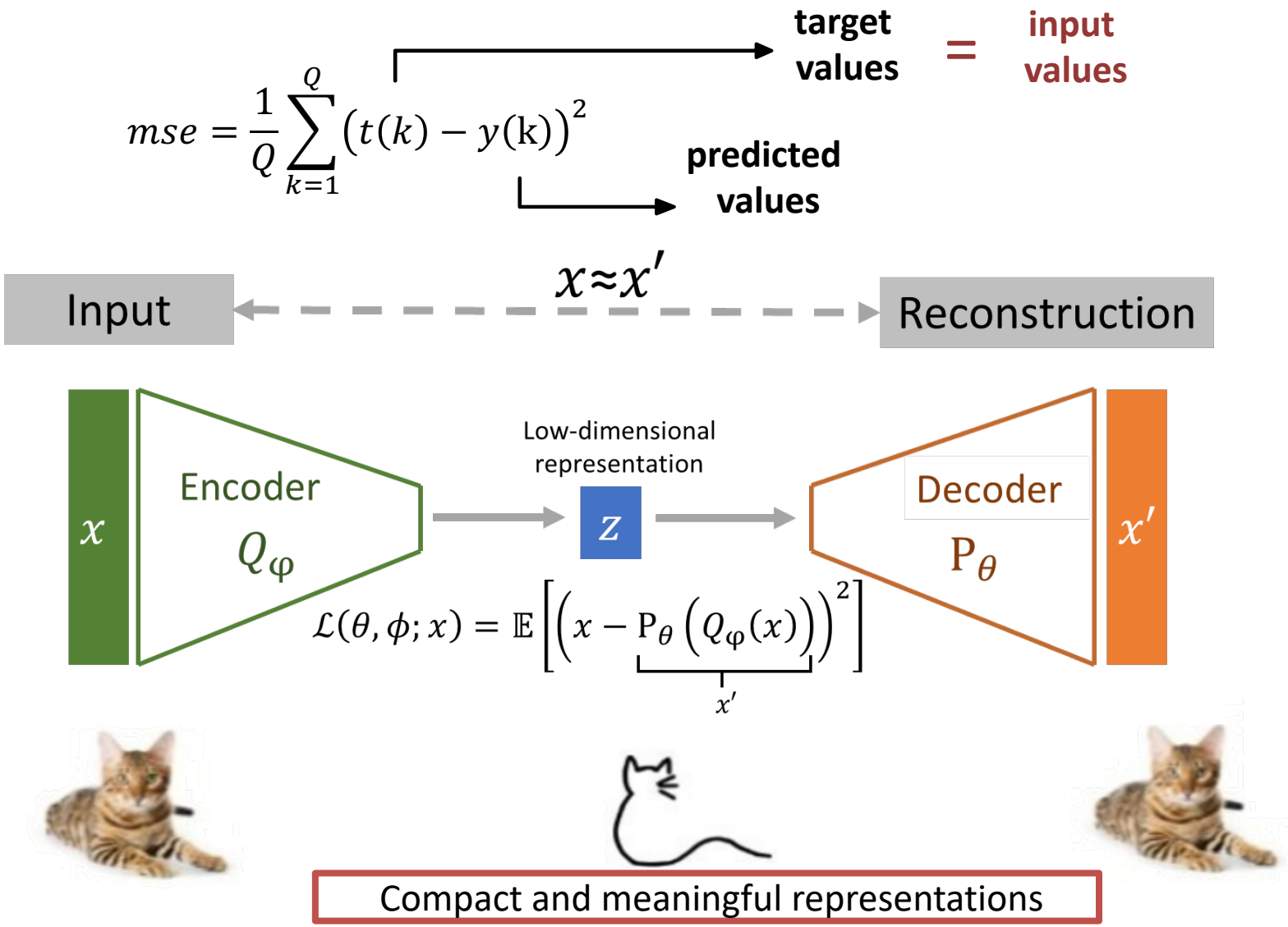
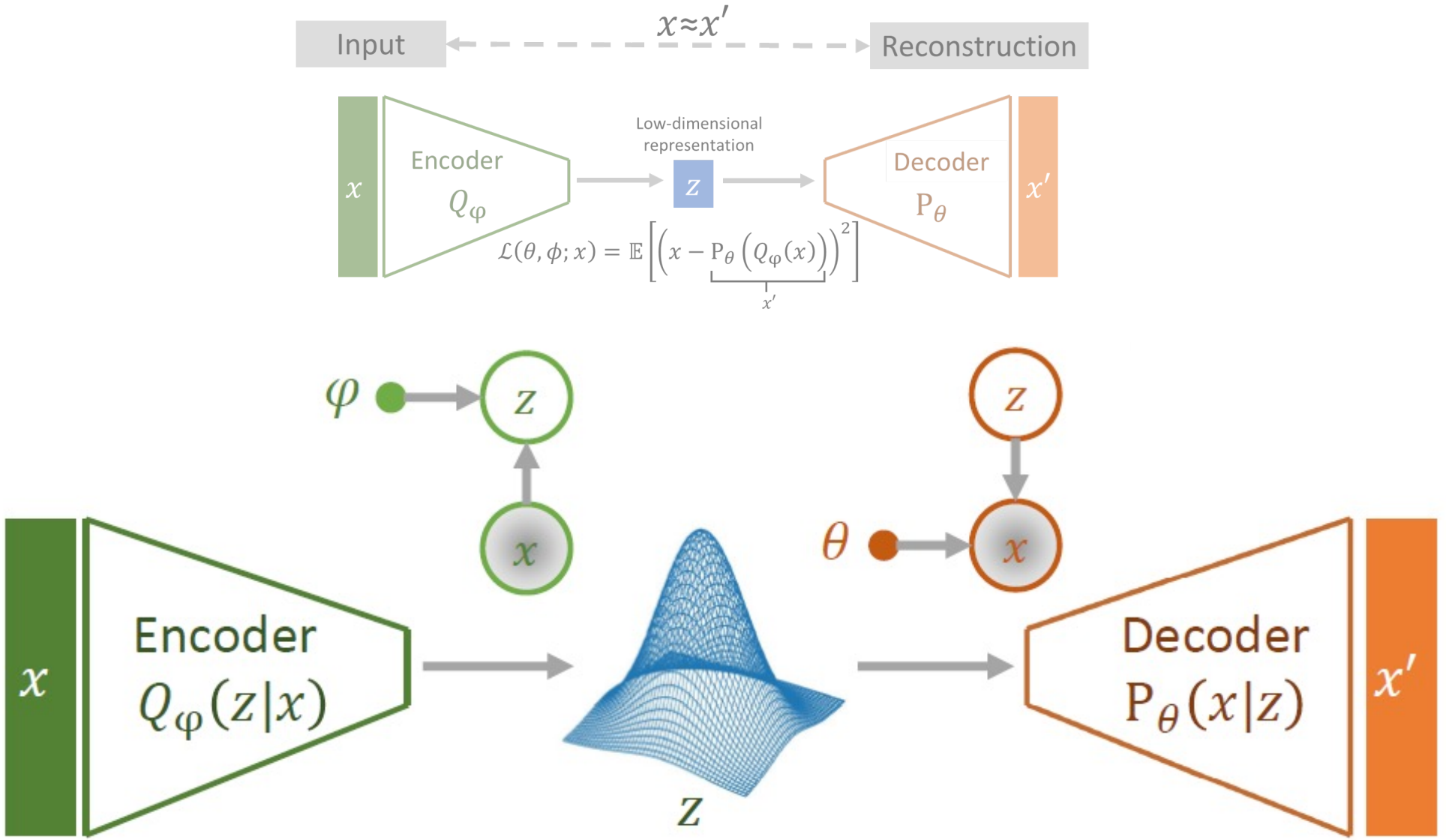


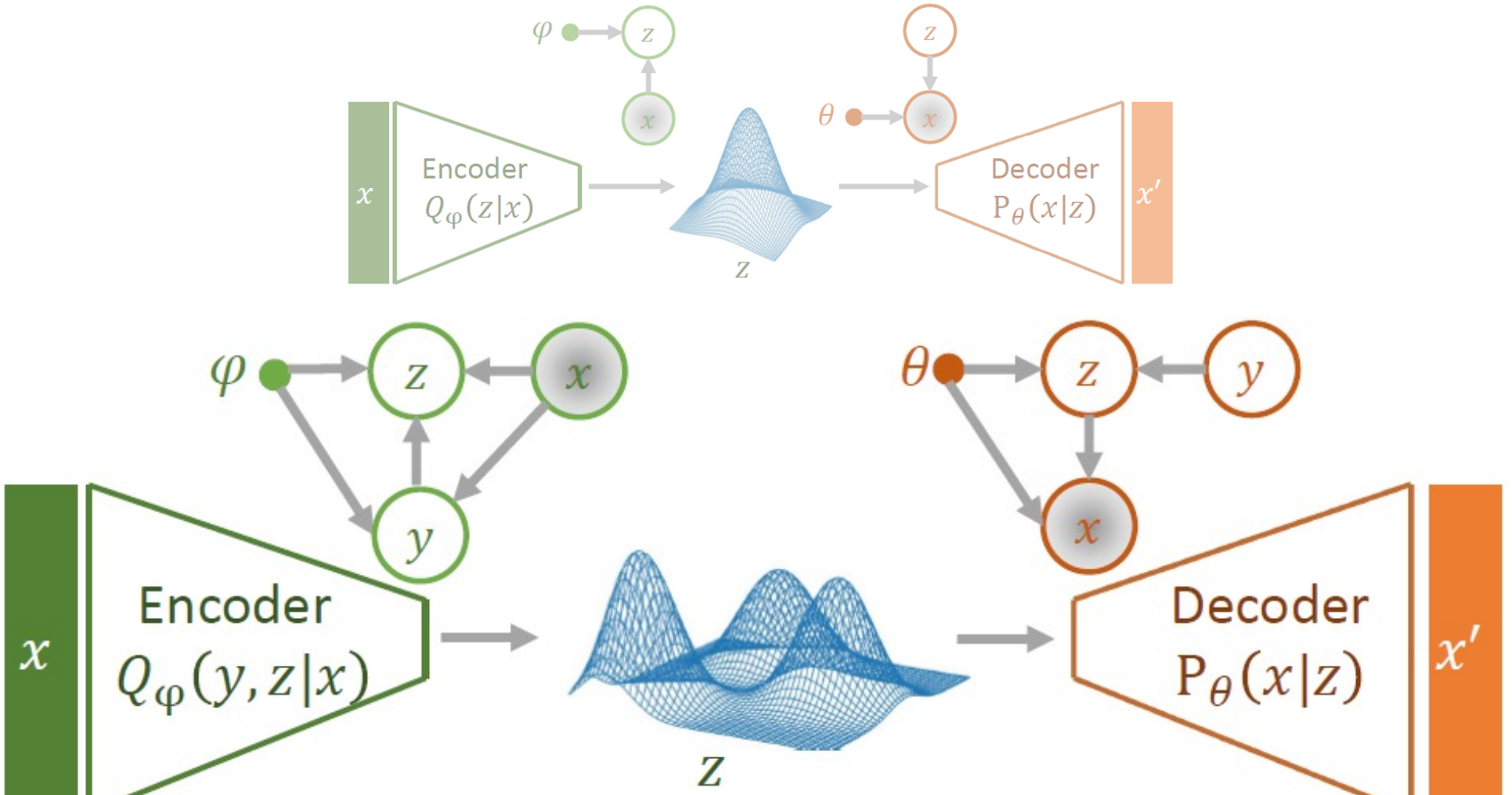
Figure credit *https://sefiks.com/2018/03/21/autoencoder-neural-networks-for-unsupervised-learning*

Variational autoencoder (VAE)



Kingma et al. "Auto-encoding Variational Bayes" arXiv, (2013)

Gaussian mixture variational autoencoder (GMVAE)



$$\mathcal{L}(\theta,\phi;x) = \mathbb{E}_{Q_\phi(y,z|x)}\left[\log P_\theta(x,y,z) - \log Q_\phi(y,z|x)\right]$$

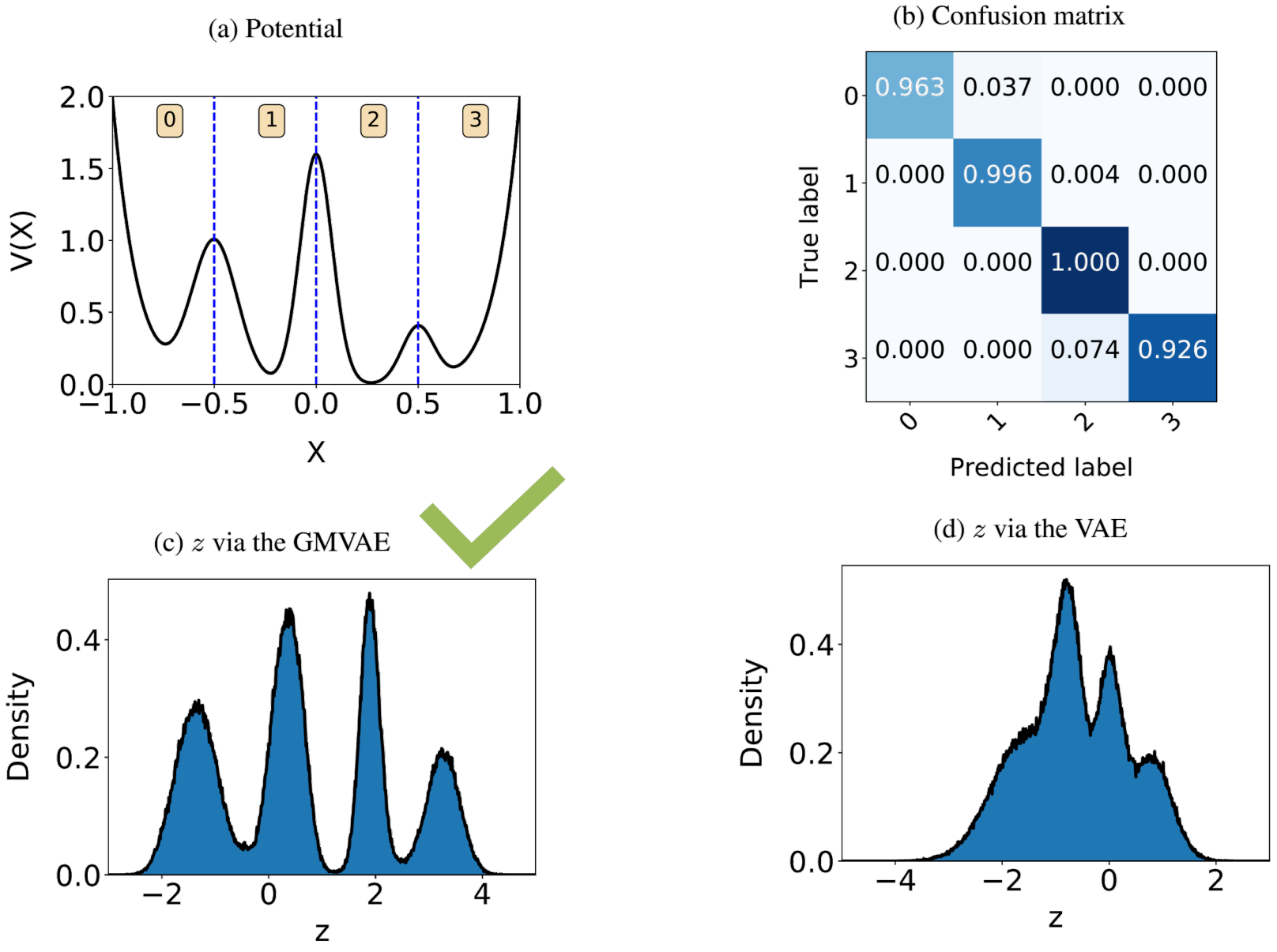
$$\mathcal{L}(\theta,\phi;x) = \mathbb{E}_{Q_\phi(y,z|x)}\left[\log\frac{P(y)}{Q_\phi(y|x)} + \log\frac{P_\theta(z|y)}{Q_\phi(z|x,y)} + \log P_\theta(x|z)\right]$$

Bozkurt et al. "Interpretable embeddings from molecular simulations using Gaussian mixture variational autoencoders." Machine Learning: Science and Technology (2020)
http://ruishu.io/2016/12/25/gmvae/

The GMVAE accurately identifies the clusters.

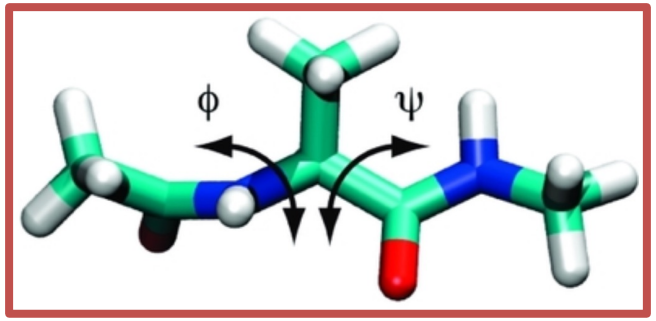
1D 4 well potential

1D ⟶ 1D



Test for conformational dynamics

Alanine dipeptide



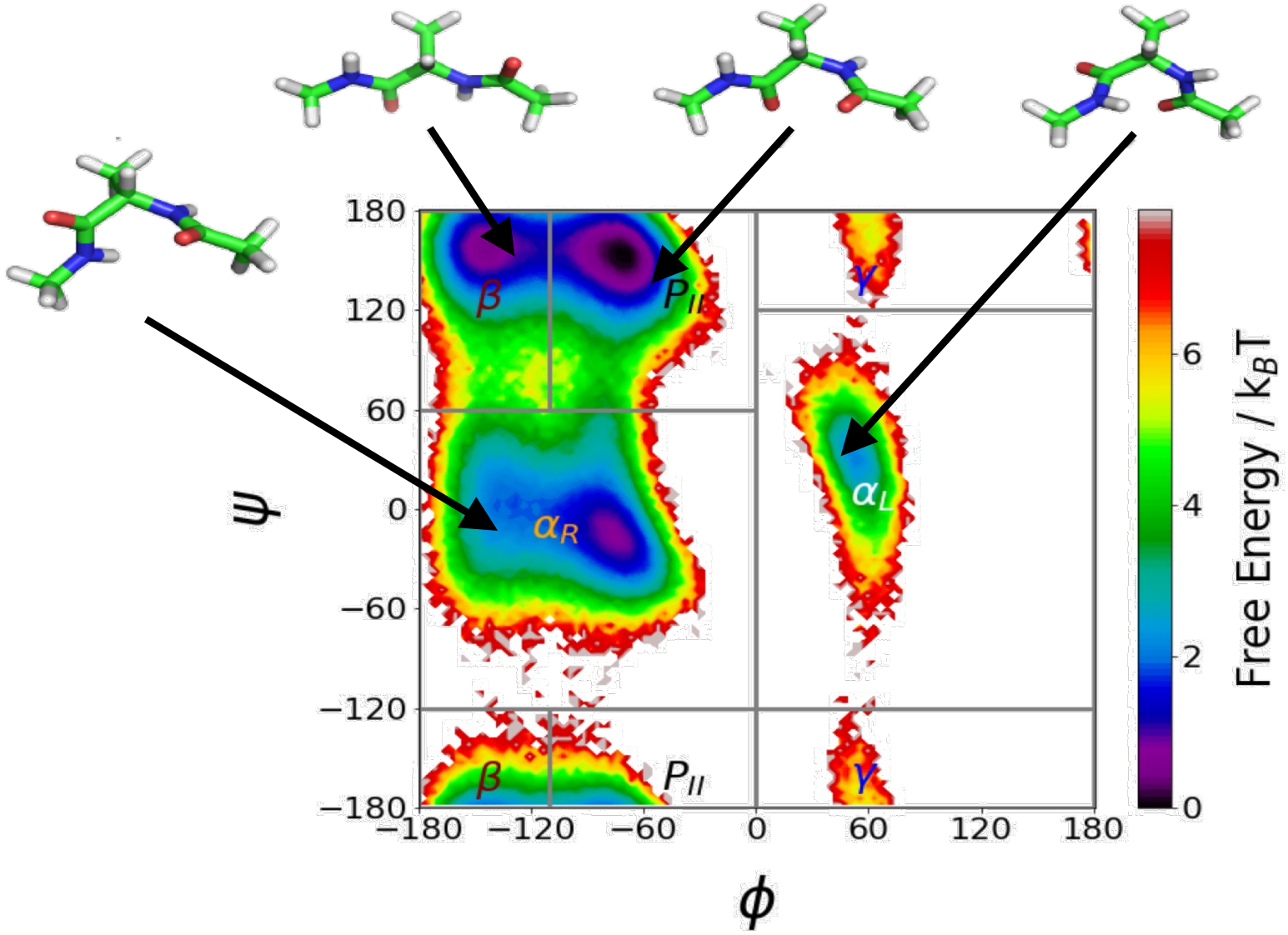
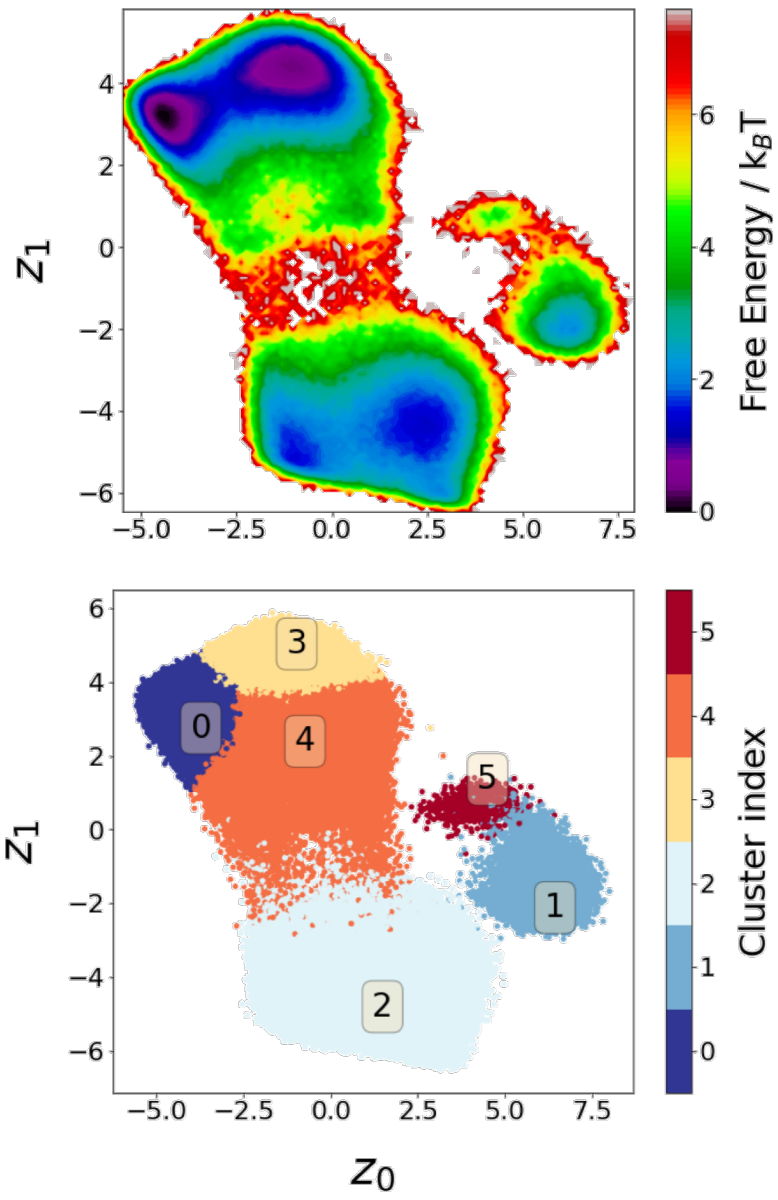


Figure credit *Devaurs et al. " A multi-tree approach to compute transition paths on energy landscapes." Workshops at the 27. AAAI Conference on Artificial Intelligence. 2013 (2017)*

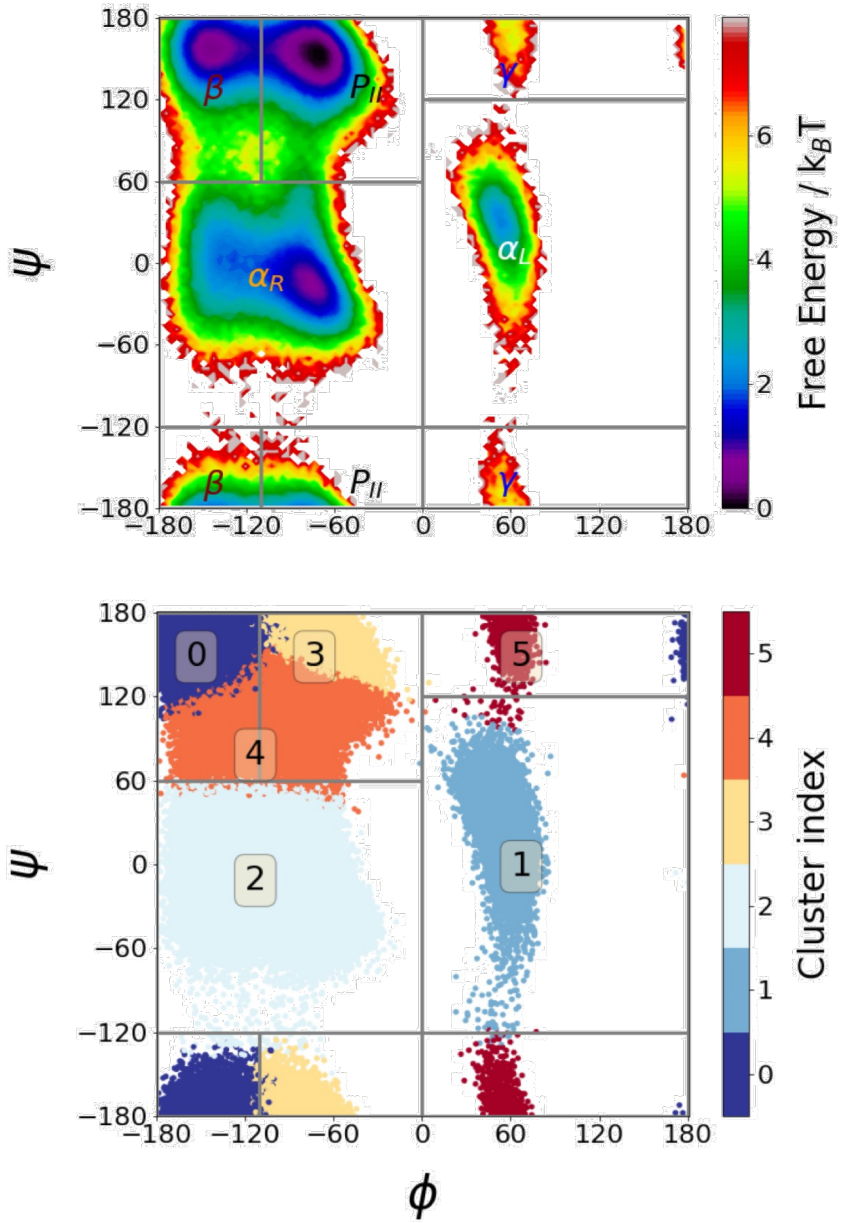
Free energy landscape of alanine dipeptide

25D → 2D

GMVAE landscape

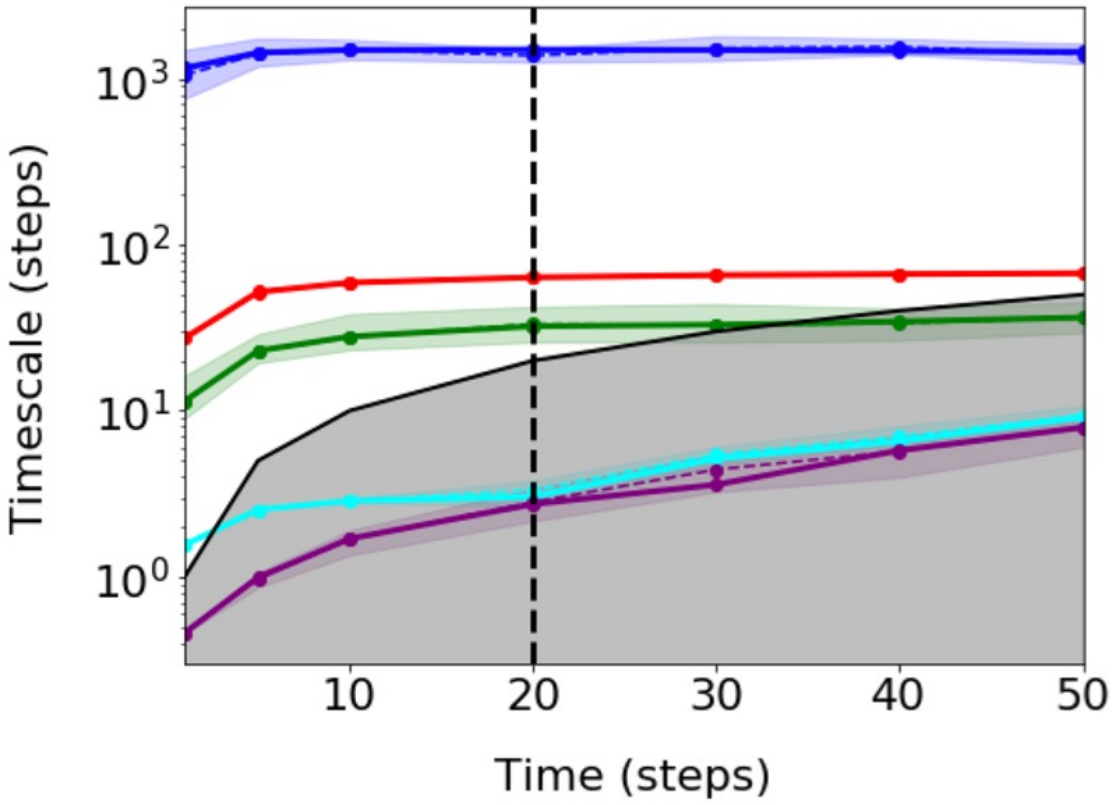


Ramachandran plot

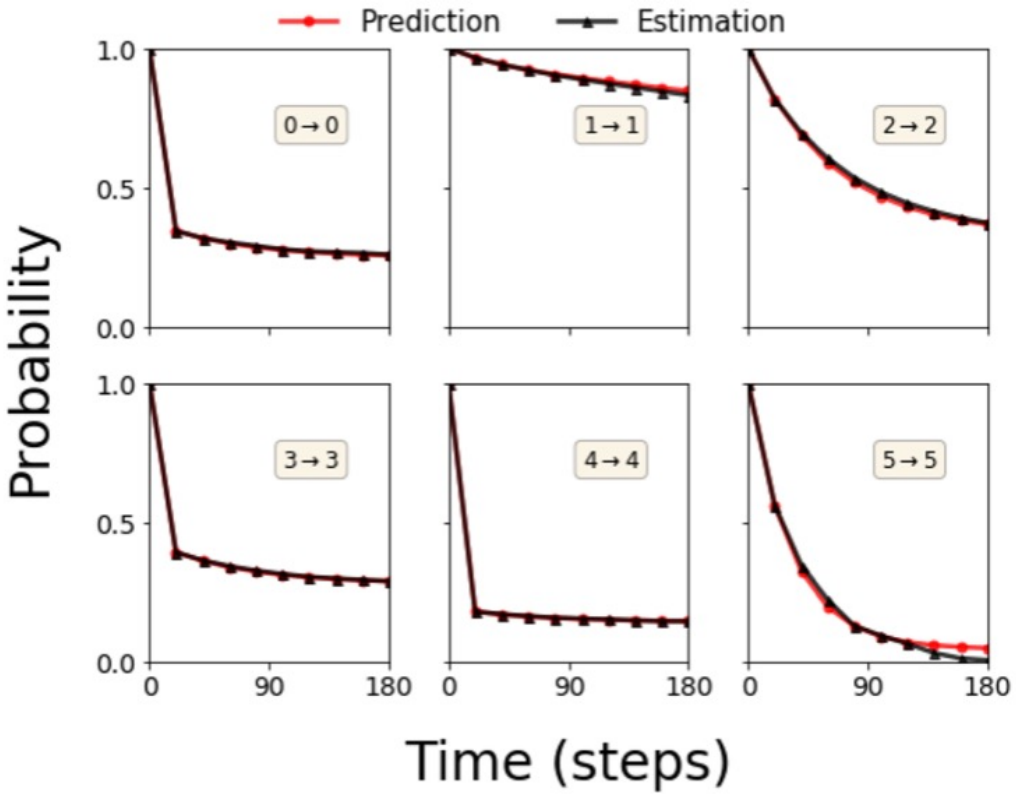


Markovian model

(a) Implied timescales



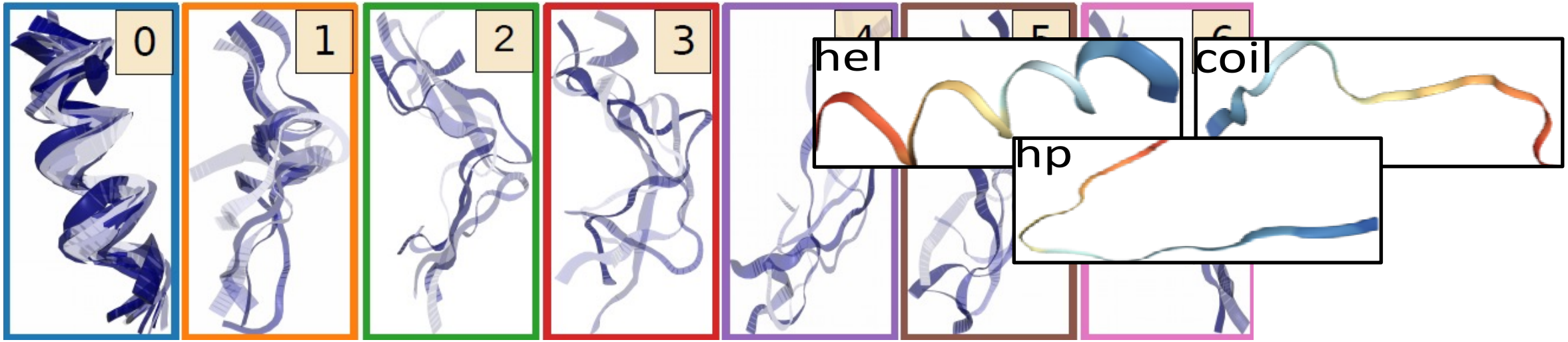
(b) Chapman-Kolmogorov test



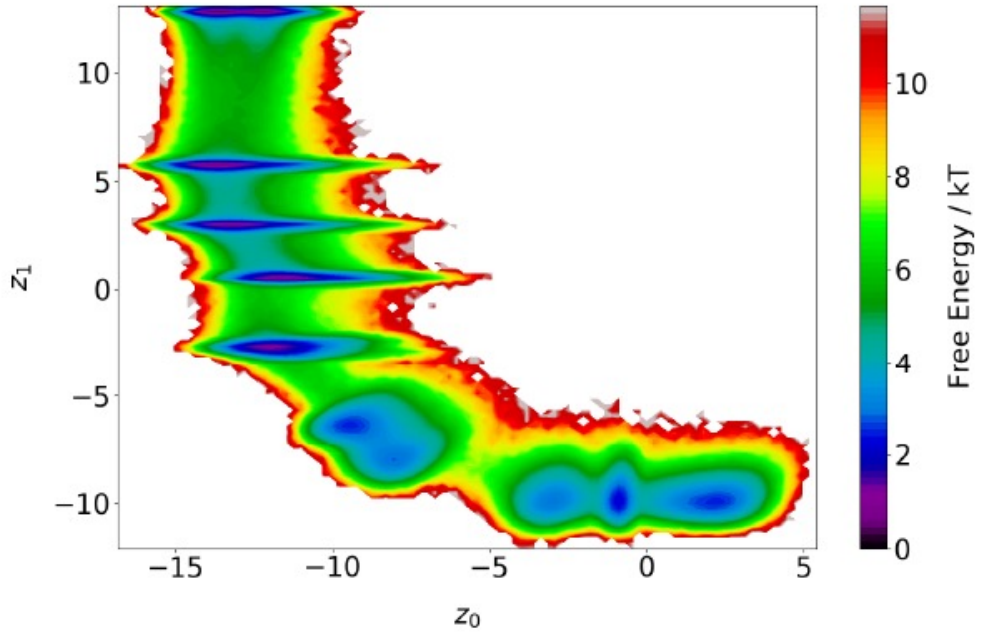
$$t_i(\tau) = -\frac{\tau}{\ln|\lambda_i(\tau)|}$$

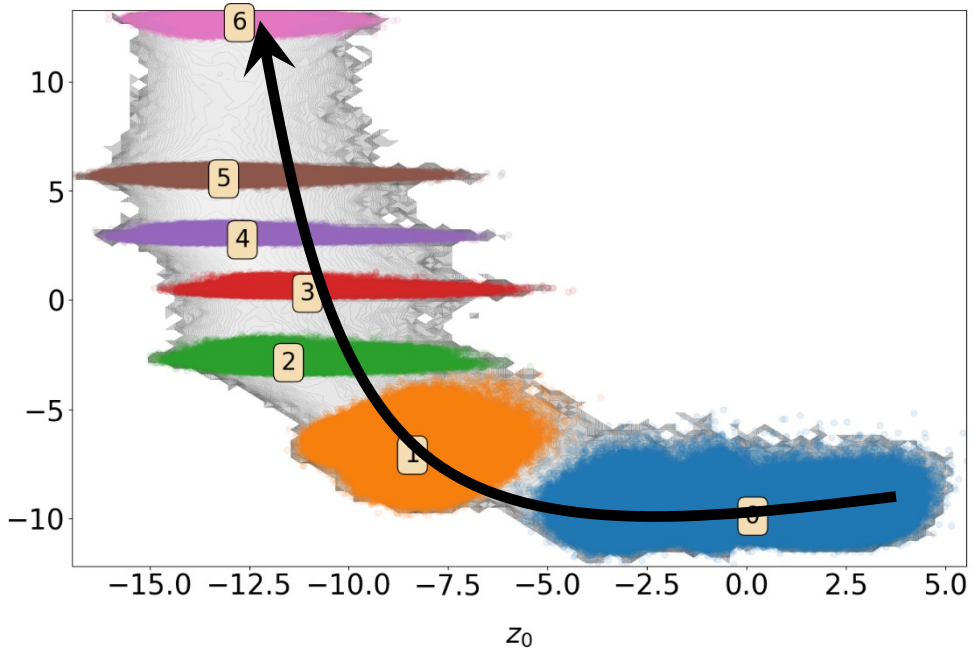
$$p_{ij}(m\tau) = p_{ij}^m(\tau)$$

$AAQAA_3$ peptide

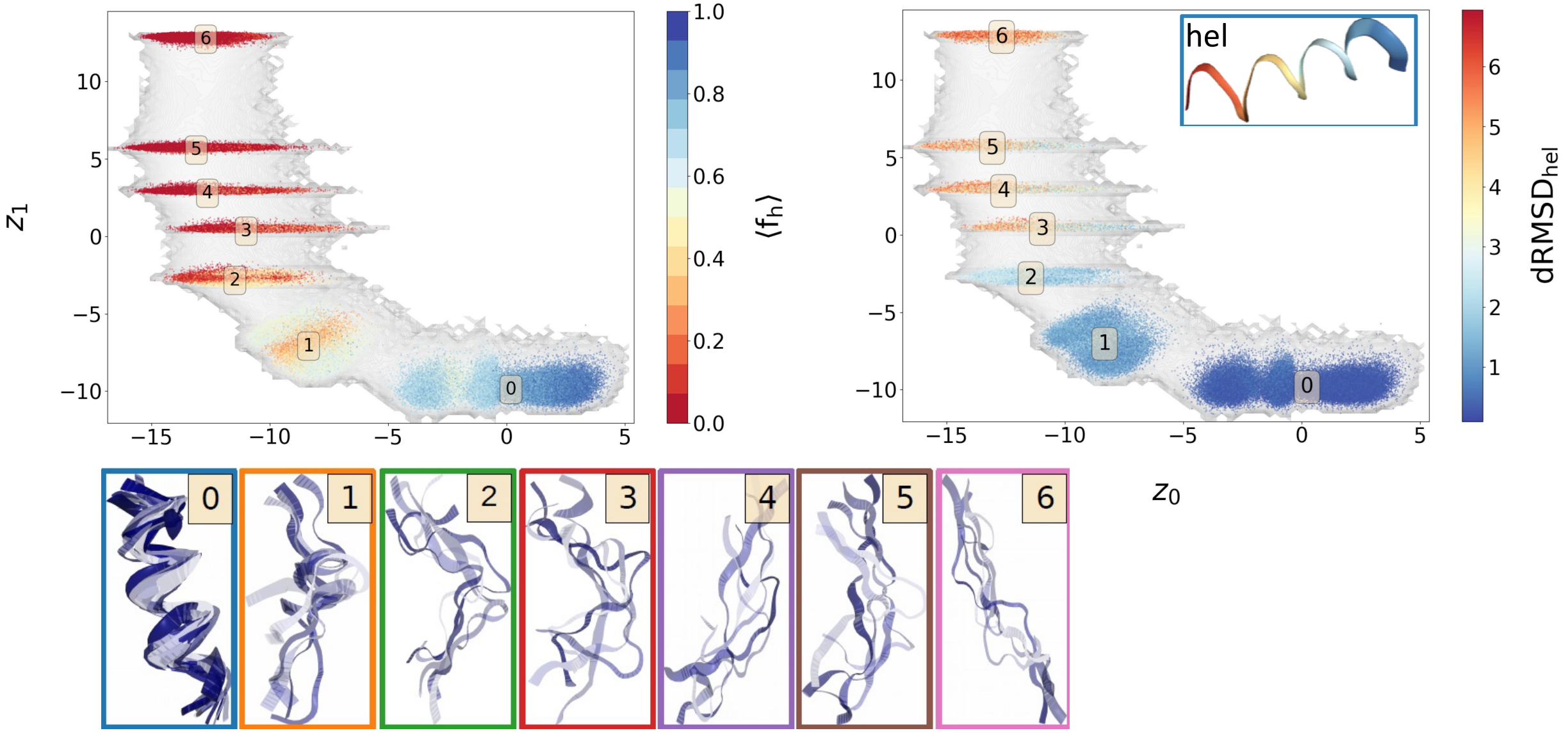


60D ⟶ 2D





Checking the projections in more detail

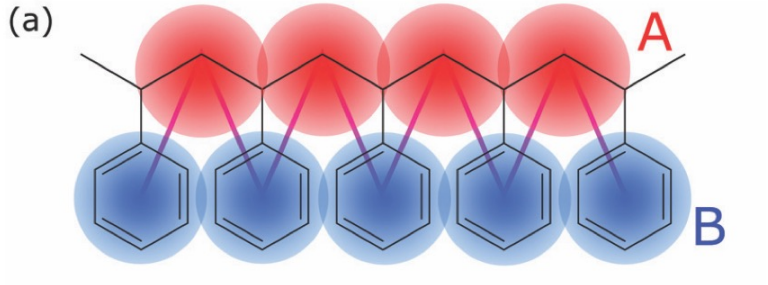


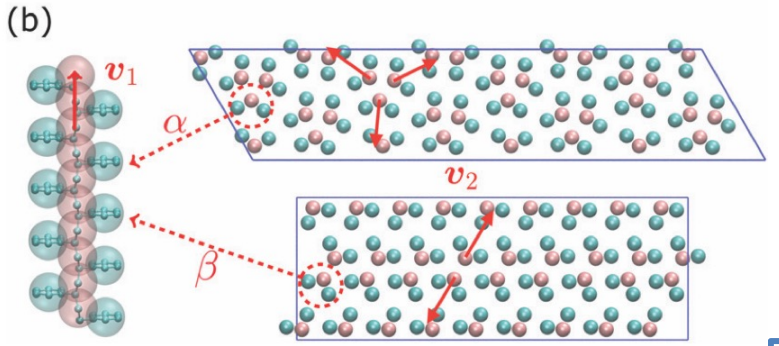
Material science application: polystyrene

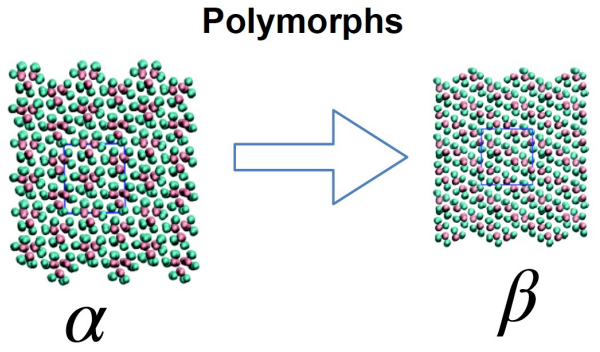
Polymorphs are different crystalline forms of the same substance in which molecules have **different arrangements** and/or different molecular conformations → Display **different properties**

Figure credit: https://letstalkscience.ca/educational-resources/stem-in-context/polystyrene-pros-cons-chemistry

Material science application: Polymorphism in polystyrene

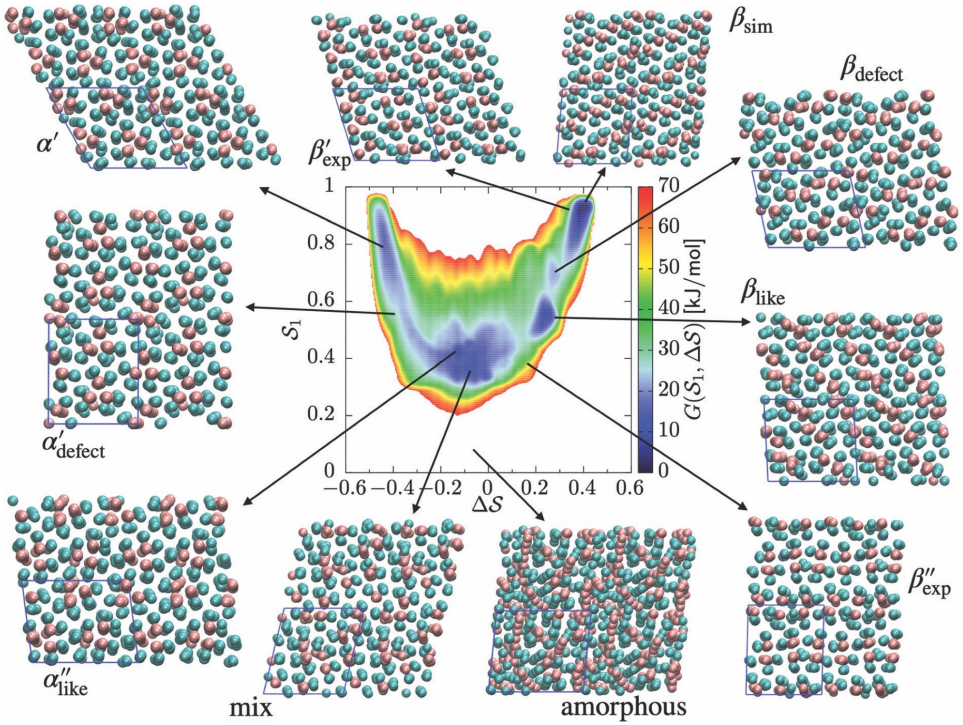






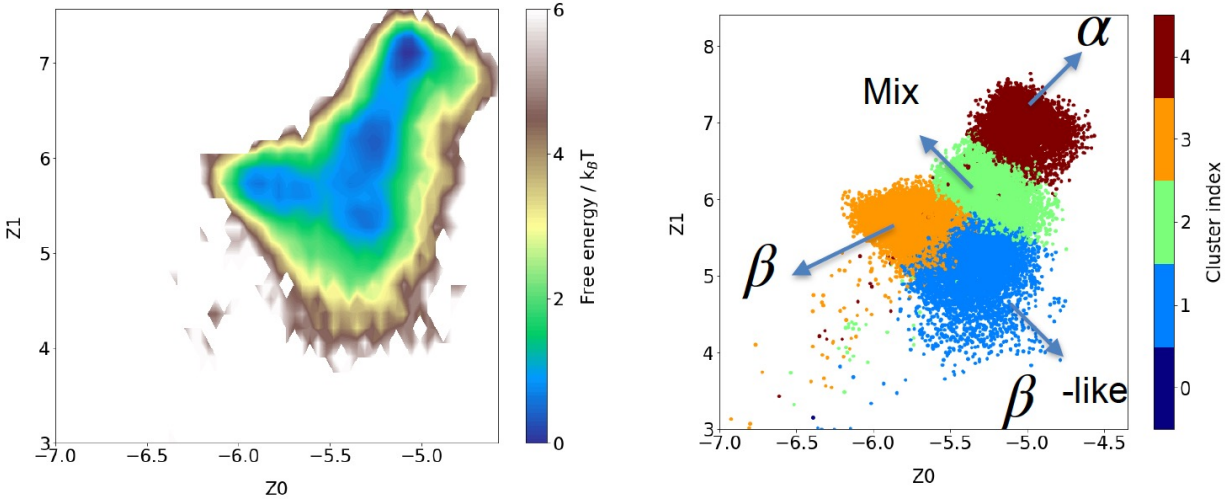
Rugged free energy landscape leads to long timescales of α to β transition.

Conformational landscape of polystyrene



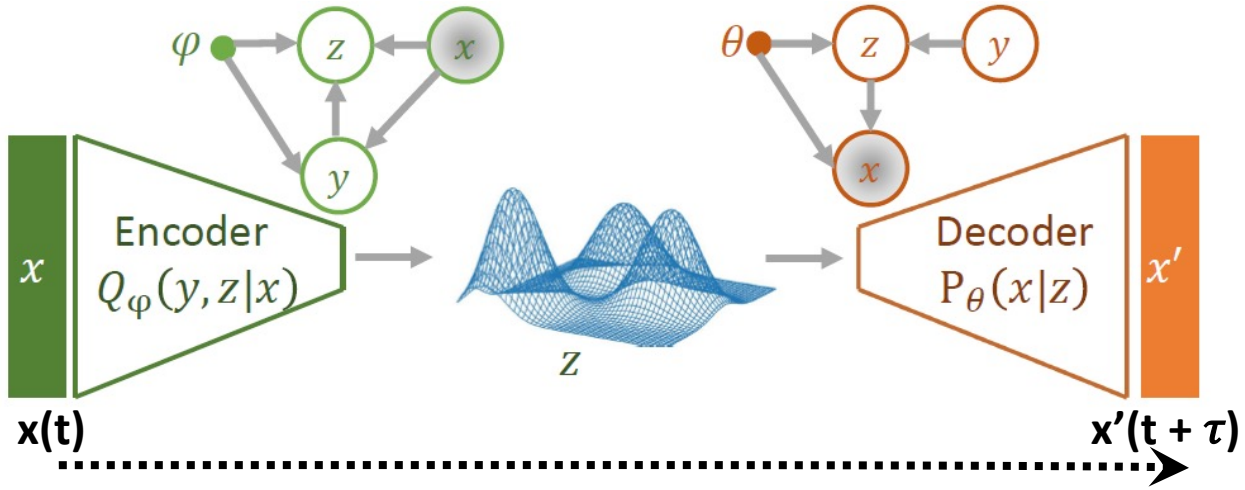
Identifying good CVs that describe the full conformational landscape is challenging.

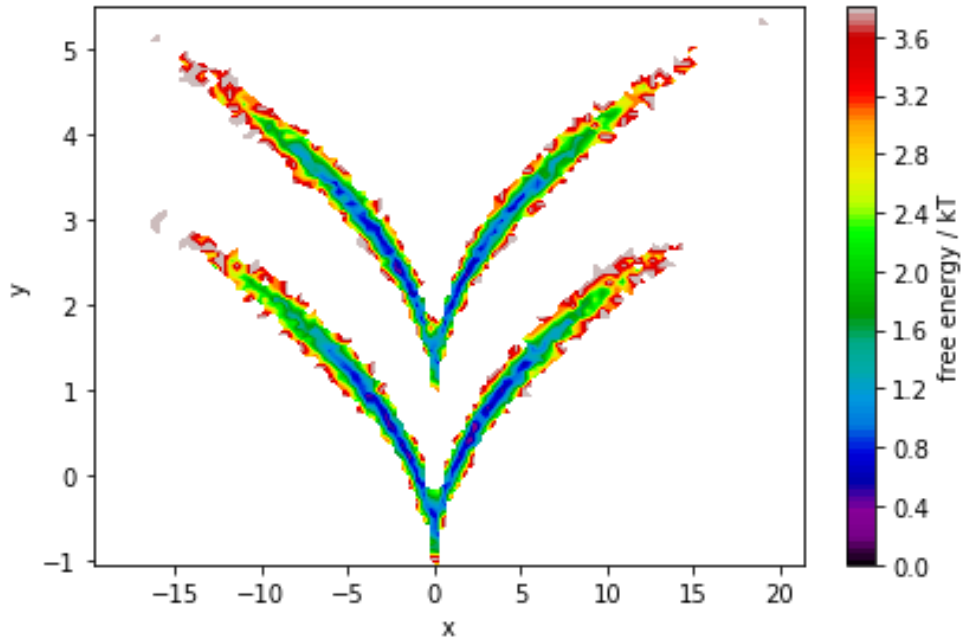
GMVAE results (preliminary)

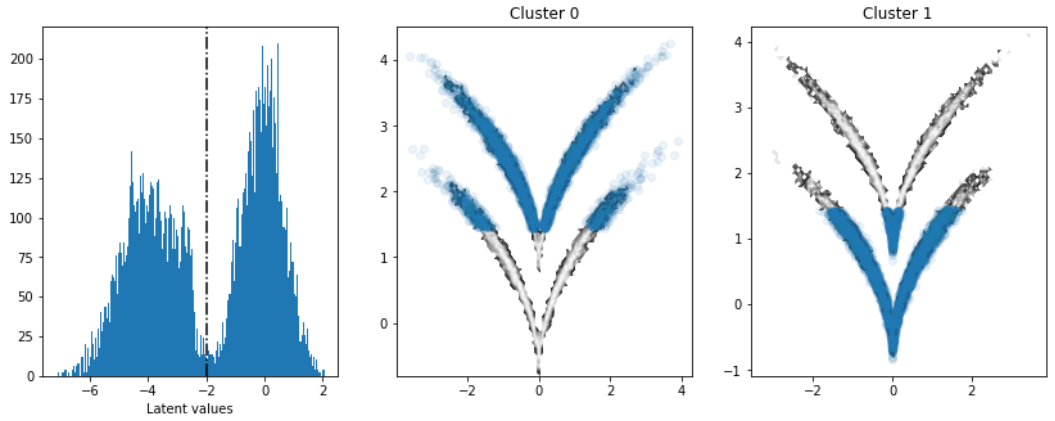


Liu et al. Soft Matter (2020)

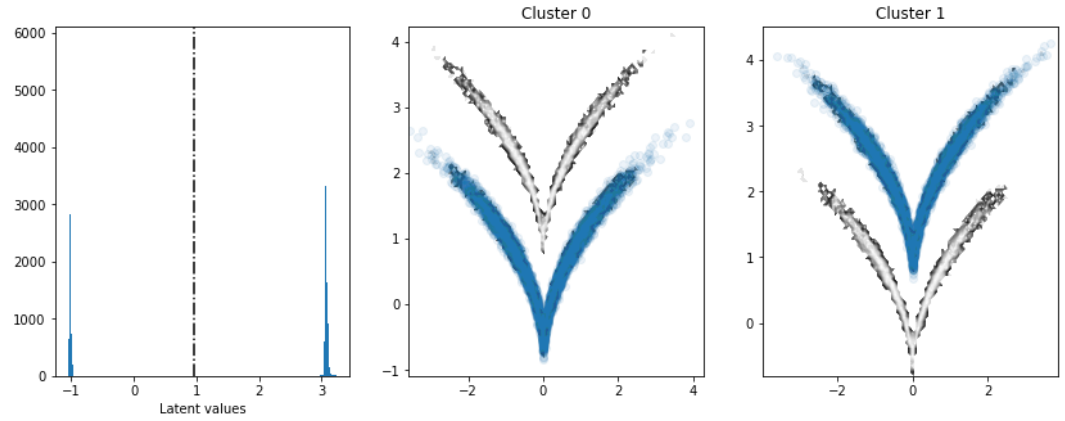
Incorporation of time-information into the GMVAE







no lag



with time-lag

Conclusion

- The GMVAE
  - restricts the latent space representation with a Gaussian mixture model,
    - ✓promoting the separation of metastable states
    - ✓allowing simultaneous dimensionality reduction and clustering.

Bozkurt et al. "Interpretable embeddings from molecular simulations using Gaussian mixture variational autoencoders." Machine Learning: Science and Technology (2020)

github.com/yabozkurt/gmvae

Acknowledgments

Assist. Prof. Tristan Bereau
University of Amsterdam, MPIP

Dr. Joseph F. Rudzinski
MPIP

Dr. Atreyee Banerjee
MPIP

MAX PLANCK INSTITUTE FOR POLYMER RESEARCH

TÜBİTAK

Thank you for your attention!

Connect with me!

Yasemin Bozkurt Varolgüneş

Max Planck Institute for Polymer Research