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

FACULTAT DE CIÈNCIES, TECNOLOGIA I ENGINYERIES
UVIC | UVIC-UCC

Contents

Notation used throughout the notes

- Phase-space point: $z = (q, p) \in \mathbb{R}^{3N} \times \mathbb{R}^{3N}$, with positions q and momenta p.
- Potential energy (force field): $U(q)$; forces $\mathbf{F}(q) = -\nabla_q U(q)$.
- Hamiltonian: $H(q, p) = K(p) + U(q)$ with $K(p) = \sum_i \frac{\|p_i\|^2}{2m_i}$.
- Temperature T, Boltzmann constant k_B, inverse temperature $\beta = (k_\mathrm{B} T)^{-1}$.
- Time step Δt; lag time for kinetic models τ.

Markov state models: from trajectories to a transition matrix

Choose a state discretization (clustering) $S(q) \in \{1, \ldots, n\}$ and a lag time τ. The MSM estimates a transition matrix $P(\tau)$ with entries

$$P_{ij}(\tau) = \mathbb{P}(S(q_{t+\tau}) = j \mid S(q_t) = i).$$

For reversible dynamics (detailed balance) one often enforces

$$\pi_i P_{ij} = \pi_j P_{ji},$$

which improves statistical efficiency and yields real eigenvalues. Implied timescales: $t_k = -\tau / \ln |\lambda_k|$ for eigenvalues $1 = \lambda_1 > \lambda_2 \geq \cdots$.

Suggested figure: Markov chain / HMM trellis diagram

Figure: A trellis diagram is a compact way to visualize a Markov process with observations (HMM), closely related to state discretization in MSMs. Source: Wikipedia article on Hidden Markov model.

Transform data

- simulateAmber.py and simulateCharmm.py produce DCD trajectories that feed into PyEMMA.
- Align, trim solvent, and save to HDF5 for storage efficiency.
- Energy and RMSD reports serve as checks before discretization.

Feature extraction

- Distances, dihedrals, and ligand–protein contacts form the vector $\mathbf{x}$.
- The covariance matrix $C = \langle (\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^T \rangle$ gets projected.
- All guide scripts (simulateGromacs, simulateTinker) share this output to feed PyEMMA.

tICA and time-lagged reduction

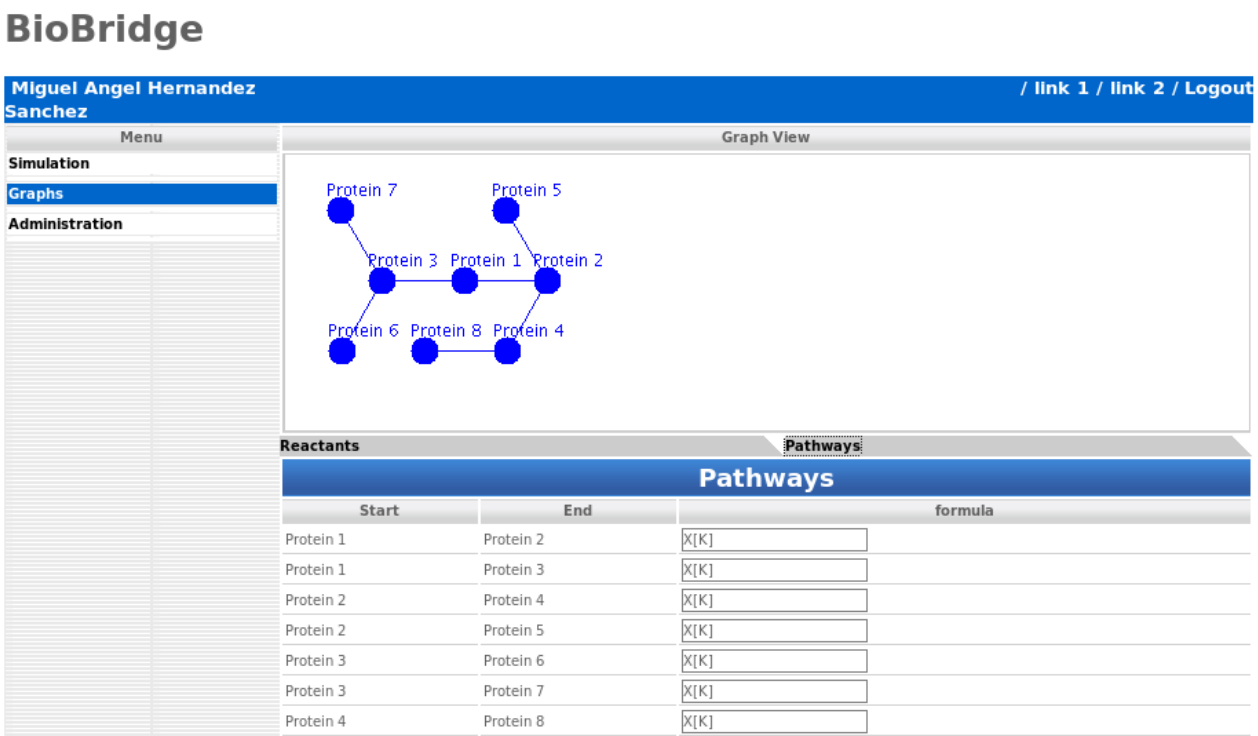
$$C_\tau \mathbf{w} = \lambda C_0 \mathbf{w}.$$

- Choose τ by comparing plateaus in implicit times $t_i = -\frac{\tau}{\ln \lambda_i}$.
- tICA captures the slow modes relevant for MSM and Deeptime.

Microstate clustering

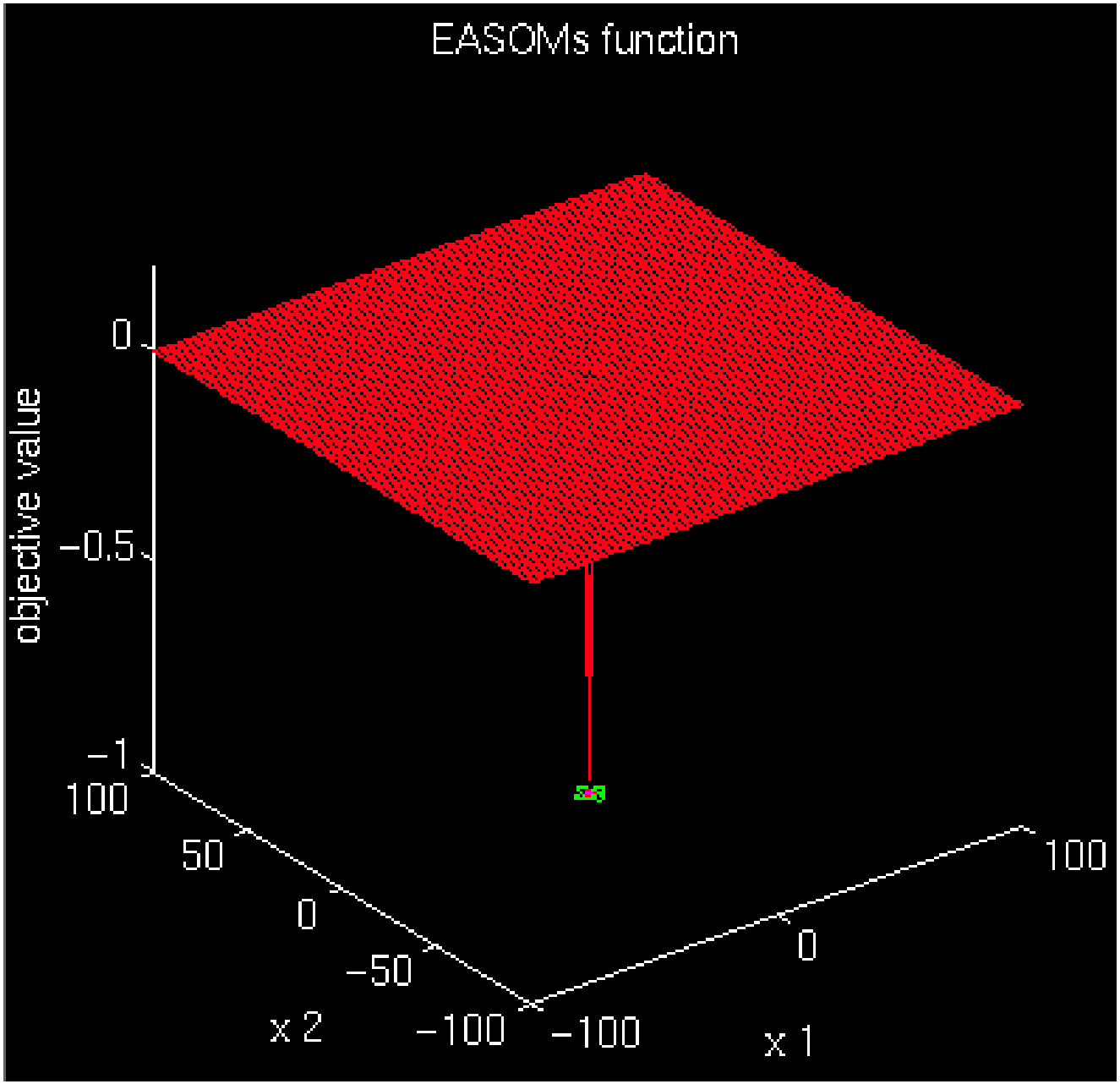
- k-means/minibatch create microstates; silhouette helps choose k.
- Favor dense states to avoid empty microstates (OpenMM User Guide §3.7).

State map



Transition graph influenced by the REST tutorial results. [1]

Energy landscape and partitions



Transition matrix

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

- Rows sum to 1 and the matrix becomes reversible stochastic when detailed balance is enforced.
- Use Bayesian estimators with pseudocounts to avoid zeros.

Markov property

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

- Chapman-Kolmogorov tests and pseudo-time diagnostics (PyEMMA) validate Markovianity.

Stationarity

$$\boldsymbol{\pi}^T T = \boldsymbol{\pi}^T.$$

- The stationary distribution π defines relative populations.
- Relate to free energies $F_i = -k_\mathrm{B} T \ln \pi_i$ and compare with alchemical ΔG.

Implied times

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

- Eigenvalues near 1 represent slow processes (Markovianity).
- Large gaps between t_i support scale separation.

Uncertainty and regularization

- Resample discretized trajectories to obtain reliable intervals.
- PyEMMA stores weights and allows reruns with a different seed.
- Enforce reversibility to reduce noise.
- Tune priors using forces detected in `simulateAmber.py`.

TPT, committor, and fluxes

- Flows $f_{ij} = \pi_i T_{ij} q_i (1 - q_j)$ identify reactive pathways.
- The committor $q_i = P(\text{reaching B before A} \mid X_0 = i)$ gauges progress.
- A transition graph colored by flux highlights kinetic bottlenecks.
- `simulateGromacs.py` provides the reference trajectories.

MSM checklist

- Informative features.
- Validated τ and eigenvalues with a clear gap.
- Error bootstrap and referenced reporting.

References I

[1] OpenMM Cookbook. *Grafos de estados en Replica Exchange Solute Tempering*. CC BY-SA 4.0. URL: https://openmm.github.io/openmm-cookbook/latest/notebooks/tutorials/Running_a_REST_simulation.html (visited on 01/12/2026).

[2] Visualización interna del curso. *Embudo energético de muestreo*. Figura adaptada del archivo function_funnel.png del repositorio. 2024.