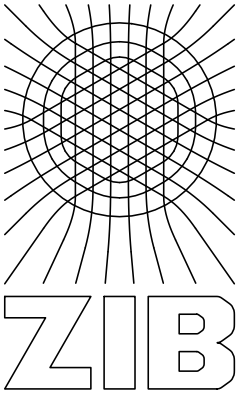
Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany



HAN WANG[1,2] AND CHRISTOF SCHÜTTE[1,3]

[1] *Department of Mathematics and Computer Science, Freie Universität Berlin, Germany*

[2] *CAEP Software Center for High Performance Numerical Simulation, Beijing, China*

[3] *Zuse Institute Berlin, Germany*

Building Markov State Models for Periodically Driven Non-Equilibrium Systems

ZIB-Report 14-46 (November 2014)

Herausgegeben vom
Konrad-Zuse-Zentrum für Informationstechnik Berlin
Takustraße 7
D-14195 Berlin-Dahlem

Telefon: 030-84185-0
Telefax: 030-84185-125

e-mail: bibliothek@zib.de
URL: http://www.zib.de

ZIB-Report (Print) ISSN 1438-0064
ZIB-Report (Internet) ISSN 2192-7782

Building Markov State Models for Periodically Driven Non-Equilibrium Systems

Han Wang[*,†,‡] and Christof Schütte[*,¶,‡]

CAEP Software Center for High Performance Numerical Simulation, Beijing, China, Zuse Institute Berlin (ZIB), Germany, and Institute for Mathematics, Freie Universität Berlin, Germany

E-mail: han.wang@fu-berlin.de; schuette@zib.de

Abstract

Recent years have seen an increased interest in non-equilibrium molecular dynamics (NEMD) simulations, especially for molecular systems with periodic forcing by external fields, e.g., in the context of studying effects of electromagnetic radiation on the human body tissue. Lately, an NEMD methods with local thermostating has been proposed that allows for studying non-equilibrium processes in a statistically reliable and thermodynamically consistent way. In this article, we demonstrate how to construct Markov State Models (MSMs) for such NEMD simulations. MSM building has been well-established for systems in equilibrium where MSMs with just a few (macro-)states allow for accurate reproduction of the essential kinetics of the molecular system under consideration. Non-equilibrium MSMs have been lacking so far. The article presents how to construct such MSMs and illustrates their validity and usefulness for the case of conformation dynamics of alanine dipeptide in an external electric field.

[*]To whom correspondence should be addressed
[†]CAEP Software Center for High Performance Numerical Simulation, Beijing, China
[‡]Zuse Institute Berlin (ZIB), Germany
[¶]Institute for Mathematics, Freie Universität Berlin, Germany

Keywords: Non-equilibrium, Markov states model, Alanine dipeptide, Electric field, Floquet theory

1 Introduction

Biomolecular systems under non-equilibrium conditions caused by external fields, especially systems under periodic forcing, have attracted increasing interest recently. For example, the potential effects of electromagnetic radiation on the human body tissue (e.g. DNA, protein, and membrane) has been extensively investigated in a vast number of articles, with the following list just representing an incomplete selection.[1–14] Molecular dynamics (MD) simulations have proved particularly useful for understanding the response of biomolecular conformations to external fields because of their ability to resolve molecular details that sometimes cannot be resolved in experiments. Only recently, a non-equilibrium MD simulation (D-NEMD) method with local thermostating has been proposed[15] that allows for studying non-equilibrium processes in a statistically reliable and thermodynamically consistent way. Despite the significance of the non-equilibrium phenomena, the analysis of the non-equilibrium MD simulations mainly follows standard approaches, and reliable tools for quantitative description of the essential conformational dynamics of the molecular system under external forcing are still lacking.

Despite their many advantages, MD simulations have severe limitations. For example, one has to assume that the underlying force fields are appropriately describing the internal and external molecular interactions, and the maximal possible simulation length often is shorter than the timescale of interest. This article is mainly concerned with circumventing the latter obstacle by introducing non-equilibrium Markov State Models. Markov State Models (MSM) have been well developed over the past decade in theory[16,17] and applications (see the recent book[18] for an overview), and software implementations,[19,20] but for systems under equilibrium conditions only! The principal idea of equilibrium MSMs is to

approximate the original high-dimensional MD system by a reduced Markovian dynamics over a finite number of (macro-)states. These (macro-)states have to be identified with the dominant metastable sets, in the sense that typical MD trajectories stay in the vicinity of a metastable set substantially longer than the systems needs for a transition to another such state.[16,21] In this case, the metastable sets are the main conformations of the molecular system under consideration which, often enough, are given by the main wells in the energy landscape. It has been shown that for molecular systems exhibiting such metastable sets the Markovian dynamics given by an MSM allows very close approximation of the longest relaxation processes of the underlying molecular system, at least under equilibrium conditions.[22,23] Moreover, it has been demonstrated that in such cases MSM building requires short MD trajectories only, much shorter than the timescales of interest.[24,25] Thus, MSM building often allows to study the dynamical behavior on long timescales without requiring MD trajectories of comparable length. Moreover, MSMs are utilized for understanding very long MD simulations: Extracting the essential structures and dynamical properties from long MD runs is becoming increasingly difficult as the system size and trajectory length grow; MSMs have been used to construct kinetic fingerprints from MD simulations[26] that help understanding the essential dynamics and in comparison with experiments.[27]

To the knowledge of the authors this work is the first attempt of using MSMs for analyzing non-equilibrium systems under periodic external forcing. More precisely, we will demonstrate how to use MSMs for investigate the conformational dynamics of a peptide (alanine dipeptide) under an oscillating electric field (EF). To this end, we will show how to generalize Markov state modeling to periodic non-equilibrium conditions where one cannot assume reversibility of the dynamics as it is mostly done in the literature on MSM building.

The outline of the article is as follows: In Sec. 2, we discuss the temporal and spatial discretizations needed to construct an MSM. We consider spatial discretization of the dihedral angle space in the traditional sense of full partition MSMs.[28,29] In the temporal direction, the non-equilibrium process is discretized utilizing Floquet's theorem. This results in a time-

homogeneous, but not necessarily reversible, that is, irreversible Markov process. Since full spatial partition does not make any sense for high-dimensional systems, we next show how to construct a *few-state* MSM based on milestoning[21,29] of the discretized irreversible Markov process in Sec. 3. The validity of the discretizations and the resulting MSM is checked in Sec. 4 by comparing the kinetic fingerprints given by the MSM to brute force non-equilibrium MD simulations of alanine dipeptide under oscillating EF. The findings are summarized in Sec. 5 including a chart presenting the workflow of MSM building for non-equilibrium systems, and a list of open questions.

2 Non-equilibrium molecular dynamics and its discretization

We consider diffusive molecular dynamics in an energy landscape V driven by the time-dependent external driving force $E(t)D(x_t)$ with the T-periodic external field $E(t)$:

$$dx_t = \Big(-\nabla V(x_t) + E(t)D(x_t)\Big)dt + \sqrt{2\beta^{-1}}dw_t, \tag{1}$$

where $x_t \in \Omega$ denotes the state of the molecular system at time t in state space Ω, w_t denotes standard n-dimensional Brownian motion, and β the inverse temperature, i.e. $\beta = 1/(k_B\mathcal{T})$. Thermostatted Hamiltonain or Langevin dynamics can be treated in the same way as explained herein, so for sake of simplicity we focus on the discussion of diffusive dynamics. The propagation of probability densities $\rho = \rho(x,t)$ based on this kind of dynamics in the sense of $\rho(x,t)dx = \mathrm{P}[x_t \in [x, x+dx)]$ is governed by Fokker-Planck equation:

$$\frac{\partial \rho}{\partial t} = \mathcal{L}^\dagger(t)\rho, \tag{2}$$

where $\mathcal{L}^\dagger(t)$ is the adjoint of the generator

$$\mathcal{L}(t) = \beta^{-1}\Delta_x + \Big(-\nabla_x V(x) + E(t)D(x)\Big) \cdot \nabla_x, \tag{3}$$

where Δ_x denotes the Laplacian operator and ∇_x the nabla-operator wrt to x. The periodicity of the external driving force induces the periodicity of the generator, i.e. $\mathcal{L}(t) = \mathcal{L}(t+T)$.

2.1 Spatial discretization: Master equation

We will now introduce an appropriate spatial discretization of this kind of non-equilibrium MD – this is done for reasons of simplicity only; we could completely avoid it for the price of more technical arguments. For achieving this discretization, we introduce a partition of state space Ω into a finite number of disjoint sets $\{\Omega_1, \cdots, \Omega_N\}$ satisfying $\Omega = \cup_i \Omega_i$, $\Omega_j \cap \Omega_j = \emptyset,\ \forall i \neq j$. Utilizing the procedure described in Ref.[30] the original Fokker-Planck equation (2) is discretized, resulting in a time-inhomogeneous Markov jump process in state space $S = \{1, \cdots, N\}$ with time-dependent rate matrix $L(t) \in \mathbb{R}^{N\times N}$ satisfying

$$\sum_{j=1}^{N} L_{ij}(t) = 0 \tag{4}$$

$$L_{ij}(t) \geq 0, \quad i \neq j \tag{5}$$

$$L_{ij}(t) = L_{ij}(t+T) \tag{6}$$

for all real time $t \geq 0$. Moreover, the rate matrix L has the form $L(t) = L_0 + E(t)L_1$ where $E(t)$ is periodic with period $T > 0$. In analogy to (2), the Markov jump process generated by $L(t)$ transports probability distributions according to the associated Master equation

$$\frac{dp(t)}{dt} = L^\top(t) \cdot p(t) \tag{7}$$

where $L^\top(t)$ denotes the matrix transpose of $L(t)$, $p(t)$ is an N-vector denoting the probability distribution on S at time t, $p(i,t)$, for example, the probability to be in state i (which corresponds to set Ω_i) at time t. As usual the properties (4) and (5) of $L(t)$ guarantee that the total probability mass is conserved, i.e., if $p(i,0) \geq 0$ componentwise, then $p(i,t) \geq 0$ and $\sum_i p(i,t) = \sum_i p(i,0)$. The temporal evolution of the probability distribution $p(t)$ can be formally written

$$p(t) = \Phi(t)p(0) \tag{8}$$

by using the associated propagator matrix $\Phi(t) \in \mathbb{R}^{N\times N}$ that solves

$$\frac{d}{dt}\Phi(t) = L^\top(t)\Phi(t), \quad \Phi(0) = \mathrm{Id}. \tag{9}$$

Since the last equation can be considered column-wise, the propagator matrix inherits column-wise conservation properties: $\Phi_{ij}(t) \geq 0$ and $\sum_{i=1}^{N} \Phi_{ij}(t) = 1$, that is, $\Phi^\top(t)$ is a stochastic matrix satisfying $\Phi^\top(t)e = e$ with $e = (1,\ldots,1)^\top \in \mathbb{R}^N$. Regarding these considerations, we find

$$\Phi_{ij}(t) = \mathrm{P}\left(X_t = i \mid X_0 = j\right), \tag{10}$$

where X_t denotes the Markov process generated by $L(t)$.

The discretization sets that we used to go from x_t and $\mathcal{L}(t)$ to X_t and $L(t)$, respectively, can be assumed to provide an arbitrarily fine partition of the original state space; then the transport properties of $L(t)$ are almost perfect approximations of the transport properties of $\mathcal{L}(t)$, in particular the approximation $p(i,t) \approx \mathrm{P}(x_t \in \Omega_i)$ is almost perfect.

2.2 Temporal discretization: Floquet theorem

As an effect of the periodicity of $L(t)$ the propagator $\Phi(t+T)$ satisfies

$$\Phi(t+T) = \Phi(t)\Phi(T), \tag{11}$$

for all $t \geq 0$. This can be seen by considering $Y(t) = \Phi(t+T)$. It satisfies

$$\frac{d}{dt}Y(t) = L^\top(t+T)Y(t) = L^\top(t)Y(t), \quad Y(0) = \Phi(T).$$

When we consider this identity column-wise and use the propagator property of $\Phi(t)$ we get $\Phi(t+T) = Y(t) = \Phi(t)\Phi(T)$. As a consequence of (11) we get for all integers $m = 0, 1, 2, \ldots$ that

$$\Phi(t+mT) = \Phi(t)\Phi^m(T). \tag{12}$$

In combination with Eq. (8), we therefore know the solution $p(t)$ of the Master equation for all $t \geq 0$, if we can compute $\Phi(t)$ for $t \in (0, T)$. This is known as the Floquet theorem.[31] In particular we get the long-term evolution of the propagator:

$$\Phi(mT) = \Phi^m(T), \tag{13}$$

where $\Phi^m(T)$ denotes the mth power of $\Phi(T)$. Thus, for the probability at integral periods we have

$$p(mT) = \Phi(mT)\, p(0) = \Phi^m(T)\, p(0). \tag{14}$$

Using the Floquet theorem, the time-inhomogeneous Markov jump process X_t is therefore discretized into a *time-homogeneous* (not necessarily reversible) Markov jump process $\tilde{X}_m =$

$X_{mT},\ m \in \mathbb{N}$, which is generated by transition matrix

$$P = \Phi^{\top}(T). \tag{15}$$

We prefer to consider the discrete-time process $\tilde{X}_m$ instead of the time-continuous process X_t because the powerful theories and computational tools for time-homogeneous Markov processes can be directly applied. It worth noting that many of these tools require the transition matrix P to satisfy the detailed balance condition. The computations in the Appendix A show that P will in general *not* satisfy this condition; in fact the deviation from reversibility can be estimated from the work of the periodic driving does to the system. There is no doubt that information within one period is lost by using this temporal discretization, however, information regarding the long-term behavior of the system on timescales much longer than the period will be perfectly described because of $\tilde{X}_m = X_{mT} \approx x_{mT}$ whenever our spatial discretization is fine enough. At the same time, the computational cost of generating $\tilde{X}_m$ is much less demanding than the brute force simulations of NEMD, which implies lower statistically uncertainty in calculating the observables of interest.

Since P is a stochastic matrix, its eigenvalues are contained in the unit circle in the complex plane, i.e., each eigenvalue λ (potentially complex-valued) satisfies $|\lambda| \leq 1$. Furthermore $\lambda = 1$ is an eigenvalue with right eigenvector $e = (1, \ldots, 1)^{\top}$ and a left eigenvector μ satisfying $\mu^{\top} P = \mu^{\top}$. From now on, we assume P to be irreducible and aperiodic such that the Perron-Frobenius theorem holds, so the eigenvector corresponding to the eigenvalue $\lambda = 1$ is non-negative componentwise, and unique (up to normalization $\sum_i \mu(i) = 1$). In this case μ is the stationary measure in the sense that $\mu^{\top} P^m = \mu^{\top},\ m \in \mathbb{N}$, and (more precisely) the asymptotic evolution of an initial probability distribution $p(t = 0)$ by the process satisfies $p^{\top}(0)\, P^m \to \mu^{\top},\ m \to \infty$, so that μ can be seen as the quasi-stationary distribution of the non-stationary process.

3 Markov State Model

If the discretization cells Ω_i, $i = 1, \ldots, N$ form a fine partition of the molecular state space, the Markov chain defined via the transition matrix P is discrete in time, but in space it still is a fine-scale description of the transport properties of the dynamics with a very large number N of states. Now we want to coarse our description much further by constructing a Markov State Model (MSM) for P with $K \ll N$ *macrostates* that should be the metastable states of the system: The resulting $K \times K$ MSM transition matrix $\hat{P}$ then defines the coarse grained long term kinetics that shall approximate the original long term kinetics well. The idea behind MSM building is that given the molecular system under consideration exhibits metastable conformations then it is usually possible to construct a relatively small number of discrete sets –the metastable sets that form the so-called macrostates– that correctly describe the slow dynamics, and in each set the fast dynamics relaxes on some timescales significantly shorter than the metastable timescales. Then if the MSM dynamics reproduces the slow timescales and the corresponding transitions of the original dynamics (1), the former is considered to be a good approximation of the latter.

MSM building has been attracted a lot of attention recently, and theory[16] as well as algorithms,[18] applications (see e.g.[24,32] for two of hundreds of articles) and software[19,20] have been developed to quite an extend. However, by far most of the literature is related to building *standard* MSMs for equilibrium MD. In standard MSM also the transition region has to be discretized, a feature that often forces the user to incorporate more macrostates than essentially needed to approximate the long-term kinetics. In Ref.[16,21,22,33] it has been shown how to construct *non-standard* MSM that avoid this problem for equilibrium MD, i.e., if P satisfies the detailed balance condition: (1) Identify the cores of the metastable sets of the dynamics, (2) use them as milestones to construct an MSM in which the macrostates are the metastable core sets and $\hat{P}$ is the transition matrix of the milestone process[16,21,29] that models the jumping behavior of the original dynamics between the metastable regions.

However, since we cannot assume P to satisfy detailed balance, we instead follow the

approach to non-standard MSMs recently proposed in Ref.[34] which allows to identify the metastable core sets for the non-reversible transition matrix P. Assume that this approach leads to the K core sets $C_1, \ldots, C_K \subset S$ that are appropriate metastable sets. Following[16,21] the process $(\tilde{X}_m)$ associated with P is coarse grained into the so-called milestone process $(\hat{X}_m)$ in the following way:

- $\hat{X}_m$ just has K states associated with the sets C_j, $j = 1, \ldots, K$.
- The sequence of random variables $(\hat{X}_m)$ is defined via the sequence $(\tilde{X}_m)$, i.e., trajectories of $(\tilde{X}_m)$ induce trajectories of $(\hat{X}_m)$: We set $\hat{X}_m = j$ if the last core set that the process $(\tilde{X}_m)$ entered prior to or at time m has been the core set C_j.

Now consider an arbitrary infinitely long trajectory of $(\tilde{X}_m)$. Because of ergodicity we know that the states in this trajectory will be distributed due to the quasi-stationary distribution μ. Based on such an infinitely long trajectory we can consider the probability $q_j^-(i)$ that conditioned on $\tilde{X}_m = i$ the last core set hit has been C_j. This function is called the backward committor of $(\tilde{X}_m)$ associated with the set C_j and is associated with the milestone process via

$$q_j^-(i) = \mathrm{P}_\mu(\hat{X}_m = j \mid X_m = i), \tag{16}$$

where the index μ refers to the fact that $\hat{X}_m$ is distributed due to μ. From the last equation we get that the stationary distribution of the milestone process is given by

$$\hat{\mu}_j = \sum_{i \in S} q_j^-(i)\mu(i), \tag{17}$$

that is, the probability to find $\hat{X}_m = j$ in (infinitely) long trajectories of the milestoning process is $\hat{\mu}_j$.

Following[35] one also has to consider the forward committor $q_j^+(i)$ identical to the probability that conditioned on $\tilde{X}_m = i$ the next core set to be hit will be C_j. The forward and

backward committors q_j^+ and q_j^- for each core set C_j can be computed from P by solving the linear equations[35]

$$\begin{aligned}(P - \mathrm{Id})q_j^+(i) &= 0, \quad i \in C \\ q_j^+(i) &= 1, \quad i \in C_j \\ q_j^+(i) &= 0, \quad i \in C_k, k \neq j\end{aligned} \tag{18}$$

where $C = S \setminus \cup_j C_j$, and

$$\begin{aligned}(P^b - \mathrm{Id})q_j^-(i) &= 0, \quad i \in C \\ q_j^-(i) &= 1, \quad i \in C_j \\ q_j^-(i) &= 0, \quad i \in C_k, k \neq j\end{aligned} \tag{19}$$

where P^b denotes the transition matrix of the time-reversed process given by $P^b_{ji} = \mu(i)P_{ij}/\mu(j)$.

We define the one-step transition matrix $\hat{P}$ for the milestone process by

$$\hat{P}_{jk} = \mathrm{P}_\mu\Big(\hat{X}_{m+1} = k \,|\, \hat{X}_m = j\Big). \tag{20}$$

Then Following Ref.,[36] Thm. 3.1, $\hat{P}$ can be computed by matrix multiplication using the committors:

$$\begin{aligned}\hat{P}_{jk} =& \frac{1}{\hat{\mu}_j}\langle (P^b - \mathrm{Id})q_j^-, q_k^+\rangle_\mu, \qquad j \neq k, \\ \hat{P}_{jj} =& 1 - \sum_{k \neq j} \hat{P}_{jk}\end{aligned} \tag{21}$$

where the inner product is defined by $\langle u, v\rangle_\mu = \sum_{i \in S} u(i)v(i)\mu(i)$. In general the milestone process need not be a Markov process. The results in[16,34] show, however, that it is an approximate Markov process as long as the core sets are proper metastable sets, i.e., if the typical timescale on which (X_m) leaves C is much smaller than the typical expected hitting

times between the core sets. Thus, by taking $\hat{P}$ as our MSM transition matrix, we introduce an additional modeling error that is the smaller the more metastable the core sets are. With this MSM transition matrix, we can define the MSM kinetics: If we start from some initial probability $\hat{p}_j(0)$ of being in state j at time $t = 0$ then its evolution $\hat{p}_j(t)$ in time is discrete in multiples of period T and given by

$$\hat{p}_j(T) = \sum_k \hat{p}_k(0)\hat{P}_{kj}. \tag{22}$$

In our approach $\hat{p}_j(mT)$ is a good approximation of $\mathrm{P}(\hat{X}_m = j)$ (for appropriately chosen core sets).

Remark 1: Our definition of the milestoning process in terms of the process $(\tilde{X}_m)$ generated by P guarantees that we can directly compute $\hat{P}$ via (20) from trajectories of $(\tilde{X}_m)$ without computing the committor functions. This is of importance if the spatial discretization underlying $(\tilde{X}_m)$ is fine enough, because then the kinetics of $(\tilde{X}_m)$ approximates the original kinetics of (x_{mT}) so that we can directly compute $\hat{P}$ via (20) from NEMD trajectories without computing P first (which substantially simplified the MSM building if the core sets are already known).

Remark 2: Following[16] we can also define another pair of stochastic MSM matrices:

$$\hat{T}_{jk} = \frac{\langle q_j^-, Pq_k^+\rangle_\mu}{\hat{\mu}_j}, \tag{23}$$

$$\hat{M}_{jk} = \frac{\langle q_j^-, q_k^+\rangle_\mu}{\hat{\mu}_j}, \tag{24}$$

that are connected to $\hat{P}$ by the following identity

$$\hat{P} = \hat{T} - \hat{M} + \mathrm{Id},$$

That can be seen by means of direct computation: For the off-diagonal entries we have

$$\hat{T}_{jk} - \hat{M}_{jk} = \frac{\langle q_j^-, Pq_k^+\rangle_\mu}{\hat{\mu}_j} - \frac{\langle q_j^-, q_k^+\rangle_\mu}{\hat{\mu}_j} = \frac{\langle (P^b - \mathrm{Id})q_j^-, q_k^+\rangle_\mu}{\hat{\mu}_j} = \hat{P}_{jk}, \quad j \neq k \tag{25}$$

Stochasticity yields $\hat{T}_{jj} - \hat{M}_{jj} + 1 = \hat{P}_{jj}$ for the diagonal entries. Furthermore, as shown in the Appendix B, $\hat{T}$ as well as $\hat{M}$ can be computed from trajectories without need to have the committors. The importance of the pair $\hat{T}$ and $\hat{M}$ for MSM building comes from the following observation: The main NEMD relaxation timescales are given by the dominant eigenvalues of P.[16,18] These dominant eigenvalues can be approximated by discretizing the related eigenvalue problem $Pu = \lambda u$ by means of a Galerkin approximation with the finite dimensional ansatz space spanned by the forward committors q_j^+, $j = 1, \ldots, K$ together with the finite dimensional test function space spanned by the backward committors q_j^-, $j = 1, \ldots, K$ (test functions multiplied from the left by the inner product $\langle \cdot, \cdot \rangle_\mu$). The thus discretize eigenproblem takes the form of a generalized eigenproblem

$$\hat{T}\hat{u} = \hat{\lambda}\hat{M}\hat{u}, \quad \text{or, equivalently} \quad \hat{M}^{-1}\hat{T}\hat{u} = \hat{\lambda}\hat{u}, . \tag{26}$$

For the reversible case it is known that its k eigenvalues $\hat{\lambda}$ are very good approximation of the dominant eigenvalues λ of the original problem if the core sets are proper metastable sets.[23] Whether this is true for the non-reversible case is not known yet, but if the deviation from reversibility is weak and the dominant eigenvalues of P are real-valued then the results should hold analogously, see,[16] Thm. 4.19.

If all of its entries are positive such that it is a stochastic matrix, $\hat{M}^{-1}\hat{T}$ thus can also be taken as MSM transition matrices. In the case of $\hat{M}^{-1}\hat{T}$ the MSM modeling error results from Galerkin discretization, while the MSM modeling error of $\hat{P}$ results from ignoring the potential non-Markovianity of $\hat{X}_m$.

Remark 3: As a matter of fact, if $\hat{X}_m$ were Markovian, the following identity would hold: $\hat{P} = \hat{T}\hat{M}^{-1}$ (see Appendix C for the proof). In this case, we would have $\hat{M}^{-1}\hat{T} = \hat{M}^{-1}\hat{P}\hat{M}$

and the eigenvalues of $\hat{P}$ and $\hat{M}^{-1}\hat{T}$ would be identical. Therefore, in practice, the deviation of the eigenvalues of $\hat{P}$ from those of $\hat{M}^{-1}\hat{T}$ indicates the deviation from Markovianity regarding the process $\hat{X}_m$.

4 Numerical example: Alanine dipeptide under oscillatory electric field

We know that –in theory– whenever the spatial discretization is fine enough, the Markov jump process X_t associated with the Master equation (7) is a good approximation to the original MD process x_t governed by (1). In practice, however, it is difficult to predict how many discrete sets we need to be fine enough. Moreover, since the total dimension of x_t is $3N_{\text{atom}}$ (N_{atom} being the number of atoms), it is prohibitive to do a really fine discretization over all degrees of freedom for most systems of practical interest.

One possible way to define appropriate discretization sets is firstly to find a few collective variables, and then to discretize these collective variables as finely as needed either by uniform or adaptive discretization.[17,37] However, it is difficult to give a general answer in prior regarding how to choose the collect variables and how fine their discretization should be. For large or high dimensional systems, these questions usually become non-trivial.

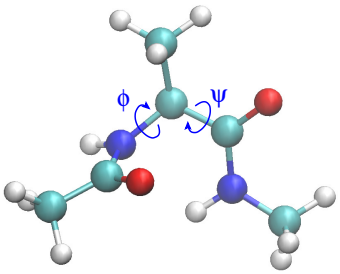


Figure 1: A schematic plot of the alanine dipeptide molecule and the dihedral angles ϕ and ψ.

To illustrate how the discretization works in practice we take the alanine dipeptide system under an oscillatory EF, as an example, the NEMD simulation of which was extensively

studied in Ref.[15] The system was simulated in a $2.7 \times 2.7 \times 2.7$ nm^3 periodic simulation region, with one alanine dipeptide molecules described by the CHARMM27 force field[38] dissolved in 641 TIP3P water molecules.[39] The grid-based energy correction map (CMAP)[40] was used to correct the backbone dihedral angle energies. All simulations were performed by a home-modified Gromacs 4.6.5[41] with CHARMM27 force field implemented.[42] The alanine dipeptide was put into the local thermostating environment, with a spherical dynamical region of radius 1.0 nm centered at the alpha-carbon. The Langevin thermostat with target temperature $\mathcal{T} = 300$ K and timescale $\tau_T = 0.1$ ps was coupled to the thermostated region. The whole system was coupled to the Parrinello-Rahman barostat[43] (in standard Gromacs implementation) with $\tau_P = 2.0$ ps to keep the system at 1 Bar. The non-equilibrium trajectories were integrated by the Leap-frog scheme with a time-step of 0.002 ps. The short-range van der Waals interactions were cut-off at 1.00 nm, and were smoothed from 0.95 nm to 1.00 nm by the "`shift`" method provide by Gromacs. The energy conserving Particle Mesh Ewald (PME)[44,45] method ("`pme-switch`") was used to compute the long-range electrostatic interaction, with the same real-space cut-off radius as the van der Waals interactions. The Gromacs default Fourier spacing of 0.12 nm and B-spline interpolation order of 4 were adopted. The splitting parameter was optimized with respect to the electrostatic force computing accuracy by Gromacs tool `g_pme_error`.[46] The neighbor list was updated every 5 time-steps with a list-building radius 1.20 nm. All hydrogen involving covalent bonds were constrained by the LINCS algorithm,[47] except the water molecules that were constrained by the SETTLE algorithm.[48] The whole system was driven by a periodic electric field $E(t) = E_0 \sin(2\pi t/T)$ and $D(x) = (1, 0, 0)^\top$ with intensity of the field being $E_0 = 1.0$ V/nm and period being $T = 10$ ps. The 20,000 branching trajectories were simulated from 20,000 initial configurations that sample the equilibrium distribution. The equilibrium configurations were prepared by an equilibrium MD simulation of length 10^6 ps, along which snapshots were saved every 50 ps. The branching NEMD trajectories were each 4,000 ps long, and the system reached non-equilibrium quasi-stationary state in roughly

300 ps.

For this periodically driven molecular system we will first show how to choose an appropriately fine spatial discretization. After validating this discretization we will consider the time-discretized dynamics generated by the Floquet transition matrix P in comparison to the original NEMD simulation. Finally we will coarse grain this description further by construction of a 3 state Markov State Model that is able to describe the long-term kinetics of the system correctly.

4.1 Spatial discretization

We choose the two dihedral angles ϕ and ψ as collective variables (see Fig. 1), and the discretization is a uniform partition of the ϕ–ψ plane. We denote the number of discretization intervals on each dihedral by N_{dih}, then we get $N = N_{\mathrm{dih}}^2$ discretization sets $\{\Omega_i\},\ i \in S = \{1, \cdots, N_{\mathrm{dih}}^2\}$.

Based on a given spatial discretization we can aggregate the transition matrix $P = \Phi^\top(T)$ just by counting the transition behavior of MD trajectories

$$P_{ij} = \mathrm{P}\left(X_T = j \mid X_0 = i\right). \tag{27}$$

However, P allows to approximate the original dynamics on multiples mT of the period only. In order to have a time-continuous description we need the generator $L(t)$ of the Master equation. If the discretization is fine enough one possible approximation to $L(t)$ is via the following forward finite difference scheme:

$$L_{ij}(t) \approx \frac{1}{\tau}\left[\,\mathrm{P}(X_{t+\tau} = j \mid X_t = i) - \delta_{ji}\,\right], \quad i, j \in S \tag{28}$$

and τ is an appropriate small enough lag-time. Since the dimensionality is reduced by using only a few collective variables, the lag-time should be chosen large enough so that the original dynamics x_t is properly relaxed with regard to the unresolved degrees of freedoms

on timescales shorter than the lag-time (assuming that the collective variables capture the slow dynamics). In the following we investigate the discretization quality with respect to the choice of N_{dih} and lag-time τ.

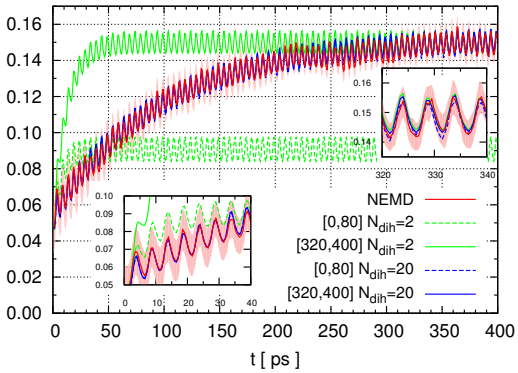


Figure 2: Time-dependent probability $\mathrm{P}\big(\phi_t \in [0, 180), \psi_t \in [0, 180)\big)$. The brute force NEMD simulation is compared with different spatial discretization methods. The red shadow region indicates the statistical uncertainty of the NEMD simulation.

We estimate the discretized generator $L(t),\ t \in [0, T)$ from NEMD trajectories generated by $\mathcal{L}(t)$ in different time intervals $[t_1, t_2]$. As discussed above, whenever the discretized dynamics approximates the original dynamics well, the time-periodic generator $L(t)$ should not depend on the choice of the interval $[t_1, t_2]$ in the estimation procedure (28), provided that the initial state of the system is not very far from the stationary state at long-time limit. Therefore, this is an indicator for calibrating the discretization quality. We compute $L(t)$ by two discretizations $N_{\text{dih}} = 2$ and $N_{\text{dih}} = 20$, and two choices of time intervals $[0, 80]$ ps and $[320, 400]$ ps, and then compare the time-dependent probability $\mathrm{P}\big(\phi_t \in [0, 180), \psi_t \in [0, 180)\big)$ with the (brute force) NEMD result in Fig. 2 using a lag-time $\tau = 0.5$ ps. Using $N_{\text{dih}} = 2$ the dynamics depends on the time interval used for calculating the generator: using time interval $[0, 80]$ ps the discretized dynamics deviates from the NEMD result, while using time interval $[320, 400]$ ps the discretized dynamics can only reproduces the NEMD result after 300 ps. This therefore indicates poor approximations to the original dynamics with $N_{\text{dih}} = 2$. The reason is that the discretization with $N_{\text{dih}} = 2$ is too coarse so that the dynamics cannot be fully equilibriated within the lag-time τ in each discretized set, therefore, the discretization

presents state dependency. For $N_{\text{dih}} = 20$, the discretized dynamics does not depend on the time interval of calculating the generator, and is consistent with the NEMD simulation within the error bar. Therefore, throughout this paper we use $N_{\text{dih}} = 20$ to discretize the dihedral angle space of alanine dipeptide.

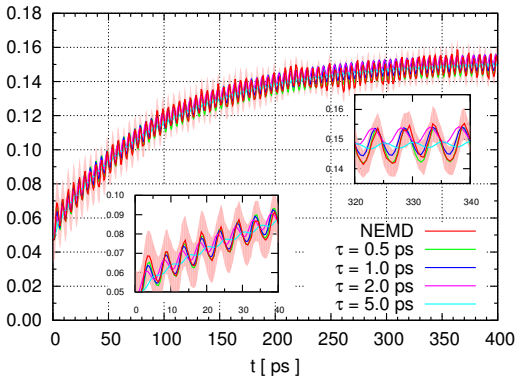


Figure 3: Time-dependent probability $\mathrm{P}\big(\phi \in [0, 180), \psi \in [0, 180)\big)$. The NEMD simulation is compared with the Master equation using generators discretized with $N_{\text{dih}} = 20$ and different lag time τ. The red shadow region indicates the statistical uncertainty of the NEMD simulation.

Next we discuss the effect of the lag time τ on the estimation of the generator. Therefore, we consider different choices of τ (0.5, 1.0, 2.0 and 5.0 ps) (see Fig. 3), all based on the identical dihedral angle discretization using $N_{\text{dih}} = 20$. It is clear that when the lag-time is close to the period (10 ps), the discretized dynamics cannot resolve the probability change within a period. However, it is surprising that even quite large lag-times are able to capture the the overall long time behavior of the original dynamics. We observe no significant difference between $\tau = 0.5$ and $\tau = 1.0$ ps, which means the discretized dynamics is not very sensitive to the choice of τ. Therefore, throughout this paper $\tau = 0.5$ ps will be used.

4.2 Quasi-stationary distribution μ

After having validated the fine-scale spatial discretization we will now consider the time-homogeneous process $\tilde{X}_m$ generated by the Floquet transition matrix $P = \Phi^{\top}(T)$, and investigate whether it reproduces the properties of the original non-equilibrium process x_t.

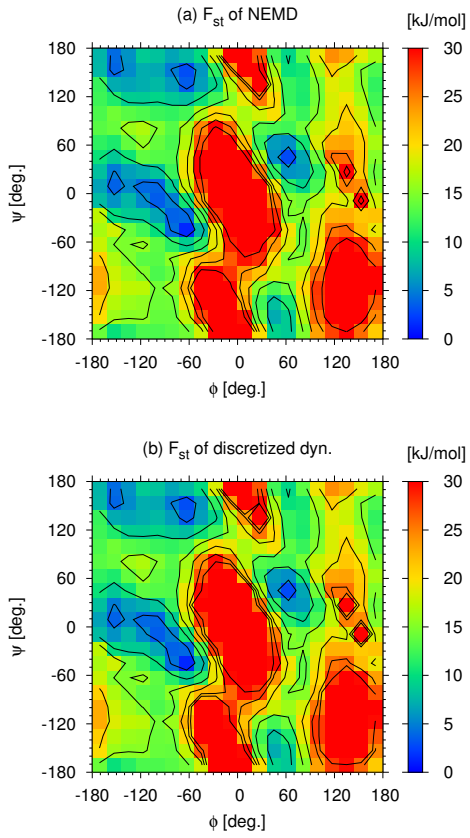


Figure 4: The color-scale plot of the logarithmic quasi-stationary distribution μ of (a) the NEMD and (b) the discretized dynamics governed by Floquet transition matrix $P = \Phi^{\top}(T)$.

In this context, only the configurations at the integral periods mT along the original process are taken into consider.

An important check is the consistency between the stationary probability density of $\Phi(T)$ (i.e. the leading eigenvector μ) and that estimated form the original NEMD simulation,

$$\rho_{\mathrm{st}}(\phi, \psi) = \lim_{m \to \infty} \rho(\phi, \psi, mT), \tag{29}$$

On each NEMD branching trajectory the initial 320 ps are discarded and the rest of the trajectory in time interval $[320, 4000]$ ps is averaged to estimate the quasi-stationary probability distribution ρ_{st}. P is computed as described above, and then μ is computed as its leading eigenvector. In order to make it comparable to the free energy in the equilibrium case, we take the logarithm of the distributions, i.e. $F_{\mathrm{st}}(\phi, \psi) = -k_B \mathcal{T} \log \rho_{\mathrm{st}}(\phi, \psi)$ for NEMD and $F_{\mathrm{st}}(\phi, \psi) = -k_B \mathcal{T} \log \mu(\phi, \psi)$ for P, where k_B is the Boltzmann constant and $\mathcal{T}$ is the temperature of the system. The results are compared in Fig. 4. A good consistency between the NEMD simulation and P is observed.

4.3 Core set identification

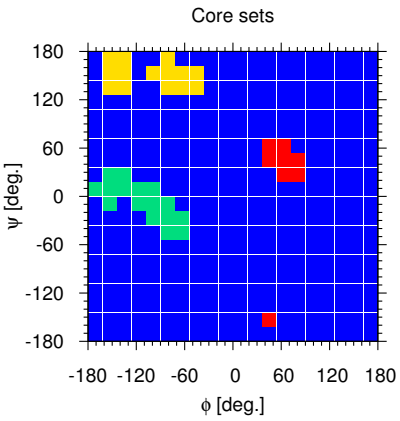


Figure 5: The core set identification. Different colors indicate different core sets: $C_{\alpha_{\mathrm{R}}}$ (green), C_β (yellow) and $C_{\alpha_{\mathrm{L}}}$ (red). The blue color indicates the transition region C that does not belong to one of the core sets.

The procedure for identifying good metastable core sets of the irreversible Markov process associated with P is described in detail in Ref.;[34] here we just provide the fundamental idea behind it: If strong metastable sets C_j, $j = 1, \ldots, K$ exist they should have one main property: When starting from a state in C_i the expected hitting time of a state in C_i should be much shorter than that of any state in one of the other sets C_j, $j \neq i$; in fact, the hitting time distribution should exhibit roughly constant levels in each set C_j and should vary significantly in the transition region $C = \Omega \setminus \cup_j C_j$ between the metastable sets. If starting from some randomly chosen initial states, one thus can identify the metastable core sets and the transition region by analyzing the hitting time distributions. This procedure is similar to the procedures used for reversible processes[17,18,49] but utilizes hitting time distributions instead of any eigenvector information.

The metastable core sets identified by this procedure based on the estimate of P are illustrated in Fig. 5 and denoted by C_{α_R} (green), C_β (yellow) and C_{α_L} (red). They correspond to the centers of the wells in the free energy landscapes shown in Fig. 4 and to the right-handed alpha-helix, beta-sheet and left-handed alpha-helix conformations of the peptide, respectively.

4.4 First mean hitting times

The first mean hitting time as a function of the dihedral angles (ϕ, ψ), is defined by the expected first time needed for hitting a certain core set C_j, $j \in \{\alpha_R, \beta, \alpha_L\}$ conditioned on starting from the conformation (ϕ, ψ), more exactly from equilibrium conformations $(\phi, \psi) \in \Omega_i$. Since the largest first mean hitting time (starting from states in core set α_R and hitting α_L) is longer than 600 ps the results will be biased if we use the NEMD trajectories of length 4000 ps for brute force Monte Carlo estimation of the hitting time. Therefore, we base our Monte Carlo estimate on 100 NEMD trajectories of 2×10^5 ps instead. For comparison we compute the first mean hitting time of the discretized dynamics via its transition matrix P: the first mean hitting time $h_{C_j}(i)$ of core set C_j starting in Ω_i can be computed by means of

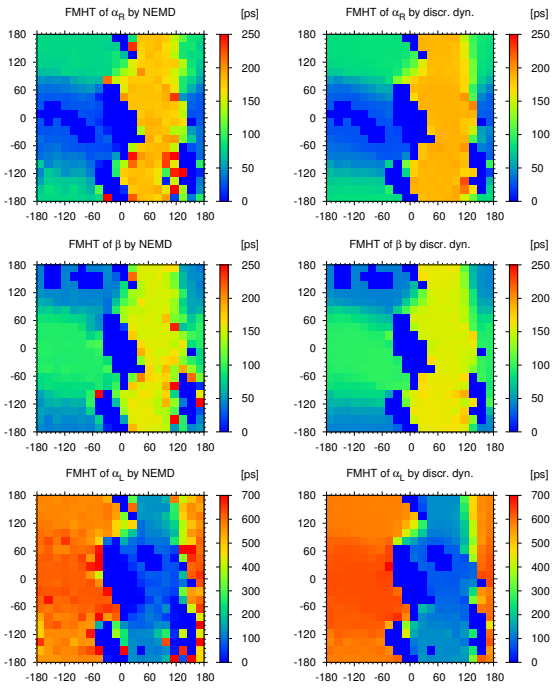


Figure 6: Comparisons of the first mean hitting time (FMHT) based on NEMD simulation (left column) and discretized dynamics (right column). From top to bottom the first mean hitting times to the core sets C_{α_R}, C_β, and C_{α_L} are shown, respectively

solving the linear problem[16]

$$(P - \mathrm{Id})h_{C_j}(i) = -1, \qquad \text{if } C_j \cap \Omega_i = \emptyset.$$

The resulting first mean hitting times are presented in Fig. 6. The good consistency between the NEMD estimate and the discretized Markov process $\tilde{X}_m$ indicates a good approximation quality. One should note that the NEMD estimate of the first mean hitting time is subject to statistical sampling errors while the first mean hitting times h_{C_j} only contain the statistical errors coming from the estimation of P. Thus, using P helps in calculating the observables in a smoother manner (less statistical error, no additional sampling). Additionally, the computational cost of the discretized process, if the cost for estimating $\Phi(T)$ is not included, is essentially smaller than NEMD: the computation of h_{C_j} is an issue of milliseconds on a laptop, while the NEMD trajectories took 1.6×10^4 core hours for Intel Xeon E5-4650 CPUs.

4.5 Forward and backward committors

Committors are very important statistical properties of Markov processes,[24,50] and play an important role in MSM building[16,29,36] (see below). Therefore, it is worth checking if the discretized process $\tilde{X}_m$ reproduces the NEMD committors. The forward committor $q_j^+(i)$ of a core set $C_j,\ j \in \{\alpha_\mathrm{R}, \beta, \alpha_\mathrm{L}\}$ is defined as the probability of visiting core set C_j next conditioned on starting at conformation $(\phi, \psi) \in \Omega_i$. The backward committor $q_j^-(i)$ of a core set $C_j,\ j \in \{\alpha_\mathrm{R}, \beta, \alpha_\mathrm{L}\}$ is defined as the probability of last coming from C_j conditioned on having arrived presently at configuration $(\phi, \psi) \in \Omega_i$. For reversible Markov processes, the forward and backward committors are identical, however, this is in general not the case for irreversible processes. The committors estimated from NEMD simulations (20000 trajectories, 4000 ps each) are compared with those computed from P by means of solving the linear equations (18) and (19). Fig. 7–9 presents both committors as well as their difference corresponding to different core sets. The committors of the discretized process

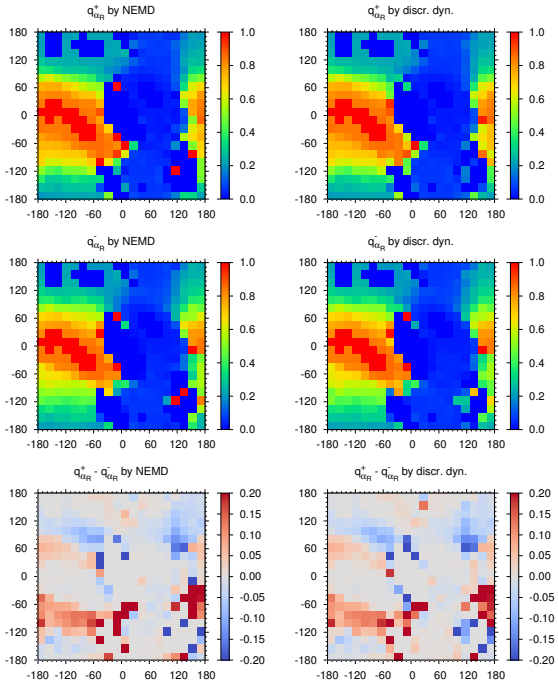


Figure 7: Forward $q^+_{\alpha_\mathrm{R}}$ and backward $q^-_{\alpha_\mathrm{R}}$ committors and their difference $q^+_{\alpha_\mathrm{R}} - q^-_{\alpha_\mathrm{R}}$ computed from the NEMD trajectories (left column) and the discretized dynamics (right column).

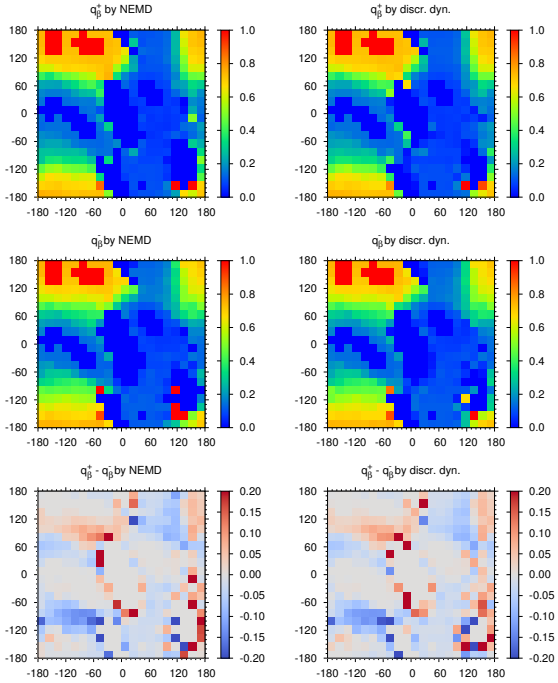


Figure 8: Forward q^+_β and backward q^-_β committors and their difference $q^+_\beta - q^-_\beta$ computed from NEMD trajectories (left column) and the discretized dynamics (right column).

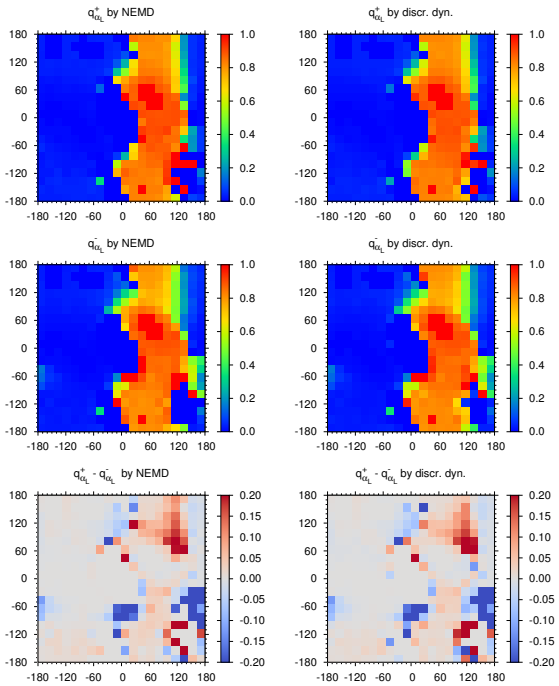


Figure 9: Forward $q^+_{\alpha_L}$ and backward $q^-_{\alpha_L}$ committors and their difference $q^+_{\alpha_L} - q^-_{\alpha_L}$ computed from NEMD trajectories (left column) and the discretized dynamics (right column).

are in good consistency with those of the NEMD simulations. The non-zero values in the committor differences indicate that the NEMD process, projected on the discretized dihedral angle space, is irreversible, and the discretized process is able to correctly describe this irreversiblity. In addition, and subsequently of central importance, the accurate reproduction of the committors indicates it is reasonable to build the MSM out of the committors of the discretized process.

4.6 MSM building and validation

Following the process described in Sec. 3, we are able to build a three state MSM for the externally driven alanine dipeptide system, where the quasi-stationary probability distribution μ, the three core sets, and the forward and backward committors are estimated as described above, and the MSM transition matrices $\hat{P}$, $\hat{M}$, and $\hat{T}$ are then evaluated using Eq. (21), (23) and (24), correspondingly. Alternatively, the MSM transition matrix $\hat{P}$ is calculated directly

from the NEMD trajectories using Eq. (20) (very good agreement with the one computed from the committors). The leading eigenvalues of $P = \Phi^{\top}(T)$ are compared with those of $\hat{P}$ and $\hat{M}^{-1}\hat{T}$ in Tab. 1. Without surprising, the two approaches for MSM building are consistent. The MSM is able to accurately reproduce the largest non-trivial eigenvalue, which means a precise reproduction of the longest non-trivial implied timescale. The accuracy of the second non-trivial timescale is not as good as the first, but is still acceptable. The reason for the lower accuracy may also be that the corresponding time scale is 26.5 ps (calculated by $-T/\log(\lambda_2)$), which is NOT significantly longer than the temporal resolution given by the period $T = 10$ ps of the external driving force. The difference between the eigenvalues of $\hat{P}$ and $\hat{M}^{-1}\hat{T}$ can be taken as an indication for the non-Markovianity of the milestone process $\hat{X}_m$. This non-Markovianity seems to have stronger influence on the second non-trivial timescale in comparison to the first one; this may be caused by shorter decorrelation timescales due to weaker metastability of the core sets involved. Numerically the two MSM transition matrices are:

$$\hat{M}^{-1}\hat{T} = \begin{pmatrix} 0.860 & 0.133 & 0.008 \\ 0.192 & 0.775 & 0.033 \\ 0.019 & 0.066 & 0.915 \end{pmatrix}, \qquad \hat{P} = \begin{pmatrix} 0.882 & 0.110 & 0.008 \\ 0.158 & 0.815 & 0.028 \\ 0.023 & 0.055 & 0.922 \end{pmatrix},$$

from which we see that the left-handed alpha-helix conformations of the peptide exhibits the strongest metastability.

It is worth noting that although the discretized process $\tilde{X}_m$ is irreversible, the MSM built out of it is almost reversible: The magnitude of the anti-symmetric part of the matrix $\mathrm{diag}(\hat{\mu}) \cdot \hat{P}$ is only of order 10^{-4}.

In fact, $\hat{P}$ can be considered as the fingerprint of the long-term kinetics (cf.[26,27]) of alanine dipeptide in an oscillatory electric field. In order to provide further validation of this

Table 1: Comparison of second and third eigenvalues of P and 3×3 MSM transition matrices $\hat{P}$ and $\hat{M}^{-1}\hat{T}$, respectively, from the two different MSM approaches.

	λ_2	λ_3	λ_4
P	0.905	0.668	0.551
$\hat{P}$	0.909	0.710	–
$\hat{M}^{-1}\hat{T}$	0.901	0.649	–

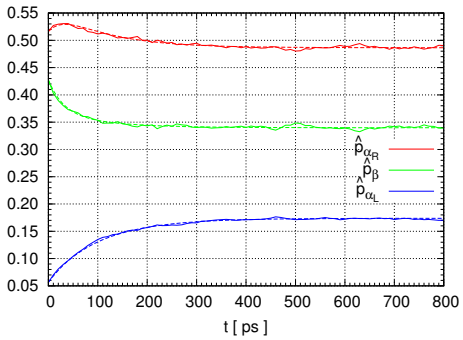


Figure 10: Comparison of NEMD and MSM long-term kinetics of alanine dipeptide in an oscillatory electric field. The plots show the time-dependent probability $\hat{p}_j(t)$ to be assigned to core set C_j (corresponding to the observable $\mathcal{A}$ given in (31) with $\alpha_j = 1$ and $\alpha_k = 0$ for $k \neq j$). Solid lines are from brute force NEMD simulations, while the dashed lines are from our MSM.

statement, we study time-dependent expectation values of the form

$$\mathcal{A}(t) = \langle A(i) \rangle_t = \sum_{i \in S} A(i) p(i,t), \tag{30}$$

where $p(i,t) = \mathrm{P}(X_t = i)$ is the probability to be in set Ω_i at time t as governed by the Master equation, and the observable $\mathcal{A}$ is spanned by the backward committors, i.e.,

$$A(i) = \sum_{j=1}^{K} \alpha_j q_j^-(i). \tag{31}$$

Then

$$\begin{aligned} \mathcal{A}(t) &= \sum_{i \in S} \sum_{j=1}^{K} \alpha_j q_j^-(i) p(i,t) = \sum_{i \in S} \sum_{j=1}^{K} \alpha_j \mathrm{P}(\hat{X}_t = j | X_t = i) \mathrm{P}(X_t = i) \\ &= \sum_{i \in S} \sum_{j=1}^{K} \alpha_j \mathrm{P}(\hat{X}_t = j, X_t = i) = \sum_{j=1}^{K} \alpha_j \mathrm{P}(\hat{X}_t = j) = \sum_{j=1}^{K} \alpha_j \, \hat{p}_j(t), \end{aligned} \tag{32}$$

where the time-dependent probability $\hat{p}_j(t)$ of being assigned to MSM macrostate j at time t can be computed by means of the MSM via simple matrix multiplications using (22).

In Fig. 10 we compare the numerical calculation of $\hat{p}_j(mT),\ m \in \mathbb{N}$ from NEMD and MSM calculations. In the NEMD case, the identity $\hat{p}_j(mT) = \sum_{i \in S} q_j^-(i) p(i, mT)$ is used, and the backward committor and the probability density on the R.H.S. are estimated directly from the NEMD trajectories. For the MSM, the projection of the initial probability is applied, $\hat{p}_j(0) = \sum_{i \in S} q_j^-(i) p(i,0)$, then the time-dependent probability at mT is generated by Eq. (22), i.e., by simple matrix multiplication. The agreement is almost perfect.

5 Concluding remarks and discussion

In this paper we proposed methods of MSM building for a periodically driven non-equilibrium system. We demonstrated their validity and performance by application to alanine dipeptide

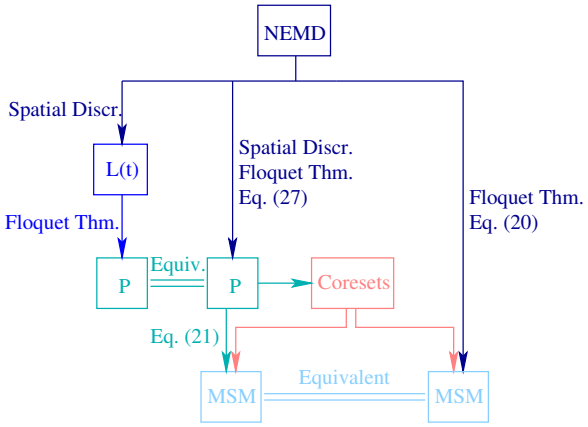


Figure 11: Flowchart showing optional procedures for MSM building based on NEMD trajectories. A lighter color indicates a "coarser" approximation to the original NEMD dynamics. Please note that "equivalence" only means that the respective procedures result in the same matrix/MSM if one assumes perfect sampling.

under an oscillatory electric field. We showed that the proposed methods allow for capturing the long-time behavior of the original non-equilibrium dynamics.

We provided effective methods for discretizing the original NEMD dynamics both temporally and spatially; the end-product is a time-homogeneous, in general irreversible Markov jump process. Discretization was done via two equivalent approaches: either by a two-step version, i.e. first spatial and then temporal discretization, or a one-step version that involves both discretizations simultaneously. These two version are shown by the left-most and middle branches of the diagram in Fig. 11.

Although the end-product of the two-step and one-step discretization procedures are formally equivalent, it is clear that the one-step discretization does not preserve any information *within* one period, because only the states at integral multiples of the period of the external forcing are considered. This is no serious problem whenever the long-time behavior of the system is of interest, and the corresponding timescales are significantly longer than one period. However, if the timescale of interest is comparable to the period, the two-step discretization is preferable because it allows to recover the dynamics between multiples of the period.

Building the final MSM based on the time-homogeneous discretized dynamics is straight-

forward by using Eq. (21) and a set of core sets that are derived from the discretized dynamics. In application to alanine dipeptide numerical results show that a three-state MSM can reproduce the leading non-trivial eigenvalue with very good accuracy, and the second non-trivial eigenvalue with acceptable accuracy. The lower accuracy regarding the second non-trivial eigenvalue may result from the fact that the second slowest timescale is not significantly longer than the period. By means of this three-state MSM we can reproduce, with almost perfect accuracy, the time-evolution of the population of the main conformations induced by the periodic forcing when starting from the equilibrium distribution of the unforced molecular system.

The right-most branch in Fig. 11 presents an equivalent, and seemingly much simpler alternative to the middle branch: MSM building directly from NEMD trajectories. In practice, however, this method may not be applicable, because it requires pre-defined core sets, and the identification of core sets usually is no trivial task, especially for molecular dynamics under non-equilibrium conditions. This task is substantially simplified when an accurate time-homogeneous discretization to the original NEMD process is available (that is, has been constructed in advance). In the present work, we computed the core sets by finding almost constant levels of the hitting time distribution for the discretized dynamics. This procedures is itself a novelty since it does not require any spectral information like eigenvectors as in standard approaches, cf.[17,18]

In this paper, we mainly focus on the development of the first available methods for MSM building in non-equilibrium systems. However, the application of these methods to the conformation dynamics of alanine dipeptide results in some observations that are interesting in itself: Under an oscillatory EF, the population of the left-handed α-helical conformation significantly increases relative to the equilibrium population (see also the discussions in Ref.[15]), and the leading relaxation timescale of the system is much shorter than in the equilibrium case.

A final remark concerning the utilization of the proposed methodology may be in order.

Based on an MSM for a given periodic forcing optimal control problems may come into reach: Using available methods like non-equilibrium linear response theory [51] or computational alchemy for MSMs [52] one can construct the MSM for slightly changed parameters (e.g., period and amplitude) of the external forcing by appropriate reweighting of the MSM for the given forcing. This in principle allows for answering questions like the following: For which parameters of the external forcing does one achieve maximal population of the left-handed α-helix? Such questions, however, will be treated in the future studies.

A Reversibility of the original dynamics

We consider the governing dynamics Eq. (1). For simplicity we denote the force by $F(x_t, t) = -\nabla_x V(x_t) + E(t)D(x_t)$. We denote $\sigma = \sqrt{2\beta^{-1}}$. According to Girsanov, we have

$$\frac{dp[x_t]}{dw[x_t]} = \exp\left\{\frac{1}{\sigma^2}\int_0^T F(x_t,t)dx_t - \frac{1}{2\sigma^2}\int_0^T F^2(x_t,t)dt\right\} \tag{33}$$

where dp is the probability measure of trajectory x_t, while dw is the probability measure of the standard Wiener process $dx_t = \sigma dw_t$. Assuming a discretization of the stochastic process at time $0 < t_1 < t_2 < \cdots < t_N = T$, where $t_i = iT/N$. We denote $x_i = x_{t_i}$, and $w_i = w_{t_i}$, then we have, in the sense of Ito,

$$\frac{dp[x_t]}{dw[x_t]} \approx \exp\left\{\frac{1}{\sigma^2}\sum_{i=0}^{N-1} F(x_i,t_i)(x_{i+1}-x_i) - \frac{1}{2\sigma^2}\sum_{i=0}^{N-1} F^2(x_i,t_i)\Delta t\right\} \tag{34}$$

Now, consider a conjugate trajectory $x_t^\dagger = x_{T-t}$ that starts at x_T, ends at x_0. The conjugate dynamics is driven by $F^\dagger(x_t^\dagger, t) = F(x_t^\dagger, T-t)$. Writing the Girsanov for the conjugate

dynamics

$$\frac{dp^\dagger[x_t^\dagger]}{dw[x_t^\dagger]} \approx \exp\left\{\frac{1}{\sigma^2}\sum_{i=0}^{N-1} F^\dagger(x_i^\dagger,t_i)(x_{i+1}^\dagger - x_i^\dagger) - \frac{1}{2\sigma^2}\sum_{i=0}^{N-1}[F^\dagger(x_i^\dagger,t_i)]^2\Delta t\right\} \tag{35}$$

$$= \exp\left\{\frac{1}{\sigma^2}\sum_{i=0}^{N-1} F(x_i^\dagger,T-t_i)(x_{i+1}^\dagger - x_i^\dagger) - \frac{1}{2\sigma^2}\sum_{i=0}^{N-1}[F(x_i^\dagger,T-t_i)]^2\Delta t\right\}$$

$$= \exp\left\{\frac{1}{\sigma^2}\sum_{i=0}^{N-1} F(x_{N-i},t_{N-i})(x_{N-i-1} - x_{N-i}) - \frac{1}{2\sigma^2}\sum_{i=0}^{N-1}[F(x_{N-i},t_{N-i})]^2\Delta t\right\}$$

$$= \exp\left\{\frac{1}{\sigma^2}\sum_{i=N}^{1} F(x_i,t_i)(x_{i-1} - x_i) - \frac{1}{2\sigma^2}\sum_{i=N}^{1} F^2(x_i,t_i)\Delta t\right\} \tag{36}$$

Since it is obvious that $dw[x_t^\dagger]/dw[x_t] = 1$,

$$\frac{dp^\dagger[x_t^\dagger]}{dw[x_t]} \approx \exp\left\{\frac{1}{\sigma^2}\sum_{i=1}^{N} F(x_i,t_i)(x_{i-1} - x_i) - \frac{1}{2\sigma^2}\sum_{i=1}^{N} F^2(x_i,t_i)\Delta t\right\} \tag{37}$$

The difference between the single trajectory probabilities is

$$\frac{dp^\dagger[x_t^\dagger]}{dp[x_t]} \approx \exp\left\{-\frac{1}{\sigma^2}\sum_{i=1}^{N-1}\left[F(x_i,t_i)(x_{i+1} - x_i) + F(x_i,t_i)(x_i - x_{i-1})\right]\right\} \tag{38}$$

Assuming the smoothness of the external perturbation, consider the differentiation:

$$\begin{aligned} F(x_i,t_i) - F(x_{i-1},t_{i-1}) &= F(x_i,t_i) - F(x_{i-1},t_i) + F(x_{i-1},t_i) - F(x_{i-1},t_{i-1}) \\ &= \nabla_x F(x_{i-1},t_i)(x_i - x_{i-1}) + \mathcal{O}(\Delta t) \\ &= \nabla_x F(x_{i-1},t_{i-1})(x_i - x_{i-1}) + \mathcal{O}(\Delta t) \end{aligned} \tag{39}$$

The second order expansion w.r.t. $x_i - x_{i-1}$ is of order Δt, so it is absorbed into $\mathcal{O}(\Delta t)$. Then the (38) becomes

$$\frac{dp^\dagger[x_t^\dagger]}{dp[x_t]} \approx \exp\left\{-\frac{2}{\sigma^2}\sum_{i=0}^{N-1} F(x_i,t_i)(x_{i+1} - x_i) - \frac{1}{\sigma^2}\sum_{i=0}^{N-1}\nabla_x F(x_i,t_i)(x_{i+1} - x_i)^2\right\} \tag{40}$$

Using the identity $dt = (dw_t)^2 = \sigma^{-2} dx_t^2$, Eq. (40) is written in the integral form

$$\frac{dp^\dagger[x_t^\dagger]}{dp[x_t]} \approx \exp\left\{ -\frac{2}{\sigma^2}\int_0^T F(x_t,t)dx_t - \int_0^T \nabla_x F(x_t,t)dt \right\} \tag{41}$$

One would not have the second integral on the exponent if the first integral of the exponent were defined in the sense of Stratonovich.

We notice that

$$\begin{aligned} dV(x,t) &= \frac{\partial V}{\partial x}dx + \frac{\partial V}{\partial t}dt \\ &= \frac{1}{2}\sigma^2\nabla_x^2 V dt + \nabla V dx_t + \frac{\partial V}{\partial t}dt \\ &= -\frac{1}{2}\sigma^2\nabla_x F dt - F dx_t + \frac{\partial V}{\partial t}dt \end{aligned} \tag{42}$$

Eq. (41) becomes

$$\frac{dp^\dagger[x_t^\dagger]}{dp[x_t]} = \exp\left\{ \frac{2}{\sigma^2}\left[V(x_T,T) - V(x_0,t_0) - \int_0^T \partial_t V(x_t,t)dt \right] \right\} \tag{43}$$

Take the limit of infinite small time interval, notice the equilibrium invariant probability density with respect to potential $V(x,0)$ satisfies $\mu(x) \propto e^{-\beta V(x,0)}$, and replace σ^2 by $2\beta^{-1}$,

$$\frac{dp^\dagger[x_t^\dagger]}{dp[x_t]} = \frac{\mu(x_0)}{\mu(x_T)} \times \exp\left\{ -\beta\int_0^T \partial_t V(x_t,t)dt \right\} \tag{44}$$

A.1 Irreversibility of the periodic time-symmetric dynamics

In Eq. (35), we assume the periodicity of the perturbation $F(x,t) = F(x,t+T)$, and the symmetry of the external perturbation, i.e. $F(x,-t) = F(x,t)$, we have

$$\frac{dp^\dagger[x_t^\dagger]}{dw[x_t^\dagger]} \approx \exp\left\{ \frac{1}{\sigma^2}\sum_{i=0}^{N-1} F(x_i^\dagger,t_i)(x_{i+1}^\dagger - x_i^\dagger) - \frac{1}{2\sigma^2}\sum_{i=0}^{N-1}[F(x_i^\dagger,t_i)]^2\Delta t \right\} \tag{45}$$

By changing notation $x^\dagger$ back to x, and comparing with (34), the reversed dynamics is subject to the Eq. (1), i.e. $dp^\dagger = dp$. Therefore

$$\begin{aligned} p(x_0, T|x_T, 0) &= \int_{\mathcal{C}\{x_T,0;x_0,T\}} dp[x_t^\dagger] \\ &= \int_{\mathcal{C}\{x_0,0;x_T,T\}} \frac{dp[x_t^\dagger]}{dp[x_t]} \cdot dp[x_t] \\ &= \int_{\mathcal{C}\{x_0,0;x_T,T\}} \frac{dp^\dagger[x_t^\dagger]}{dp[x_t]} \cdot dp[x_t] \\ &= \frac{\mu(x_0)}{\mu(x_T)} \int_{\mathcal{C}\{x_0,0;x_T,T\}} \exp\left\{ -\beta \int_0^T \partial_t V(x_t,t)dt \right\} \cdot dp[x_t] \end{aligned} \tag{46}$$

where $\mathcal{C}\{x_0, 0; x_T, T\}$ denotes all continuous trajectories starting at x_0 and ending at x_T. If $\partial_t V = 0$, i.e. equilibrium, we have

$$p(x_0, T|x_T, 0)e^{-\beta V(x_T,T)} = p(x_T, T|x_0, 0)e^{-\beta V(x_0,0)}, \tag{47}$$

which proves the reversibility of the equilibrium dynamics. The term

$$W[x_t] = \int_0^T \partial_t V(x_t, t)dt \tag{48}$$

is the non-equilibrium work associated to all possible the dynamics x_t starting at x_0 and ending at x_T (see e.g. Ref.[53]). Therefore Eq. (46) is the detailed Jarzynski relation. Noticing that

$$p(x_T, T|x_0, 0) = \int_{\mathcal{C}\{x_0,0;x_T,T\}} dp[x_t], \tag{49}$$

From Eq. (46) we have

$$\frac{p(x_0, T|x_T, 0)}{p(x_T, T|x_0, 0)} = \frac{\mu(x_0)}{\mu(x_T)} \mathbb{E}_{x_0 \to x_T}[e^{-\beta W}] \tag{50}$$

B Computation of $\hat{T}$ and $\hat{M}$ directly from trajectories

In order to show how to compute $\hat{T}$ and $\hat{M}$ directly from trajectories let us start by denoting the first hitting time of set A starting at time t by $h_t(A)$. In addition we define $\hat{q}_j^-(i) = \mathrm{P}(X_t = i|\hat{X}_t = j)$, and always assume $t = mT$. Then due to the Bayes' Theorem we have

$$\hat{q}_j^-(i) = \frac{\mathrm{P}(\hat{X}_t = j|X_t = i)\mathrm{P}(X_t = i)}{\mathrm{P}(\hat{X}_t = j)} = \frac{q_j^-(i)\mu(i)}{\hat{\mu}_j} \tag{51}$$

Then

$$\begin{aligned}
&\mathrm{P}[\, h_t(C_k) < h_t(\cup_{l\neq k} C_l) \,|\, \hat{X}_t = j \,] \\
&= \sum_{i=1}^{N} \frac{\mathrm{P}[\, h_t(C_k) < h_t(\cup_{l\neq k} C_l), X_t = i, \hat{X}_t = j \,]}{\mathrm{P}(\hat{X}_t = j)} \\
&= \sum_{i=1}^{N} \mathrm{P}[\, h_t(C_k) < h_t(\cup_{l\neq k} C_l) \,|\, X_t = i, \hat{X}_t = j \,]\, \hat{q}_j^-(i) \\
&= \sum_{i=1}^{N} \mathrm{P}[\, h_t(C_k) < h_t(\cup_{l\neq k} C_l) \,|\, X_t = i \,]\, \hat{q}_j^-(i) \\
&= \sum_{i=1}^{N} q_k^+(i) \frac{q_j^-(i)\mu(i)}{\hat{\mu}_j} = \frac{\langle q_j^-, q_k^+ \rangle_\mu}{\hat{\mu}_j} = \hat{M}_{jk}
\end{aligned} \tag{52}$$

where the third equation holds because of the Markovianity of the process X_t, and the fourth equation is due to the definition of the forward committor. The interpretation of the result is that when starting with the milestone process in state j we have to determine the fraction of all trajectories that hit core set C_k first of all core sets to estimate $\hat{M}_{jk}$.

To see the same for $\hat{T}$, we first need

$$\begin{aligned}
&\mathrm{P}[\, h_{t+T}(C_k) < h_{t+T}(\cup_{l\neq k} C_l) \,|\, X_t = i \,] \\
&= \sum_{l=1}^{N} \frac{\mathrm{P}[\, h_{t+T}(C_k) < h_{t+T}(\cup_{l\neq k} C_l), X_{t+T} = l, X_t = i \,]}{\mathrm{P}(X_t = i)} \\
&= \sum_{l=1}^{N} \mathrm{P}[\, h_{t+T}(C_k) < h_{t+T}(\cup_{l\neq k} C_l) | X_{t+T} = l, X_t = i \,] \, \mathrm{P}(X_{t+T} = l | X_t = i) \\
&= \sum_{l=1}^{N} \mathrm{P}[\, h_{t+T}(C_k) < h_{t+T}(\cup_{l\neq k} C_l) | X_{t+T} = l \,] \, \mathrm{P}(X_{t+T} = l | X_t = i) = \sum_{l=1}^{N} q_k^+(l) P_{il}
\end{aligned}$$

We used the Markovianity of X_T at the third equation, and the time-homogeneity in the fourth equation. Therefore, following the same procedure as before we have

$$\mathrm{P}[\, h_{t+T}(C_k) < h_{t+T}(\cup_{l\neq k} C_l) \,|\, \hat{X}_t = j \,] = \sum_{i=1}^{N} \sum_{l=1}^{N} q_k^+(l) P_{il} \frac{q_j^-(i)\mu(i)}{\hat{\mu}_j} = \frac{\langle q_j^-, P q_k^+ \rangle_\mu}{\hat{\mu}_j} = \hat{T}_{jk} \quad (53)$$

Thus, when starting with the milestone process in state j at some time t we have to determine the fraction of all trajectories of length at least T that hit core set C_k first of all core sets to estimate $\hat{T}_{jk}$.

C Prove of the identity $\hat{P} = \hat{T}\hat{M}^{-1}$

In this section we prove the identity $\hat{P} = \hat{T}\hat{M}^{-1}$. Starting from the result Eq. (53),

$$\begin{aligned}
\hat{T}_{jk} &= \mathrm{P}[\, h_{t+T}(C_k) < h_{t+T}(\cup_{l\neq k} C_l) \,|\, \hat{X}_t = j \,] \\
&= \frac{\mathrm{P}[\, h_{t+T}(C_k) < h_{t+T}(\cup_{l\neq k} C_l), \hat{X}_t = j \,]}{\mathrm{P}(\hat{X}_t = j)} \\
&= \sum_l \frac{\mathrm{P}[\, h_{t+T}(C_k) < h_{t+T}(\cup_{l\neq k} C_l), \hat{X}_{t+T} = l, \hat{X}_t = j \,]}{\mathrm{P}(\hat{X}_t = j)} \\
&= \sum_l \mathrm{P}[\, h_{t+T}(C_k) < h_{t+T}(\cup_{l\neq k} C_l) | \hat{X}_{t+T} = l, \hat{X}_t = j \,] \hat{P}_{jl} \\
&= \sum_l \mathrm{P}[\, h_{t+T}(C_k) < h_{t+T}(\cup_{l\neq k} C_l) | \hat{X}_{t+T} = l \,] \hat{P}_{jl} \\
&= \sum_l \hat{P}_{jl} \hat{M}_{lk}.
\end{aligned}$$

This proves the identity $\hat{P} = \hat{T}\hat{M}^{-1}$. Notice that in the fifth equation we assumes the Markovianity of the milestone process $\hat{X}_m$. In the sixth equation we use the result Eq. (52), and the time-homogeneity of the milestone process.

References

(1) Henrik Bohr and Jakob Bohr. Microwave-enhanced folding and denaturation of globular proteins. *Phys. Rev. E*, 61(4):4310, 2000.

(2) Henrik Bohr and Jakob Bohr. Microwave enhanced kinetics observed in ord studies of a protein. *Bioelectromagnetics*, 21(1):68–72, 2000.

(3) David de Pomerai, Clare Daniells, Helen David, Joanna Allan, Ian Duce, Mohammed Mutwakil, David Thomas, Phillip Sewell, John Tattersall, Don Jones, and Peter Candido. Cell biology: Non-thermal heat-shock response to microwaves. *Nature*, 405:417–418, 2000.

(4) David I de Pomerai, Brette Smith, Adam Dawe, Kate North, Tim Smith, David B Archer, Ian R Duce, Donald Jones, and E Peter M Candido. Microwave radiation can alter protein conformation without bulk heating. *FEBS letters*, 543(1):93–97, 2003.

(5) Fabrizio Mancinelli, Michele Caraglia, Alberto Abbruzzese, Guglielmo d'Ambrosio, Rita Massa, and Ettore Bismuto. Non-thermal effects of electromagnetic fields at mobile phone frequency on the refolding of an intracellular protein: Myoglobin. *J. Cell. Biochem.*, 93(1):188–196, 2004.

(6) Peter D Inskip, Robert E Tarone, Elizabeth E Hatch, Timothy C Wilcosky, William R Shapiro, Robert G Selker, Howard A Fine, Peter M Black, Jay S Loeffler, and Martha S Linet. Cellular-telephone use and brain tumors. *N. Engl. J. Med.*, 344(2):79–86, 2001.

(7) I. Bekard and D.E. Dunstan. Electric field induced changes in protein conformation. *Soft Matter*, pages –, 2014.

(8) A. Budi, F.S. Legge, H. Treutlein, and I. Yarovsky. Electric field effects on insulin chain-b conformation. *J. Phys. Chem. B*, 109(47):22641–22648, 2005.

(9) A. Budi, F.S. Legge, H. Treutlein, and I. Yarovsky. Effect of frequency on insulin response to electric field stress. *J. Phys. Chem. B*, 111(20):5748–5756, 2007.

(10) Akin Budi, F Sue Legge, Herbert Treutlein, and Irene Yarovsky. Comparative study of insulin chain-b in isolated and monomeric environments under external stress. *J. Phys. Chem. B*, 112(26):7916–7924, 2008.

(11) Loukas G Astrakas, Christos Gousias, and Margaret Tzaphlidou. Structural destabilization of chignolin under the influence of oscillating electric fields. *J. Appl. Phys.*, 111(7):074702–074702, 2012.

(12) Markus Damm, Christoph Nusshold, David Cantillo, Gerald N Rechberger, Karl Gruber, Wolfgang Sattler, and C Oliver Kappe. Can electromagnetic fields influence the

structure and enzymatic digest of proteins? a critical evaluation of microwave-assisted proteomics protocols. *J. Proteomics*, 75:5533–5543, 2012.

(13) N.J. English, G.Y. Solomentsev, and P. O'Brien. Nonequilibrium molecular dynamics study of electric and low-frequency microwave fields on hen egg white lysozyme. *J. Chem. Phys.*, 131:035106, 2009.

(14) G.Y. Solomentsev, N.J. English, and D.A. Mooney. Effects of external electromagnetic fields on the conformational sampling of a short alanine peptide. *J. Comput. Chem.*, 33:917–923, 2012.

(15) Han Wang, Christof Schütte, Giovanni Ciccotti, and Luigi Delle Site. Exploring the conformational dynamics of alanine dipeptide in solution subjected to an external electric field: A nonequilibrium molecular dynamics simulation. *Journal of Chemical Theory and Computation*, 10(4):1376–1386, 2014.

(16) Ch. Schütte and M. Sarich. *Metastability and Markov State Models in Molecular Dynamics: Modeling, Analysis, Algorithmic Approaches*, volume 24 of *Courant Lecture Notes*. American Mathematical Society, December 2013.

(17) J.H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J.D. Chodera, C. Schütte, and F. Noé. Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.*, 134:174105, 2011.

(18) G. R. Bowman, V. S. Pande, and F. Noé, editors. *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*, volume 797 of *Advances in Experimental Medicine and Biology*. Springer, 2014.

(19) M. Senne, B. Trendelkamp-Schroer, A. S. J. S. Mey, Ch. Schütte, and F. Noé. Emma - a software package for Markov model building and analysis. *J. Chem. Theory Comput.*, 8:2223–2238, 2012.

(20) Kyle A Beauchamp, Gregory R Bowman andThomas J Lane, Lutz Maibaum, Imran S Haque, and Vijay S Pande. MSMBuilder2: Modeling conformational dynamics at the picosecond to millisecond scale. *J Chem Theor Comput*, 2011.

(21) C. Schütte, F. Noé, J. Lu, M. Sarich, and E. Vanden-Eijnden. Markov State Models based on Milestoning. *J. Chem. Phys.*, 134:204105, 2011.

(22) Marco Sarich, Frank Noé, and Christof Schütte. On the approximation quality of Markov state models. *Multiscale Modeling & Simulation*, 8(4):1154–1177, 2010.

(23) N. Djurdjevac, M. Sarich, and Ch Schütte. Estimating the eigenvalue error of Markov state models. *Multiscale Modeling & Simulation*, 10(1):61–81, 2012.

(24) F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl. Constructing the full ensemble of folding pathways from short off-equilibrium simulations. *Proc. Natl. Acad. Sci. USA*, 106:19011–19016, 2009.

(25) Kai J Kohlhoff, Diwakar Shukla, Morgan Lawrenz, Gregory R Bowman, David E Konerding, Dan Belov, Russ B Altman, and Vijay S Pande. Cloud-based simulations on google exacycle reveal ligand modulation of gpcr activation pathways. *Nature chemistry*, 6(1):15–21, 2014.

(26) B. G. Keller, J.-H. Prinz, and F. Noé. Markov models and dynamical fingerprints: Unraveling the complexity of molecular kinetics. *Chem. Phys.*, 396:92–107, 2012.

(27) Jan-Hendrik Prinz, Bettina G. Keller, and Frank Noé. Probing molecular kinetics with Markov models: Metastable states, transition pathways and spectroscopic observables. *Phys. Chem. Chem. Phys.*, 13:16912–16927, 2011.

(28) V.S. Pande, K. Beauchamp, and G.R. Bowman. Everything you wanted to know about Markov state models but were afraid to ask. *Methods*, 52(1):99–105, 2010.

(29) M. Sarich, Ralf Banisch, C. Hartmann, and Ch. Schütte. Markov state models for rare events in molecular dynamics. *Entropy (Special Issue)*, 16(1):258–286, December 2013.

(30) Juan C Latorre, Philipp Metzner, Carsten Hartmann, and Christof Schütte. A structure-preserving numerical discretization of reversible diffusions. *Commun. Math. Sci*, 9(4):1051–1072, 2011.

(31) G Floquet. Sur les équations différentielles linéaires à coefficients périodiques. *Annales scientifiques de l'École Normale Supérieure*, 12:47–82, 1883.

(32) F. Noé, H. Wu, J.-H. Prinz, and N. Plattner. Projected and hidden Markov models for calculating kinetics and metastable states of complex molecules. *J. Chem. Phys.*, 139:184114, 2013.

(33) Nicaolae V. Buchete and Gerhard Hummer. Coarse master equations for peptide folding dynamics. *J. Phys. Chem. B*, 112:6057–6069, 2008.

(34) Marco Sarich and Christof Schütte. Utilizing hitting times for finding metastable sets in non-reversible Markov chains. *submitted to the Journal of Computational Dynamics*, 2014.

(35) P. Metzner, Ch. Schütte, and E. Vanden-Eijnden. Transition path theory for Markov jump processes. *Multiscale Modeling and Simulation*, 7(3):1192–1219, 2009.

(36) N. Djurdjevac, M. Sarich, and C. Schütte. On Markov state models for metastable processes. In *Proceedings of the International Congress of Mathematicians*, volume 901, pages 3105–3131, 2010.

(37) J.D. Chodera, N. Singhal, V.S. Pande, K.A. Dill, and W.C. Swope. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J. Chem. Phys.*, 126:155101, 2007.

(38) Nicolas Foloppe and Alexander D MacKerell Jr. All-atom empirical force field for nucleic acids: I. parameter optimization based on small molecule and condensed phase macromolecular target data. *J. Comput. Chem.*, 21(2):86–104, 2000.

(39) William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2):926–935, 1983.

(40) Alexander D MacKerell, Michael Feig, and Charles L Brooks III. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.*, 25(11):1400–1415, 2004.

(41) S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M.R. Shirts, J.C. Smith, P.M. Kasson, D. van der Spoel, Hess B., and Lindahl E. Gromacs 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, pages 1–10, 2013.

(42) Pär Bjelkmar, Per Larsson, Michel A Cuendet, Berk Hess, and Erik Lindahl. Implementation of the charmm force field in gromacs: Analysis of protein stability effects from correction maps, virtual interaction sites, and water models. *J. Chem. Theory Comput.*, 6(2):459–466, 2010.

(43) M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.*, 52:7182, 1981.

(44) T. Darden, D. York, and L. Pedersen. Particle mesh Ewald: An $N \log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.*, 98:10089, 1993.

(45) U. Essmann, L. Perera, M.L. Berkowitz, T. Darden, H. Lee, and L.G. Pedersen. A smooth Particle Mesh Ewald method. *J. Chem. Phys.*, 103(19):8577, 1995.

(46) Han Wang, Florian Dommert, and Christian Holm. Optimizing working parameters of the smooth particle mesh ewald algorithm in terms of accuracy and efficiency. *The Journal of chemical physics*, 133(3):034117, 2010.

(47) B. Hess, H. Bekker, H.J.C. Berendsen, and J.G.E.M. Fraaije. LINCS: a linear constraint solver for molecular simulations. *J. Comput. Chem.*, 18(12):1463–1472, 1997.

(48) S. Miyamoto and P.A. Kollman. SETTLE: an analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.*, 13(8):952–962, 2004.

(49) P. Deuflhard and M. Weber. Robust Perron cluster analysis in conformation dynamics. *Linear Algebra and its Applications*, 161(184), 2005. 398 Special issue on matrices and mathematical biology.

(50) Jan-Hendrik Prinz, Martin Held, Jeremy C. Smith, and Frank Noé. Efficient computation of committor probabilities and transition state ensembles. *SIAM Multiscale Model. Simul.*, 9:545, 2011.

(51) Han Wang, Carsten Hartmann, and Christof Schütte. Linear response theory and optimal control for a molecular system under non-equilibrium conditions. *Molecular Physics*, 111(22-23):3555–3564, 2013.

(52) Christof Schütte, Adam Nielsen, and Marcus Weber. Markov state models and molecular alchemy. *Molecular Physics*, ahead-of-print, 2014.

(53) Udo Seifert. Stochastic thermodynamics, fluctuation theorems and molecular machines. *Reports on Progress in Physics*, 75(12):126001, 2012.