

# Practical Course on Molecular Dynamics and Trajectory Analysis

## Episode 6: Markov models with PyEMMA

Jordi Villà i Freixa

Universitat de Vic - Universitat Central de Catalunya  
Facultat de Ciències, Tecnologia i Enginyeries (FCTE)

*jordi.villa@uvic.cat*

MD Course and Trajectory Analysis  
Concepcion, January 2026

# Contents

- 1 Episode 6: Markov models with PyEMMA
  - Motivation and theoretical background
  - Trajectory preparation
  - Feature representation
  - Time-lagged Independent Component Analysis
  - Discretization
  - MSM estimation
  - Spectral analysis and validation
  - Transition Path Theory

# Why Markov State Models?

- Molecular dynamics (MD) simulations generate high-dimensional trajectories  $\{\mathbf{X}_t\}_{t=0}^T$ .
- Relevant molecular processes occur on timescales much longer than MD timesteps.
- MSMs provide a statistical coarse-graining into discrete states with Markovian dynamics.
- Enable computation of long-timescale kinetics, populations, and pathways.

# From continuous dynamics to a Markov chain

- Consider a stochastic process  $\mathbf{X}_t$  in phase space  $\Omega$ .
- Partition  $\Omega$  into disjoint sets  $\{S_1, \dots, S_N\}$ .
- Define a discrete process  $X_t \in \{1, \dots, N\}$ :

$$X_t = i \quad \text{if } \mathbf{X}_t \in S_i.$$

- Markov assumption at lag time  $\tau$ :

$$P(X_{t+\tau} = j \mid X_t = i, \dots) \approx P(X_{t+\tau} = j \mid X_t = i).$$

# Preparing molecular trajectories

- Input trajectories from MD engines (OpenMM, Gromacs, AMBER, ...).
- Preprocessing:
  - Remove periodic boundary artifacts.
  - Align structures to a reference.
  - Remove solvent if not used as features.
  - Subsample to a uniform timestep  $\Delta t$ .
- Validate trajectories: energy stability, RMSD convergence.

# Feature extraction

- Each frame is mapped to a feature vector  $\mathbf{x}_t \in \mathbb{R}^d$ .
- Typical features:
  - Interatomic distances or contacts.
  - Dihedral angles.
  - Ligand–protein distances.
- Features should resolve slow collective motions.

## Covariance structure

$$\bar{\mathbf{x}} = \langle \mathbf{x}_t \rangle,$$
$$C_0 = \langle (\mathbf{x}_t - \bar{\mathbf{x}})(\mathbf{x}_t - \bar{\mathbf{x}})^T \rangle.$$

- Averages over all frames and trajectories.
- $C_0$  captures instantaneous correlations.

# Time-lagged covariance

$$C_\tau = \langle (\mathbf{x}_t - \bar{\mathbf{x}})(\mathbf{x}_{t+\tau} - \bar{\mathbf{x}})^T \rangle.$$

- Measures correlations persisting over lag time  $\tau$ .
- Slow processes correspond to large time-lagged correlations.
- This procedure is known as time-lagged independent component analysis (TICA): the eigenvectors of  $C_\tau$  define the slow collective coordinates (tICs).

# Generalized eigenvalue problem

$$C_\tau \mathbf{w}_i = \lambda_i C_0 \mathbf{w}_i.$$

- Eigenvectors  $\mathbf{w}_i$  define tICs.
- Projection:

$$y_{i,t} = \mathbf{w}_i^T \mathbf{x}_t.$$

# Implied timescales

$$t_i = -\frac{\tau}{\ln \lambda_i}.$$

- Estimates relaxation timescales of slow modes.
- Plateaus vs.  $\tau$  indicate robust dynamics.

# Clustering into microstates

- Project data onto first  $m$  tICs.
- Cluster in reduced space (e.g.  $k$ -means).
- Each frame assigned to a discrete state  $i$ .

## Discrete trajectories

- Continuous trajectories become symbol sequences:

$$X^{(n)} = (x_0^{(n)}, \dots, x_{T_n}^{(n)}).$$

- These sequences define the MSM input.

## Transition counts

$$C_{ij}(\tau) = \sum_t \mathbb{I}(X_t = i, X_{t+\tau} = j).$$

- Counts transitions from  $i$  to  $j$  at lag time  $\tau$ .

# Transition matrix

$$T_{ij}(\tau) = \frac{C_{ij}(\tau)}{\sum_k C_{ik}(\tau)}.$$

- Row-stochastic matrix.
- Interpreted as conditional probabilities.

## Detailed balance

$$\pi_i T_{ij} = \pi_j T_{ji}.$$

- Expected for equilibrium simulations.
- Enforcing reversibility reduces statistical noise.

# Stationary distribution

$$\pi^T T = \pi^T.$$

- $\pi_i$  gives equilibrium populations.
- Free energies:

$$F_i = -k_B T \ln \pi_i + \text{const.}$$

# Eigenvalues and timescales

$$T\mathbf{r}_i = \lambda_i \mathbf{r}_i, \quad t_i = -\frac{\tau}{\ln \lambda_i}.$$

- $\lambda_1 = 1$  corresponds to equilibrium.
- Spectral gap indicates timescale separation.

# Chapman–Kolmogorov test

- Markovianity check:

$$T(n\tau) \approx T(\tau)^n.$$

- Agreement validates chosen lag time.
- Passing CK ensures that the implied kinetics remain invariant when propagation is computed at multiples of the base lag, confirming the MSM describes the same slow modes.

## Committor function

- Define reactant set  $A$  and product set  $B$ .
- Forward committer:

$$q_i = \sum_j T_{ij} q_j, \quad q_i = 0 \ (i \in A), \ q_i = 1 \ (i \in B).$$

- The committer is the probability to reach  $B$  before returning to  $A$  and defines reactive surfaces.

# Reactive fluxes

- TPT flux:

$$f_{ij} = \pi_i T_{ij} q_i (1 - q_j)$$

- Measures net current along transition tubes and highlights dominant pathways.

# Reactive flux

$$f_{ij} = \pi_i T_{ij} q_i (1 - q_j).$$

- Identifies dominant reactive pathways.

# References I