Markov field models of molecular kinetics

Dissertation
zur Erlangung des Grades eines Doktors der Naturwissenschaften

am Fachbereich Physik der Freien Universität Berlin

vorgelegt von
Tim Hempel

Berlin, 2023

Erstgutachter: Prof. Dr. Frank Noé
Zweitgutachter: Prof. Dr. Roland Netz
Tag der Disputation: 04.09.2023

©2023 – Tim Hempel
all rights reserved.

This work is dedicated to Gerhard Hempel (1950 – 2021).

Contents

Since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice when there are tigers abroad.

George E.P. Box [1]

1 Introduction

In molecular biology, many relevant dynamic processes happen on timescales of nanoseconds to seconds and are governed by spatial rearrangement on length-scales smaller than nanometers. However, experiments have limited spatiotemporal resolution and cannot follow the time-course of complex many-particle systems such as whole proteins in the necessary atomistic detail. Therefore, the mechanism of protein function can often not be observed by an experiment alone. Computer simulations provide a possible remedy. Not being subject to physical barriers such as the diffraction limit of optical microscopy, they provide a picture of arbitrary spatiotemporal resolution that is only bounded by the availability of compute power.

Atomistic molecular dynamics (MD) simulations, based on empirical definitions of atomistic interaction terms, are commonly used to obtain *in silico* dynamic models. The resulting simulation data can be analyzed with Markov state models (MSMs), yielding a quantitative kinetic model that, e.g., encodes state populations and transition rates. Both approaches have been successfully combined in the past to quantify the dynamics and to push the timescale boundary of classical MD simulations that is posed by compute power. However, as compute power grows and the size of investigated systems increases, another fundamental scaling problem of the MD/MSM approach arises. Rooted in the representation of the system with global descriptors, ever more sampling is necessary in order to estimate a valid dynamical model. In this work, it is shown that this scaling problem can be escaped when leveraging weak couplings between local subsystems or domains of a studied system. Decomposing a system into weakly coupled or even independent local domains drastically decreases the amount of statistics required to adequately parameterize models and opens up avenues to previously inaccessible system sizes.

Our approach, termed independent Markov decomposition (IMD), is a first-order approximation neglecting couplings, i.e., it represents a decomposition of the underlying global dynamics into a set of independent local ones. Using the example of a (truly uncoupled) ion channel, we demonstrate that the sampling necessary to obtain valid models can be reduced by three orders of magnitude. Furthermore, IMD is applied to two biomolecular systems, demonstrating its applicability to high-dimensional prob-

lems with sparse sampling. First, synaptotagmin-1 is analyzed, a rapid calcium switch from the neurotransmitter release machinery. Within its C2A domain, local conformational switches are identified and modeled with independent MSMs. Furthermore, their allosteric interplay is analyzed using information-theoretic measures like transfer entropy. This multiscale model sheds light on activation of the C2A domain and shows how populations, rates, and allosteric mechanisms change as a function of calcium binding. Second, the catalytic site of TMPRSS2 is analyzed with a local drug-binding model. TMPRSS2 is a serine protease that has been shown to be a promising target for antivirals to prevent SARS-CoV-2 or influenza infections. Equilibrium populations of different drug-binding modes are derived for three inhibitors. We show that reactive populations of drug-enzyme complexes are a good proxy for experimentally determined drug efficiencies. Finally, we have extended our IMD approach an end-to-end deep learning framework called independent VAMPnets (or iVAMPnets). It learns a domain decomposition from simulation data and simultaneously models the kinetics in the local domains. iVAMPnets have been shown to succeed in both tasks for fully uncoupled benchmark systems and for the previously mentioned MD data set of synaptotagmin-1 C2A. We finally classify IMD and iVAMPnets as Markov field models (MFM), which we define as a class of models that describe dynamics by decomposing systems into local domains. Most MFMs account for couplings, thus generally being higher-order approximations compared to IMD.

In summary, we present a local approach to Markov modeling, from an abstract decomposition of the underlying transfer operator through semi-automated domain-decompositions and local Markov modeling to a deep learning framework. This work is highly focused on the applicability to high-dimensional MD data, thus aiming to pave a path for future quantitative kinetic modeling of large biomolecular complexes.

The following introduction has the goal to provide an introduction to the field of *in silico* kinetic modeling of molecular machines. In Sec. 1.1, we start by reviewing basic statistical mechanics approaches as an abstract layer connecting methods used throughout this thesis. Specifically, we study equations governing the dynamics of thermodynamic systems. Sec. 1.2 reviews the methods to generate such dynamics by means of computer simulations: We introduce the framework of atomistic molecular dynamics simulations and relate them to problems of molecular biology. In Sec. 1.3, we consider Markov models for describing molecular dynamics simulations. Taking up on the ideas of statistical physics presented earlier, we lay out estimation procedures and discuss how Markov models describe molecular kinetics. In Sec. 1.4, we study kinetic modeling in the specific context of biomolecular machines. Finally, we lead over to the publications by reviewing conceptual problems with existing approaches and pointing out a possible remedy in Sec. 1.5, motivating the results presented in the remainder of this thesis.

1.1 On the time evolution of thermodynamic systems

This work uses concepts from classical statistical physics to describe the time-dependent properties of thermodynamic systems. The ideas that are most relevant to the kinetic modeling of biomolecular machines are reiterated here to provide an abstract skeleton for the remainder of this thesis, i.e., to contextualize the used methods and the contributions of this work.

1.1.1 The Liouville equation of statistical mechanics

As the systems studied in this thesis are thermodynamic many-particle systems and treating all degrees of freedom explicitly is not desirable, we apply a mathematical formalism that is based on densities (in contrast to a particle-based picture). To that end, it is instructive to derive such a formalism for systems obeying the Hamilton equations of classical mechanics, a description known as the Liouville equation. The derivation presented below is based on Ref. [2].

Let Γ be a phase space of a Hamiltonian N-particle system equipped with a Hamiltonian that evaluates to the total energy of the system, $\mathcal{H} = K(\mathbf{p}) + U(\mathbf{q})$. We denote the vectors encoding the positions and momenta of all particles by $\mathbf{q}$ and $\mathbf{p}$, respectively, and assume that the Hamiltonian is separable into a potential $U(\mathbf{q})$ and a kinetic energy term $K(\mathbf{p})$. Instead of focusing on a single system, we now define $\rho(\mathbf{q}, \mathbf{p}, t)$ as the density of an ensemble of systems at point $(\mathbf{q}, \mathbf{p}) \in \Gamma$ and time t, with the normalization condition

$$\int_\Gamma \mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p}\rho(\mathbf{q}, \mathbf{p}, t) = 1. \tag{1.1}$$

This conservation of density means that there is a continuity equation that balances changes of ρ by a flux into or out of adjacent regions of phase space [3],

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial q_i}\left(\frac{\partial q_i}{\partial t}\rho\right) + \frac{\partial}{\partial p_i}\left(\frac{\partial p_i}{\partial t}\rho\right) = 0. \tag{1.2}$$

Now, we use the Hamilton equations $\frac{\partial q_i}{\partial t} = \frac{\partial \mathcal{H}}{\partial p_i}$ and $\frac{\partial p_i}{\partial t} = -\frac{\partial \mathcal{H}}{\partial q_i}$. As the second derivatives cancel out due to $\frac{\partial^2 \mathcal{H}}{\partial q_i \partial p_i} = \frac{\partial^2 \mathcal{H}}{\partial p_i \partial q_i}$, we can write the time derivative of the density function as

$$\frac{\partial \rho}{\partial t} = \left[\frac{\partial \mathcal{H}}{\partial q_i}\frac{\partial}{\partial p_i} - \frac{\partial \mathcal{H}}{\partial p_i}\frac{\partial}{\partial q_i}\right] \circ \rho. \tag{1.3}$$

This equation is known as the Liouville equation; it can be simplified by introducing the Liouville operator* $\mathcal{L} = -i\left[\frac{\partial \mathcal{H}}{\partial q_i}\frac{\partial}{\partial p_i} - \frac{\partial \mathcal{H}}{\partial p_i}\frac{\partial}{\partial q_i}\right]$:

$$\frac{\partial \rho(\mathbf{q}, \mathbf{p}, t)}{\partial t} = i\mathcal{L} \circ \rho(\mathbf{q}, \mathbf{p}, t). \tag{1.4}$$

*Following Ref. [4], we define the Liouville operator with imaginary unit i.

The Liouville equation describes the time evolution of phase-space densities in any Hamiltonian system. As the Liouville operator generates this dynamics, it is often called the *generator*. Assuming ergodicity (cf. Sec. 1.1.5), it can be applied to both a single system trajectory or an ensemble of trajectories, and, in particular, can be interpreted as the equation governing any kind of classical multi-component systems. It therefore plays a crucial role in statistical physics and is fundamental to the contemporary results presented in this work. Atomistic molecular dynamics simulations are an example of particular interest as they usually model intrinsically high-dimensional complex systems (cf. Sec. 1.2).

Given an initial state $(\mathbf{q}_0, \mathbf{p}_0) \coloneqq (\mathbf{q}(t{=}0), \mathbf{p}(t{=}0))$, the formal solution [2] of the Liouville equation is given by

$$\rho(\mathbf{q}, \mathbf{p}, t) = \exp(it\mathcal{L}) \circ \rho(\mathbf{q}_0, \mathbf{p}_0), \tag{1.5}$$

i.e., the density at its initial time $\rho(\mathbf{q}_0, \mathbf{p}_0)$ is propagated by the exponential of the Liouville operator multiplied by a finite time step t, to its target density $\rho(\mathbf{q}, \mathbf{p}, t)$ at time t.

In the remainder of this work, the exponential form of the Liouville operator will be replaced by the Perron–Frobenius (or PF) operator [4],

$$\mathcal{P}(\tau) = \exp(i\tau\mathcal{L}). \tag{1.6}$$

For the sake of simplicity, we describe a system's state as a function of time (omitting $\mathbf{q}$ and $\mathbf{p}$) and additionally use a Greek letter τ instead of the formally used Latin t to denote that this argument is a fixed, finite time step intrinsic to the operator. Therefore, Eq. (1.5) becomes

$$\rho(t_0 + \tau) = \mathcal{P}(\tau) \circ \rho(t_0). \tag{1.7}$$

This equation describes the *transfer* of density from time t_0 to $t_0 + \tau$. Therefore, the PF operator $\mathcal{P}(\tau)$ is often called a *transfer operator*. Eq. (1.7) can be regarded fundamental for the modeling of kinetics of a system. The remainder of this thesis considers approximating the PF operator in various forms, for various application cases, e.g., by Markov state models (cf. Sec. 1.1.5). We note that we will omit some technical details regarding the operator here and refer to Ref. [5] for details.

1.1.2 Equilibrium

In equilibrium [6, 7], the density of the states populated by an ensemble of systems does not change with time. It is assumed that any physical system will converge to its equilibrium state, and that the equilibrium state is uniquely defined. Please note that in the remainder of this thesis, we will work with normalized and non-negative *probability* densities.

Given any initial condition $(\mathbf{q}_0, \mathbf{p}_0)$, we define equilibrium as the state that is reached after infinite time has passed. The density of that state, i.e., the equilibrium density, is

denoted by the Greek letter μ,

$$\lim_{t\to\infty} \rho(\mathbf{q}, \mathbf{p}, t|\mathbf{q}_0, \mathbf{p}_0) = \mu. \tag{1.8}$$

In particular, if our starting state is already the equilibrium distribution, it follows that it will not be altered by the PF operator $\mathcal{P}$. In other words, the equilibrium density is an eigenfunction of the PF operator with an eigenvalue of 1,

$$\mathcal{P} \circ \mu = \mu. \tag{1.9}$$

In the scenarios investigated in this thesis, it is assumed that velocities decorrelate quicker than the processes of interest [8]. This means that the investigated lag times τ of the propagator are so large that velocities have no effect and are thus omitted as explicit functional dependencies. See Ref. [5] for a thorough derivation.

The density μ describes the equilibrium state of a thermodynamic system. From the viewpoint of the Liouville equation (Eq. (1.4)), such a state is represented by a density that does not explicitly depend on time, i.e., $\frac{\partial \rho}{\partial t} = 0$. We therefore relate the equilibrium distribution to the energy through the Boltzmann distribution

$$\mu(\mathbf{q}) \sim \exp\left(-\frac{\mathcal{H}(U(\mathbf{q}))}{k_B T}\right), \tag{1.10}$$

which fulfills this requirement. Here, $\mathcal{H}(U(\mathbf{q}))$ is the Hamiltonian, i.e., the total energy of the system in configuration $\mathbf{q}$ [6]. We note that the vanishing partial time derivative is not sufficient to identify the Boltzmann distribution as the unique functional form of a distribution in the equilibrium state – a thorough derivation can be found, e.g., in Ref. [3].

1.1.3 Brownian motion and the Langevin equation

The Liouville equation (Eq. (1.4)) and its solution (Eq. (1.5)) do not model the stochasticity that arises from excluding certain degrees of freedom (e.g., those of a heat bath), a fundamental property of thermodynamic systems. As the latter and their time evolution are the main focus of this thesis, we present a formalism that describes the dynamics of systems coupled to a heat bath without explicitly modeling the heat bath. This formalism is incorporated in the Langevin equation [9–11] and will be motivated and derived following Ref. [2] below.

The canonical example for the Langevin equation is Brownian motion, which describes the random motion of particles immersed in a fluid at room temperature. It is a stochastic process can be modeled by a spherical particle of radius r immersed in a fluid that is subject to frictional forces. These forces may be described by Stokes' law, i.e., $\mathbf{F}_{\text{friction}} = -\zeta\, \mathbf{v}$ with $\zeta = 6\pi\eta r$ and η the viscosity of the fluid. If that was the total force acting on such a particle, the velocity would decay to zero, countering our observation of the (ongoing) random motion of particles immersed in a fluid. We need to

take into account the equipartition theorem [12, p. 54f.] $\langle \mathbf{v}^2 \rangle = \frac{k_B T}{m}$ which states that in an equilibrium system that is coupled to a heat bath at finite temperature $T > 0$, the expectation value of the velocity $\langle \mathbf{v} \rangle$ must be non-zero. To account for that, a random force term $\delta\mathbf{F}(t)$ is added, and Newton's equation of motion becomes

$$m\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = -\zeta\mathbf{v} + \delta\mathbf{F}(t). \tag{1.11}$$

This equation is the Langevin equation for a Brownian particle. The first term describes the internal friction, i.e., a systematic contribution, and the second is a noise term accounting for random fluctuations within the fluid. We find that the random force has the properties

$$\langle \delta\mathbf{F}(t) \rangle = 0 \qquad \langle \delta\mathbf{F}(t)\delta\mathbf{F}(t') \rangle = 2\zeta k_B T \delta(t - t'), \tag{1.12}$$

i.e., it can be modeled by white noise with zero mean and variance $2\zeta k_B T$. We note the connection between the magnitudes of fluctuations and friction, which is a simple incarnation of the fluctuation-dissipation theorem [2].

Generalizing the Langevin equation (Eq. (1.11)) to the one of the used molecular dynamics engine [13] is trivially achieved by adding a force term $\mathbf{f}$ which represents the molecular dynamics force field. For each particle i, it is given* by

$$m_i\frac{\mathrm{d}\mathbf{v}_i}{\mathrm{d}t} = \mathbf{f}_i - \zeta\mathbf{v}_i + \delta\mathbf{F}(t)_i. \tag{1.13}$$

We are now equipped with a formal description of a multi-particle system in terms of the per-particle equations of motion (Eq. (1.13)). We note that from the Langevin equation, we can derive a Fokker-Planck equation that, without friction ($\zeta = 0$), reduces to the standard Liouville equation, Eq. (1.4) (see, e.g., Ref. [12, p. 428] or [2] for a derivation). We now have a consistent description of the time-evolution of thermodynamic systems. Even though this description is still very abstract, we can analyze properties of such systems and relate them to the more applied class of methods that are used in this thesis.

1.1.4 Metastability

An example of Brownian motion in a one-dimensional potential energy landscape is shown in Fig. 1.1. It has two basins with an energy barrier separating them. Dynamically, the two basins are connected by a stochastic process (Brownian dynamics, cf. Sec. 1.1.3) with long decorrelation time† as compared to the fast exchange processes that occur within each of the basins. Usually, we are interested in the *slow* exchange kinetics between the two basins rather than the fast fluctuations within an individual one. For example, in a biomolecular system, the slow (and relevant) kinetics could take place be-

*The equation used by the OpenMM [13] engine taken from the user handbook `http://docs.openmm.org/7.6.0/userguide/theory/04_integrators.html#langevinintegator`

†Corresponding to the implied timescale of a Markov state model that is introduced later (Sec. 1.3.1).

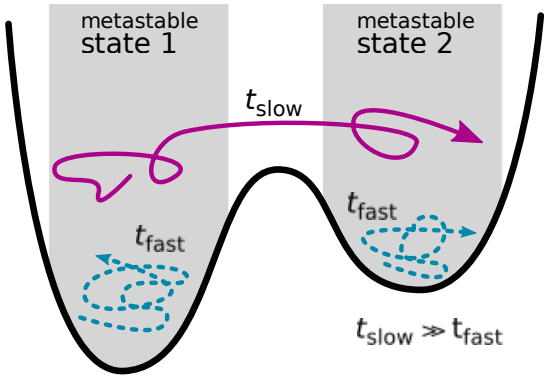


Figure 1.1: Metastability is defined by an exchange process that is very *slow* compared to other processes in a system.

tween active and nonactive protein states, whereas fast fluctuations around these states may not be relevant for a model.

This leads to metastability, an important concept for describing dynamical systems. Generalizing the above example (Fig. 1.1), transitions between different parts of an energy landscape can occur with various decorrelation times, which is particularly true for biomolecular systems. Therefore, we define a metastable state as a set of (molecular) configurations $\mathbf{q}$ that is in a quasi-equilibrium state, i.e., it rarely changes into another set of configurations. Within the metastable state, configurations exchange quickly [4]. Metastable states are often called *(molecular) conformations* [5]. Due to the connection between the eigenvalues of the PF operator and the relaxation timescales of corresponding processes, metastable dynamics can be modeled by eigenvalues λ_i that are only slightly smaller than the equilibrium eigenvalue $\lambda_1 = 1$ [5]. The metastable dynamics can therefore be seen in a spectral analysis of the PF operator and the slow ($\lambda_i \approx 1$) dynamics can be bisected from the fast fluctuations ($\lambda_i \approx 0$) in a quantitative way, resulting in a decomposition of the PF operator (cf. Eq. (1.7)) [6]

$$\rho(\tau) = \mathcal{P}_{\text{slow}}(\tau) \circ \rho_0 + \mathcal{P}_{\text{fast}}(\tau) \circ \rho_0. \tag{1.14}$$

The second part, i.e., the fast dynamics, generally describes processes that are of no interest to molecular dynamics modeling [6]. An eigendecomposition of the underlying operator can therefore be used to extract the slow, relevant processes (discarding the fast ones). We finally note that the concept of metastability can be generalized to coherent sets for non-reversible dynamics [14, 15].

1.1.5 Time averages and approximations to the Perron–Frobenius operator

The ergodicity hypothesis states that in the thermodynamic limit, the time average $\bar{o}$ of any observable o is the same as the ensemble average $\langle o \rangle$, i.e., the average over multiple

copies of the same system [16]. It assumes that a trajectory of a single system visits all states on its energy surface, yielding an equilibrium sample. In the NVE ensemble (i.e., conserved number of particles N, volume V, and total energy E), it can be written [11, p. 96 ff]

$$\langle o \rangle = \frac{\int \mathrm{d}\mathbf{q}\, o(\mathbf{q})\delta(\mathcal{H}(\mathbf{q}) - E)}{\int \mathrm{d}\mathbf{q}\, \delta(\mathcal{H}(\mathbf{q}) - E)} = \lim_{T\to\infty} \frac{1}{T} \int_0^T \mathrm{d}t\, o(\mathbf{q}(t)) = \bar{o}. \tag{1.15}$$

The ergodicity hypothesis is usually assumed to hold in the thermodynamic limit with infinite observation time, which corresponds to infinitely long molecular dynamics trajectories or infinitely many instances in the observation ensemble. However, with regard to molecular dynamics, real simulation time is finite, and very often transitions between different metastable regions are rare events, thus hampering attempts to fully and adequately sample the energy surface [11, p. 96 ff]. Consequently, most molecular simulations are too short to represent a sample of the global equilibrium state, which is what most experiments measure. Therefore, computing simple time averages is often not sufficient to obtain equilibrium properties of an *in silico* system, impeding comparability to experiments and reproducibility. This issue is broadly known as the sampling problem of molecular dynamics (cf, e.g., Refs. [17, 18]). It states that trajectory lengths necessary to appropriately approximate the right-hand side of Eq. (1.15) are often prohibitively large for any biologically interesting protein system (cf. Sec. 1.2.5).

However, we can obtain equilibrium properties by extracting them from a learned approximation to the PF operator, as estimating the latter only requires sampling the *local* equilibrium state of the system [*]. As the PF operator encodes stationary or equilibrium probabilities μ as an eigenmode, we find

$$\langle o \rangle = \int_\Gamma o(\mathbf{q})\mu(\mathrm{d}\mathbf{q}). \tag{1.16}$$

Importantly, ergodicity results in a unique stationary distribution [19, p. 93, 4] and is therefore fundamental for the analysis conducted with Markov state models as presented in Sec. 1.3.

In the remainder of this thesis, averages will be computed from approximations to PF operators. For this task, we use Markov state models (MSMs). MSMs are an approximation to a PF operator in a discrete state space [†], or more precisely, a Galerkin discretization [4]. Specifically, MSMs use indicator basis functions $\chi_i(\mathbf{q})$ that assign a discrete state i to a continuous (molecular) configuration $\mathbf{q}$ if the configuration is in a set S_i ,

$$\chi_i(\mathbf{q}) = \begin{cases} 1 & \mathbf{q} \in S_i \\ 0 & \text{else.} \end{cases} \tag{1.17}$$

[*] E.g., a Markov state model (cf. Sec. 1.3) assumes that the simulation ensemble is in equilibrium within the discrete Markov states [6], i.e., that the stochastic processes within a Markov state are very fast compared to the dynamics between different ones.

[†] MSMs approximate a reweighed spatial PF operator. For details, cf. Refs. [5, 6].

This is usually realized by lumping together molecular configurations into defined discrete states by splitting the state space by a Voronoi tesselation [6]. The Galerkin discretization yields a transition matrix that is defined given the discrete basis functions,

$$(\mathbf{P}(\tau))_{ij} = \frac{\langle \chi_j, \mathcal{P}(\tau) \circ \chi_i \rangle_\mu}{\langle \chi_i, \chi_j \rangle_\mu} \tag{1.18}$$

where $\langle \cdot, \cdot \rangle_\mu$ defines a μ-reweighed scalar product [6]. The transition matrix is a row-stochastic matrix with

$$(\mathbf{P}(\tau))_{ij} > 0 \; \forall \; i,j \qquad \text{and} \qquad \sum_i (\mathbf{P}(\tau))_{ij} = 1 \; \forall \, j. \tag{1.19}$$

In comparison to the continuous case, probability densities become vectors that encode probabilities per discrete state, i.e., the MSM transition matrix propagates probability vectors in time. Its eigenvector corresponding to eigenvalue $\lambda_1 = 1$ encodes the stationary or equilibrium state, now encoded as a vector of (equilibrium) probabilities denoted by $\boldsymbol{\pi}$. Computing averages simplifies to

$$\langle o \rangle = \sum_i^N o_i \pi_i \tag{1.20}$$

with the additional assumption that the observable o can be modeled as constant within a given discrete state i. Sec. 1.3.1 provides a more comprehensive overview of discrete state MSMs.

1.2 Atomistic molecular dynamics simulations: Generators of many-body system dynamics

Molecular dynamics (MD) simulations have been developed to gain physical insights into molecular systems on the atomistic scale. They compute the time evolution of a given molecular system by using empirical interaction potentials between its atoms. In particular, the advent of fast and affordable computing resources has enabled the method to take a remarkable leap over the last 45 years, from a 9 ps trajectory of bovine pancreatic trypsin inhibitor (58 amino acid residues) in vacuum in 1977 [20] to 0.1 seconds of the SARS-CoV-2 spike protein ($\sim$ 1200 residues) in explicit solvent in 2021 [21].

The generator of the dynamics is usually a numerical integrator that evaluates Newton's equation of motion (akin to Hamiltonian dynamics cited above) in a high-dimensional potential landscape [22]. Regardless of its complexity, the generator can be interpreted in the form of the Liouville operator in Eq. (1.5).

MD simulations yield trajectories of particle coordinates, i.e., Cartesian coordinates of every atom over time. This makes MD simulations intrinsically high-dimensional, posing a challenge to quantify system properties and to describe the dynamics in a human-readable fashion. In this work, much of these issues will be addressed using

Markov state models (cf. Sec. 1.3). This section summarizes the basics of MD simulations. For a more comprehensive overview, see, e.g., Refs. [16, 22].

1.2.1 Atomistic force fields

MD simulations define a potential function that describes particle interactions within a system [22]. This potential is usually defined for up to 4-body interaction terms. In a classical MD framework, the potential terms are split into bonded and non-bonded interactions. Non-bonded forces describe repulsion and van-der-Waals interactions (usually via Lennard-Jones potentials) and electrostatics (via a Coulomb term). Bonded interactions are composed of distances, angles, and dihedral angles between pairs, triplets, and quadruplets of atoms, respectively. Many of these terms are governed by harmonic potential terms, i.e., they are approximated by a harmonic oscillator fluctuating around a mean with some force constant. The set of functional forms and parameters is called a force field. In order to achieve biologically relevant simulation timescales, MD force fields approximate quantum-mechanical interactions using classical expressions (e.g., harmonic oscillators of atoms). This means that chemical reactions and other electronic effects cannot be described unless they are parameterized specifically [22–24].

A force field that has been used in Chapter 4 of this work is CHARMM 36 [25]. Its functional form is a sum over various energy terms including bonded and non-bonded interactions. E.g., atomic bond terms are covered by a harmonic potential of the form $\sum_{\text{bonds}} K_{\text{b}}(b - b_{\text{o}})^2$ with the force constant K_{b} that determines the strength of the interaction, the equilibrium value of the bond-length b_{o} , and b its time-dependent value. Please compare Ref. [26] for the full energy expression. The parameters of the force field are obtained by fitting them to *ab initio* quantum mechanical calculations or experimental observables such as NMR J-couplings [25, 26]. They are reported for different atom types and can automatically be assigned to a given protein structure, a task that is usually done by softwares such as OpenMM modeler [13] or GROMACS [27].

1.2.2 Integration schemes

Having defined a potential function enables us to determine particle positions and velocities over time according to the laws of classical mechanics. To generate the time series of a molecule evolving over time, the above defined potential (or the force field) is evaluated for some initial particle positions (and velocities), yielding the forces acting on each particle of the molecule. Subsequently, particles are *moved* according to the forces acting on them. This procedure is repeated to evolve the particle coordinates in time, using the last position (and velocities) as the new initial positions (and velocities). As this can be conducted by several integration schemes, we start by illustrating the general idea using a very simple integration scheme (Euler integrator) [16]: Let $\mathbf{q}(t) \in \mathbb{R}^n$ be the trajectory of a particle with mass m and $U(\mathbf{q})$ be its position-dependent potential.

We can expand Newton's equation of motion,

$$\nabla U(\mathbf{q}(t)) = -m\frac{\mathrm{d}^2\mathbf{q}(t)}{\mathrm{d}t^2} \tag{1.21}$$

by using a polynomial ansatz function $\mathbf{q}(t) \in \mathbb{P}^n$ and write $\mathbf{q}(t)$ as a Taylor series. Using the definition of the velocity $\mathbf{v} = \frac{\mathrm{d}\mathbf{q}(t)}{\mathrm{d}t}$ and Eq. (1.21), we find

$$\mathbf{q}(t_0 + \Delta t) = \mathbf{q}(t_0) + \mathbf{v}(t_0)\Delta t - \frac{1}{2m}\nabla U(\mathbf{q})\Delta t^2. \tag{1.22}$$

This equation can be evaluated for small Δt with a known force field potential $U(\mathbf{q})$, however it comes with severe limitations and serves the purpose of demonstrating the concept only. Most notably, the Euler integrator is not symplectic, i.e., it is not time-reversible or energy conserving.

A commonly used symplectic integrator is the Velocity-Verlet algorithm, which can be derived directly from the Liouville equation (Eq. (1.4)) by means of a Trotter expansion [16, p. 77ff.]. It comes at almost the same computational cost as the Euler method but conserves a so called shadow Hamiltonian [11, p. 120], i.e., the total energy oscillates around the true value and does not diverge. The Velocity-Verlet integrator therefore represents the microcanonical (or NVE) ensemble, which in many cases does not mirror experimental conditions, as the total energy is rarely controlled (see next Section 1.2.3).

1.2.3 Modeling the experimentally observed thermodynamic ensemble

So far, the presented MD methods describe classical (deterministic) mechanics at constant energy E. However, conditions for biomolecular experiments pose different requirements: Instead of working with single molecules, there is usually a large number N of molecules in a sample, e.g., a protein solution in a test tube. Furthermore, systems are usually coupled to a heat bath, i.e., have constant temperature T rather than constant (internal) energy E, and are conducted at atmospheric pressure P. Therefore, it is crucial that the computational models of such experiments reproduce these conditions, by controlling the variables N, P, and T. In statistical mechanics, this situation is known as the isothermal-isobaric or NPT ensemble [11, p. 236]. In the following, we will discuss how these conditions can be met in MD simulations.

Number of particles First, the number of particles N is kept constant trivially by not adding or removing particles to a simulation.

Temperature Second, temperature T is not directly translatable to a single particle inside a simulation box as it is defined for a thermodynamic system. In order to control the temperature, it is important to note that by the equipartition theorem, the average

kinetic energy $\langle K \rangle$ per degree of freedom scales with the temperature [16, p. 64]

$$\langle K \rangle = \frac{1}{2} k_B T. \tag{1.23}$$

Temperature can therefore be controlled by adjusting the kinetic energy, or more precisely, the particle velocities. This can algorithmically be realized by computational *thermostats*. It is important to note that velocities are distributed according to the temperature-dependent Maxwell-Boltzmann distribution [11]

$$p(|\mathbf{v}|) = \left(\frac{m}{2\pi k_B T} \right)^{3/2} \exp\left(-\frac{m|\mathbf{v}|^2}{2k_B T} \right). \tag{1.24}$$

A way of controlling temperature is to simulate a heat bath by randomly selecting particles and applying a stochastic force to them such that the resulting velocity distribution follows the Maxwell-Boltzmann distribution [16, p. 141 f.]. This method is known as the Andersen thermostat.

In the present work, temperature control is achieved by integrating the Langevin equation rather than Newton's equation of motion, directly describing the dynamics of a system with an implicit coupling to a heat bath [11] (cf. Sec. 1.1.3). Thus, no additional thermostat is necessary. The Langevin equation (Eq. (1.13)) is numerically integrated using a random force that is modeled with two Gaussian random variables. They are sampled at each integrator step and have zero mean, unit variance, and no cross correlation [11, p. 591], modeling the Langevin random force described earlier (Sec 1.1.3).

Pressure Third, the pressure P can be controlled by introducing a pressure coupling. In this work, the pressure is controlled using a Monte Carlo procedure* [28, 29] by applying variations $\Delta V = A \cdot r$ to the volume of the simulation box, with $r \sim \mathcal{U}(-1, 1)$ a uniformly distributed random number and A a scaling factor.

We are now equipped with the basic concepts to model dynamics of an N-particle system in the *NPT* ensemble.

1.2.4 Short overview of MD-related methods

There are numerous methods that are used to speed up MD simulations or to make them feasible in general. Most of them are optimized such that they do not alter the underlying physics. Therefore, we assume that these methods do not affect the results presented in this thesis and only give a very brief overview.

Periodic boundary conditions Usually, an MD simulation is conducted in a periodic box to mimic bulk behavior and to rule out surface effects [22].

*The presented Monte Carlo barostat is the one implemented in OpenMM [13], cf. user manual `http://docs.openmm.org/7.6.0/userguide/theory/02_standard_forces.html`

Neighbor lists In order to avoid computing interactions between all pairs of particles in a simulation, MD algorithms often work with neighbor lists that are updated in a low frequency fashion [30]. Short-ranged non-bonded interaction terms are only computed between neighbors in that list. Furthermore, non-bonded interaction terms from Lennard-Jones potentials are usually truncated [22].

Particle mesh Ewald (PME) summation Coulomb interactions are long-range non-bonded interaction terms that make MD simulations expensive when evaluated directly. PME splits them into a long-range and a short-range part with the long-range part being solved using a Fast Fourier Transform, reducing scaling of the computing load from N^2 to $N\log(N)$ [31].

Constraints Often, fast vibrating bonds are constrained to a fixed length in order to allow for a larger integration time step. This is usually done using the SHAKE [32] or RATTLE algorithm [33].

Hydrogen Mass Repartitioning (HMR) To artificially slow down the fastest degrees of freedom in a system, hydrogen masses can be increased while simultaneously decreasing the masses of the heavy atoms that they are bound to. Usually, a mass of 4 u is assigned to hydrogens [34]. This method allows to increase the integrator time step to up to 5 fs and does not significantly alter the resulting trajectories.

While the listed methods are among the most important ones, there are many more optimization schemes available that make MD feasible. This highlights the need of specialized software packages (such as OpenMM [13], GROMACS [27] or NAMD [35]), that have been highly optimized over many years.

1.2.5 Biological timescales and the sampling problem

One of the fundamental problems of MD simulations is the timescale gap to biological systems. Functionally relevant processes in molecular machines are intrinsically multiscale, ranging from nanoseconds (e.g., side-chain rotamers or local loop motions) to minutes (larger domain rearrangements, protein-protein association/dissociation) [36, 37], i.e., about eight orders of magnitude. However, the fastest process that is still relevant for all atom MD simulations is the vibration of bonds that contain hydrogen atoms. With a period in the femtosecond range, it determines the integration step of an MD integrator to ≤ 1 fs. This limit can be pushed up to 5 fs, e.g., by methods such as Hydrogen Mass Repartitioning [4, 34, 38]. However, we stay up to 10 orders of magnitude away from relevant biological timescales, rendering biological timescale simulations very expensive and in many cases even intractable [39, 40].

To gather significant statistics of a biological process in an MD simulation, i.e., to compute quantitative results such as expectation values or averages, the slowest processes must be observed in significant numbers [40]. To that end, a single simulation

would need to sample the global equilibrium of the system–otherwise, time averages do not represent meaningful quantities. Given that MD simulations are finite in reality, this is often hard if not impossible to achieve. However, according to the ergodicity hypothesis (Sec. 1.1.5), there is an alternative to computing time averages from a single, infinitely long trajectory. We can use an ensemble of trajectories instead that cover the slowest process in multiple instances (Sec. 1.1.5). In particular the development of modern computational hardware and graphical processing units (GPU) cards has opened the possibility to run MD simulations in a highly parallelized, multi-trajectory fashion, shifting the paradigm towards ensembles of trajectories. To deal with the resulting large number of MD trajectories, methods are needed that cast this data into quantitative, human-readable models. A popular choice in that regard are Markov state models [41].

1.3 Markov state models: Low rank approximations to the generator*

In brief, Markov state models (MSMs) of molecular dynamics are an approximation to the PF operator (Eq. (1.5)) [43] that is physically interpretable and yields a human-readable quantitative model of the dynamics. MSMs can be estimated from MD data and are capable of combining many short trajectories into a single model, making them a useful tool to cope with the MD sampling problem (Sec. 1.2.5). The MSM transition matrix describes (conditional) probabilities to jump from one state to another in a defined time, e.g., it could encode transition probabilities between different protein configurations.

Spectral analysis of the MSM transition matrix yields a rich toolbox for describing dynamic systems. In particular, the concepts of equilibrium distributions and metastability (Secs. 1.1.2 and 1.1.4) are contained in the eigenvalue spectrum of that matrix [6]. Revisiting Eq. (1.9), we see that the equilibrium distribution of the system is an eigenvector of the transition matrix with unity eigenvalue. It can thus be extracted from that matrix using standard eigendecomposition methods. Furthermore, we note that the system relaxes into equilibrium along the other eigenvectors, with timescales that can be computed from the corresponding eigenvalues λ_i with $t_i = \frac{\tau}{\ln(|\lambda_i|)}$. This gives a quantitative handle on metastability via the slow processes of the system.

*The content of this section (Sec. 1.3 including subsections) was published as
Christoph Wehmeyer*, Martin K. Scherer*, Tim Hempel*, Brooke E. Husic, Simon Olsson, and Frank Noé. "Introduction to Markov State Modeling with the PyEMMA Software [Article v1.0]". *Living Journal of Computational Molecular Science* 1.1 (2019), 5965. (*contributed equally). The article can be obtained from the publisher under `https://doi.org/10.33011/livecoms.1.1.5965`.
TH was one of three lead authors in this project. He has substantially contributed to writing the manuscript. Parts of the manuscript were modified or extended to fit the purpose of this introduction, other parts were quoted verbatim. The publication is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit `https://creativecommons.org/licenses/by/4.0/`.

1.3.1 Markov state models

In its standard formulation, the estimation of an MSM involves decomposing the phase or configuration space occupied by a (stochastic) dynamical system into a set of disjoint, discrete states, and a transition matrix (cf. Sec. 1.1.5) $\mathbf{P}(\tau) = [p_{ij}(\tau)]$ denoting the conditional probability of finding the system in state j at time $t+\tau$ given that it was in state i at time t. Let us make two remarks to avoid common misconceptions:

1. Equilibrium: While most analysis techniques require simulation trajectories to be long enough to sample from the equilibrium distribution, this is not required for MSMs. Because MSMs use the *conditional* probability $p_{ij}(\tau)$, they are useful for the analysis of short simulation trajectories with arbitrary starting points—see Ref. [44] for a thorough discussion of this matter.

2. Markovianity: An MSM is a memoryless model. Early MSM papers have argued that accurate MSMs can be found if a few states with high energy barriers between them are resolved so as to achieve a Mori-Zwanzig projection with fast-decaying memory [8, 45, 46]. The modern view, however, is that MSMs can be highly accurate if the MSM states discretize the reaction coordinates of the slowest processes well [6]. This mainly requires that the system is characterized by only a few slow processes at lag time τ, which is true for cooperative systems such as most proteins, but not for highly frustrated systems such as glasses.

In order to create an MSM for a dynamical system, each data point in the time series is assigned to a state. Given an appropriate lag time, every pairwise transition at that lag time is counted and stored in a count matrix. Then, a row-stochastic transition matrix $\mathbf{P}(\tau)$ is estimated from the count matrix. It is defined for the specified lag time. For MD simulations in equilibrium, $\mathbf{P}(\tau)$ should obey detailed balance which is enforced by constraining the estimation of $\mathbf{P}(\tau)$ to the following equations:

$$\pi_i p_{ij} = \pi_j p_{ji}, \qquad i,j = 1,\ldots N, \tag{1.25}$$

where π_i is the stationary probability of state i, p_{ij} is the probability of transitioning to state j conditional on being in state i, and N is the total number of states. The constraints (Eq. (1.25)) are omitted if MD simulations are not conducted in equilibrium, e.g., for systems experiencing a pulling force or an external potential (see Ref. [47] for a recent review on nonequilibrium MSMs). For the remainder of this section we will simplify the matter by assuming the more common scenario of MD simulations without external forces, such that Eq. (1.25) is assumed to hold.

When estimating an MSM it is critical to choose a lag time, τ, which is long enough to ensure Markovian dynamics in our state space, but short enough to resolve the dynamics in which we are interested. Plotting the implied timescales (ITS) as a function of τ can be a helpful diagnostic when selecting the MSM lag time [45]. The ITS t_i approximates the decorrelation time of the i^{th} process and is computed from the eigenvalues λ_i

of the MSM transition matrix via

$$t_i = \frac{-\tau}{\ln|\lambda_i(\tau)|}. \tag{1.26}$$

When the ITS become approximately constant with the lag time, we say that our timescales have converged and choose the smallest lag time with the converged timescales in order to maximize the model's temporal resolution.

Once we have used the ITS to choose the lag time, we can further validate whether a given transition matrix $\mathbf{P}(\tau)$ is approximately Markovian using the Chapman-Kolmogorov (CK) test [6, 48]. The CK property for a Markovian matrix is

$$\mathbf{P}(k\tau) = \mathbf{P}^k(\tau), \tag{1.27}$$

where the left-hand side of the equation corresponds to an MSM estimated at lag time $k\tau$ and k is an integer larger than 1. The right-hand side of the equation is our estimated MSM transition matrix to the k^{th} power. By assessing how well the approximated transition matrix adheres to the CK property, we can validate the appropriateness of the Markovian assumption for the model (see Sec. IV.F in Ref. [6]).

Once validated, the transition matrix can be decomposed into eigenvectors and eigenvalues. The highest eigenvalue, $\lambda_1(\tau)$, is equal to 1. As we assume that the underlying dynamics are ergodic, $\lambda_1(\tau) = 1$ is a unique eigenvector (Sec. 1.1.5) and its corresponding left eigenvector represents the stationary distribution, $\boldsymbol{\pi}$ (cf. Eq. (1.9)):

$$\boldsymbol{\pi}^\top \mathbf{P}(\tau) = \boldsymbol{\pi}^\top. \tag{1.28}$$

For the systems of interest here, the subsequent eigenvalues $\lambda_{i>1}(\tau)$ are real valued with $|\lambda_{i>1}| < 1$ [5] and are related to the *characteristic* or *implied* timescales of dynamical processes within the system (Eq. (1.26)). The dynamical processes themselves (for $i > 1$) are encoded by the right eigenvectors $\boldsymbol{\psi}_i$,

$$\mathbf{P}(\tau)\boldsymbol{\psi}_i = \lambda_i(\tau)\boldsymbol{\psi}_i, \tag{1.29}$$

where the eigenvalue-eigenvector pairs are indexed in decreasing order according to the eigenvalues. The coefficients of the eigenvectors represent the flux into and out of the Markov states that characterize the corresponding process. The right eigenvector $\boldsymbol{\psi}_1$ is a vector consisting of 1's.

1.3.2 Variational approach

The theory described in the previous section required the decomposition of the phase or configuration space occupied by a dynamical system into discrete, disjoint states. Starting from the output of an MD simulation of a protein, there are several steps that can be taken to obtain an MSM from the original configuration space, including featurization, dimension reduction, and clustering (cf. Sec. 1.3.4). These steps can be summarized as

finding appropriate basis functions or, more specifically, to optimize a set of indicator functions such that the dynamics in the discretized space is Markovian (cf. Sec. 1.1.5).

In 2013, the variational approach to conformational dynamics (VAC) was derived to quantify the process of optimizing basis functions for dynamical models in reversible setting [49]. It enables an objective comparison among different state decomposition choices for MSMs, or choices of basis functions in general.

Briefly, the VAC uses functions $\hat{l}_i$ to approximate the true left eigenfunctions of the PF operator (Eq. (1.7)). The functions $\hat{l}_i$ are therefore called approximated eigenfunctions. Under the condition that $\hat{l}_i$ are normalized and orthogonal to the true first eigenfunction μ, the autocorrelation function acf of the weighted approximated eigenfunctions have an upper limit that is given by the true eigenvalues [49]:

$$\mathrm{acf}(\mu^{-1}\hat{l}_i, \tau) \leq \lambda_i(\tau) \qquad i \geq 2. \tag{1.30}$$

As autocorrelation functions can directly be obtained from simulation data, the VAC can be used for the estimation of a Markov operator from data. Specifically, as the MSM eigenvalues are bounded from above by the true ones, maximizing the transition matrix eigenvalues represents a way of obtaining a VAC-optimal MSM [49].

However, the VAC is limited to reversible systems. A more general formalism was derived later that is the variational approach to Markov processes (VAMP) [50]. It uses the Koopman operator [51, 52] instead of the PF operator, which – in brief – is a Markov operator that propagates *observables* rather than densities, and is adjoint to the PF operator [53]. For the sake of space, we here only mention that in the special case of stationary, time-reversible dynamics, the singular value decomposition (SVD) of the Koopman operator is equivalent to the eigendecomposition of the transition matrix [50] (Sec. 1.3). Please find more details about the Koopman operator in Refs. [53, 54].

Following Ref. [50], the VAMP states that we can quantify (or *score*) the approximation quality of test functions $\mathbf{f}$ and $\mathbf{g}$ to represent the underlying stochastic processes. We make the ansatz of writing the test functions as linear combinations of basis functions χ_t and $\chi_{t+\tau}$, i.e., $\mathbf{f}(\mathbf{q}_t) = \mathbf{U}^\top \chi_t(\mathbf{q}_t)$ and $\mathbf{g}(\mathbf{q}_{t+\tau}) = \mathbf{V}^\top \chi_{t+\tau}(\mathbf{q}_{t+\tau})$. We can now optimize the basis functions without explicitly treating the linear expansion coefficients $\mathbf{U}$ and $\mathbf{V}$ using the so-called VAMP-r score $\mathcal{R}_r$ (r is a positive integer). To that end, we compute the covariance matrices of the basis functions,

$$\begin{aligned}
\mathbf{C}_{00} &= \mathbb{E}_T\left[\chi_t(\mathbf{q}_t)\chi_t(\mathbf{q}_t)^\top\right] \\
\mathbf{C}_{0\tau} &= \mathbb{E}_T\left[\chi_t(\mathbf{q}_t)\chi_{t+\tau}(\mathbf{q}_{t+\tau})^\top\right] \\
\mathbf{C}_{\tau\tau} &= \mathbb{E}_T\left[\chi_{t+\tau}(\mathbf{q}_{t+\tau})\chi_{t+\tau}(\mathbf{q}_{t+\tau})^\top\right].
\end{aligned} \tag{1.31}$$

Here, $\mathbb{E}_T[\cdot]$ denotes a simple time average. We can now compute a Koopman operator $\bar{\mathbf{K}}$ from the covariance matrices,

$$\bar{\mathbf{K}} = \mathbf{C}_{00}^{-1/2}\mathbf{C}_{0\tau}\mathbf{C}_{\tau\tau}^{-1/2}, \tag{1.32}$$

which represents an approximation to the true Koopman operator given the basis functions χ_t and $\chi_{t+\tau}$. The VAMP states that the approximation quality of that operator $\bar{\mathbf{K}}$ can be controlled via the VAMP-r score,

$$\mathcal{R}_r = \left\|\bar{\mathbf{K}}\right\|_r^r \tag{1.33}$$

with $\|\cdot\|_r$ denoting the r-Schatten norm (or the sum over the r-th power of singular values) [50]. Maximizing this score with respect to the basis functions χ_t and $\chi_{t+\tau}$ gives an optimal approximation to the true eigenfunctions of the Koopman operator in the test functions $\mathbf{f}$ and $\mathbf{g}$. Their linear expansion coefficients are encoded in the SVD of the approximated Koopman operator $\bar{\mathbf{K}}$,

$$\bar{\mathbf{K}} \approx \mathbf{U}\mathbf{K}\mathbf{V}^\top, \tag{1.34}$$

where $\mathbf{K}$ is a diagonal matrix with the singular values of the approximated Koopman operator. Please note that $\mathcal{R}_r$ is maximized if $\mathbf{f}$ and $\mathbf{g}$ are the eigenfunctions of the true Koopman operator.

1.3.3 Hidden Markov state models

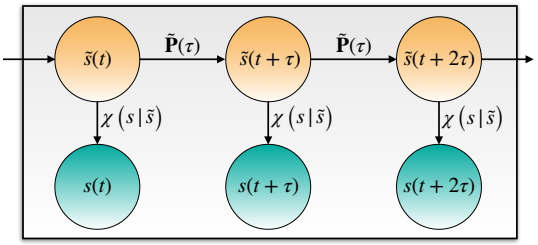


Figure 1.2: The HMM transition matrix $\tilde{\mathbf{P}}(\tau)$ propagates the hidden state trajectory $\tilde{s}(t)$ (orange circles) and, at each time step t, the emission into the observable state $s(t)$ (cyan circles) is governed by the emission probabilities $\chi(s(t)|\tilde{s}(t))$. Reprinted from [55].

The estimation of an MSM requires the dynamics between microstates to be Markovian. However, in case of a poor dimension reduction, discretization, or short trajectories, we cannot anticipate this to be the case and the Markovianity assumption is often violated [6].

An alternative, which is much less sensitive to poor discretization, is to estimate a hidden Markov model (HMM) [56–60]. HMMs are less sensitive to the discretization error as they sidestep the assumption of Markovian dynamics in the (observed) discretized space (illustrated in Fig. 1.2). Instead, HMMs assume that there is an underlying (hidden) dynamic process that is Markovian and gives rise to our observed data, i.e., the (n states) discretized trajectories $s(t)$. This is a powerful principle as we know that there is indeed an underlying process that is Markovian, which is generated by our molecular dynamics integrator.

To estimate an HMM, we need a spectral gap after the m^{th} timescale; in practice, a timescale separation of $t_m \geq 2t_{m+1}$ is sufficient [61]. The HMM then consists of a transition matrix $\tilde{\mathbf{P}}(\tau)$ between $m < n$ hidden states and a row-stochastic matrix (χ) of probabilities $\chi(s|\tilde{s})$ to emit the discrete state s conditional on being in the hidden state $\tilde{s}$.

An HMM estimation yields a model with a small number of (hidden) states. For the current application cases, these hidden states are optimized to describe metastable states of, e.g., a protein, and thus, the number of hidden states is a new hyper-parameter which needs to be chosen carefully. As the HMMs—like MSMs—approximate the full phase-space dynamics, we can similarly compute the metastable kinetics, apply TPT, visualize the network, and obtain physical observables. However, we note that HMMs tend to be harder to train to high-dimensional MD data and therefore, the decision of MSM vs. HMM needs to be adjusted to the specific problem at hand.

1.3.4 The Markov state modeling workflow

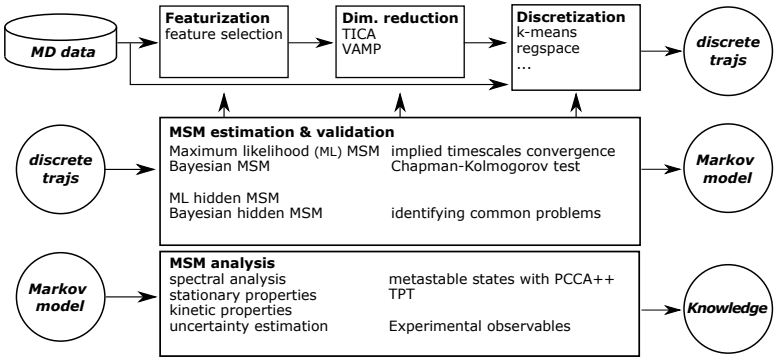


Figure 1.3: The MSM estimation workflow: MD trajectories are processed and discretized (first row). A Markov state model is estimated from the resulting discrete trajectories and validated (middle row). By iterating between data processing and MSM estimation/validation, a dynamical model is obtained that can be analyzed (last row). Reprinted and modified from [55].

In short, the workflow (Fig. 1.3) for a full analysis of an MD dataset might consist of

- extracting molecular features from the raw data,
- transforming those features into a suitable, low-dimensional subspace,
- discretizing the low-dimensional subsets into a state decomposition,
- estimating a maximum likelihood MSM from the discrete trajectories and performing validation tests,
- analyzing the stationary and kinetic properties of the MSM,
- finding metastable macrostates and applying transition path theory (TPT) to identify the pathways of conformational change,
- computing expectation values for experimental observables, and

- coarse-graining the MSM using a hidden Markov model approach.

For the remainder of this section we will walk through the example and analyze a dataset of the Trp-Leu-Ala-Leu-Leu pentapeptide (Fig. 1.4a), consisting of 25 independent MD trajectories conducted in implicit solvent with frames saved at an interval of 0.1 ns.

Note that the modeler has to select hyper-parameters at most stages throughout the workflow. This selection must be done carefully as poor choices make it hard, or even impossible, to build a good MSM. While there exist automated schemes [62] for cross-validated optimization in the full hyper-parameter space, we chose to adopt a sequential approach where only the hyper-parameters of the current stage are optimized. This approach is not only computationally cheaper but allows us to discuss the significance of the necessary modeling choices.

1.3.5 Feature selection

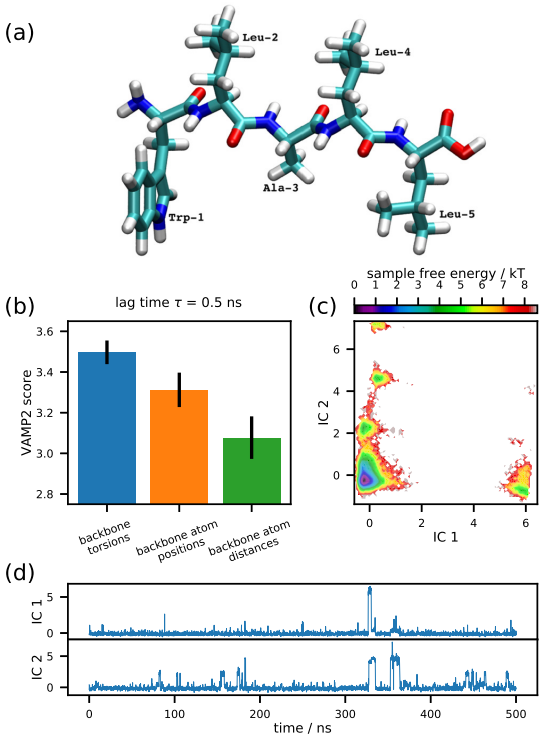


Figure 1.4: Example analysis of the conformational dynamics of a pentapeptide backbone: **(a)** The Trp-Leu-Ala-Leu-Leu pentapeptide in licorice representation [63]. **(b)** The VAMP-2 score indicates which of the tested featurizations contains the highest kinetic variance. **(c)** The sample free energy projected onto the first two time-lagged independent components (ICs) at lag time $\tau = 0.5\,\mathrm{ns}$ shows multiple minima and **(d)** the time series of the first two ICs of the first trajectory show rare transitions. Reprinted from [55].

In Markov state modeling, our objective is to model the slow dynamics of a molecular process. In order to approximate the slow dynamics in a statistically efficient manner, a lower-dimensional representation of our simulation data is necessary. However, the features (e.g., torsion angles, distances or contacts) which best represent the slow dynamical modes of a given molecular system are unknown *a priori* [64]. Fortunately, the VAC [49, 65] and the more general VAMP [50] provide a systematic means to quantitatively compare multiple representations of the simulation data (Sec. 1.3.2). In particular, we can use a scalar score obtained using VAMP to directly compare the ability of certain features to capture slow dynamical modes in a particular molecular system.

For the following analyses, we choose the VAMP-2 score (i.e., $r = 2$) as it maximizes the kinetic variance contained in the features [66], i.e., has a physical interpretation. We should always evaluate the score in a cross-validated manner to ensure that we include neither too few (under-fitting) nor too many (over-fitting) features [50, 67]. To choose among three different molecular features reflecting protein structure, we compute the (cross-validated) VAMP-2 score. Although we cannot optimize MSM lag times with a variational score [68], such as VAMP-2, it is important to ensure that the properties we optimize are robust as a function of lag time. Consequently, we compute the VAMP-2 score at several lag times. We find that for our pentapeptide system, the relative rankings of the different molecular features are highly robust as a function of lag time. We show one example of this ranking and the absolute VAMP-2 scores for lag time 0.5 ns in Fig. 1.4b. We find that backbone torsions contain more kinetic variance than the backbone heavy atom positions or the distances between them (Fig. 1.4b). This suggests that backbone torsions are the best of the options evaluated for MSM construction.

1.3.6 Dimensionality reduction

Subsequently, we perform TICA [66, 69, 70] in order to reduce the dimension from the feature space, which is typically very high-dimensional, to a lower-dimensional space that can be discretized with higher resolution and better statistical efficiency. TICA can be understood as a special case of the VAC [49, 65] and is designed to find a projection preserving the long-timescale dynamics in the dataset. Here, performing TICA on the backbone torsions at lag time 0.5 ns yields a four-dimensional subspace. The sample free energy projected onto the first two independent components (ICs) exhibits several minima (Fig. 1.4c). Discrete transitions between the minima can be observed by visualizing the transformation of the first trajectory into these ICs (Fig. 1.4d). We thus assume that our TICA-transformed backbone torsion features describe one or more metastable processes.

1.3.7 Discretization

TICA yields a representation of our molecular simulation data with a reduced dimensionality, which can greatly facilitate the decomposition of our system into the discrete Markovian states necessary for MSM estimation. Here, we use the k-means algorithm to segment the four-dimensional TICA space into $k = 75$ cluster centers. The number

of cluster centers has been chosen to optimize the VAMP-2 score in a manner identical to how the feature selection was carried out above.

1.3.8 MSM estimation and validation

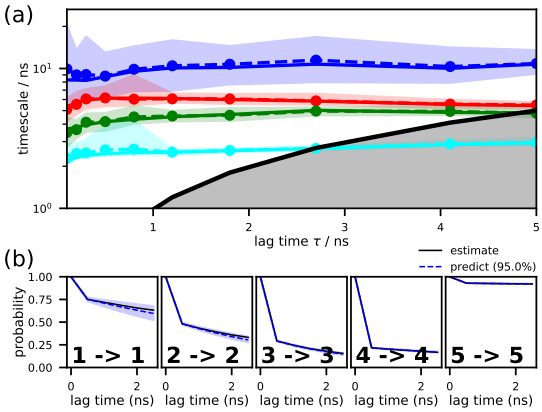


Figure 1.5: Example analysis of the conformational dynamics of a pentapeptide backbone: **(a)** The convergence behavior of the implied timescales associated with the four slowest processes. The solid lines refer to the maximum likelihood result while the dashed lines show the ensemble mean computed with a Bayesian sampling procedure [71]. The black line (marking equality of timescale and lag time) with grey area indicates the timescale horizon below which the MSM cannot resolve processes. As implied timescales are well-converged at $\tau = 0.5$ ns, this lag time is chosen for subsequent MSM estimation. **(b)** CK test computed using an MSM estimated with lag time $\tau = 0.5$ ns assuming 5 metastable states. Predictions from this model agree with higher lag time estimates within confidence intervals. Implied timescales convergence as well as a passing CK test are a necessary condition in MSM validation. In both panels, the (non-grey) shaded areas indicate 95 % confidence intervals computed with a Bayesian sampling procedure [71]. Reprinted from [55].

A necessary condition for Markovian dynamics in our reduced space is that the ITS are approximately constant as a function of τ; accordingly, we chose the smallest possible τ which fulfills this condition within the model uncertainty. The uncertainty bounds are computed using a Bayesian scheme [71, 72] with 100 samples. In our example, we find that the four slowest ITS converge quickly and are constant within a 95 % confidence interval for lag times above 0.5 ns (Fig. 1.5a).

To test the validity of our MSM, we perform a CK test. Visualizing full transition matrices $\mathbf{P}$ over a multitude of different lag-times is difficult; we therefore coarse-grain $\mathbf{P}$ into a smaller number of metastable states before performing the test. An appropriate number of metastable states can be chosen by identifying a relatively large gap in the ITS plot. For this analysis, we chose five metastable states. The CK test (Fig. 1.5b) shows that predictions from our MSM (blue-dashed lines) agrees well with MSMs estimated with longer lag times (black-solid lines). Thus, the CK test confirms that five metastable states is an appropriate choice and shows that the MSM we have estimated at lag time $\tau = 0.5$ ns indeed predicts the long-timescale behavior of our system within error (blue/shaded area).

1.3.9 MSM Analysis

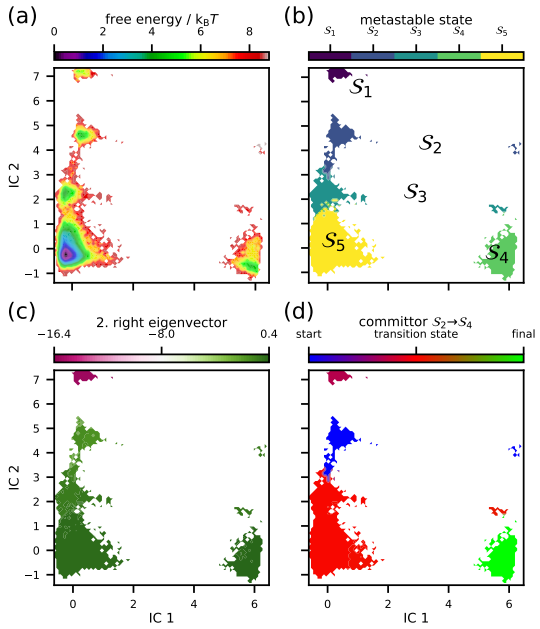


Figure 1.6: Example analysis of the conformational dynamics of a pentapeptide backbone: **(a)** The reweighted free energy surface projected onto the first two independent components exhibits five minima which **(b)** PCCA++ identifies as five metastable states. **(c)** The second right eigenvector shows that the slowest process shifts probability between the least probable state ($\mathcal{S}_1$) and the other states, in particular states ($\mathcal{S}_4, \mathcal{S}_5$), whereas **(d)** the committor $\mathcal{S}_2 \to \mathcal{S}_4$ indicates that states $\mathcal{S}_{(1,3,5)}$ act as a transition region between states $\mathcal{S}_2$ and $\mathcal{S}_4$. Reprinted from [55].

We can now directly extract several thermodynamic and kinetic properties from the estimated and validated model. An example of the former is the free energy surface in the projection onto the first two ICs (Fig. 1.6a) reweighted by the MSM stationary distribution.

A spectral clustering using the PCCA++ algorithm [73–75] allows us to coarse-grain the 75 k-means states into five metastable states (Fig. 1.6b) $\mathcal{S}_i$, $i = 1, \ldots, 5$. Often, the two different sets of states are distinguished by using the terms *microstates* for the highly-resolved k-means states, and *macrostates* for the metastable states. We here assume that each microstate is uniquely assigned to a macrostate (crisp assignment). We approximate the stationary probabilities for the macrostates and relative free energies between them (defined up to an additive constant)

macrostate $\mathcal{S}_i$	$\pi_{\mathcal{S}_i}$	$G_{\mathcal{S}_i}/\mathrm{k_B}T$
$\mathcal{S}_1$	0.004	5.567
$\mathcal{S}_2$	0.014	4.293
$\mathcal{S}_3$	0.021	3.841
$\mathcal{S}_4$	0.021	3.875
$\mathcal{S}_5$	0.940	0.062

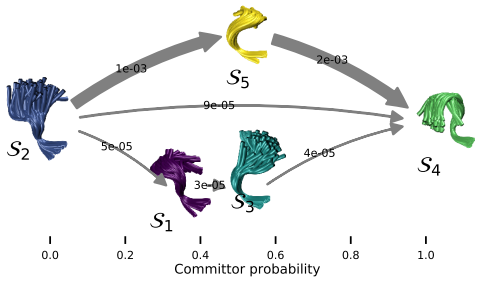


Figure 1.7: Example analysis of the conformational dynamics of a pentapeptide backbone: visualization of the transition paths from $\mathcal{S}_2$ to $\mathcal{S}_4$. Metastable states $\mathcal{S}_{(1-5)}$ are represented by an ensemble of representative structures and are arranged along the horizonal axis according to their committor probabilities. The three main transition pathways starting from $\mathcal{S}_2$ and ending in $\mathcal{S}_4$ are depicted by gray arrows with thickness proportional to the transition flux. The dominant pathway proceeds through $\mathcal{S}_5$.

using the relation

$$G_{\mathcal{S}_i} = -\mathrm{k_B} T \ln \sum_{j \in \mathcal{S}_i} \pi_j, \tag{1.35}$$

where π_j denotes the MSM stationary probability of the j^{th} microstate.

In order to interpret the slowest relaxation timescales, we refer to the (right) eigenvectors. This enables us to specifically study what conformational changes are happening on a particular timescale independently of the equilibrium distribution. The first right eigenvector corresponds to the stationary process with unity eigenvalue. The second right eigenvector, on the other hand, corresponds to the slowest non-trivial process in the system. Note that the eigenvectors are real because detailed balance (Eq. (1.25)) has been enforced during MSM estimation [71]. The minimal and maximal components of the second (and higher) right eigenvector(s) indicate the microstates between which the process shifts probability density. The relaxation timescale of this exchange process is exactly the corresponding implied timescale, which can be computed from its corresponding eigenvalue using Eq. (1.26). In the projection onto the first two TICA components, we identify the slowest MSM process as a probability shift between macrostate $\mathcal{S}_1$ and the rest of the system, with macrostates $\mathcal{S}_4$ and $\mathcal{S}_5$ in particular (Fig. 1.6c).

The mean first passage times (MFPTs) out of and into the macrostate $\mathcal{S}_1$ compute to

direction	mean / ns		std / ns
$\mathcal{S}_1 \to \mathcal{S}_{(2,3,4,5)}$	9.0	$\pm$	1.9
$\mathcal{S}_{(2,3,4,5)} \to \mathcal{S}_1$	2496.4	$\pm$	470.0

using the Bayesian MSM.

TPT [76, 77] is a method used to analyze the statistics of transition pathways. TPT as implemented in Ref. [48] can be conveniently applied to the estimated MSM. Here, we compute the TPT flux between macrostates $\mathcal{S}_2$ and $\mathcal{S}_4$ (Fig. 1.6d). The committor projection onto the first two TICA components shows that it is constant within the metastable

states defined above. Transition regions (macrostates $\mathcal{S}_{(1,3,5)}$) can be identified by committor values $\approx \frac{1}{2}$.

The transition network can be additionally visualized by plotting representative structures of the five metastable states $\mathcal{S}_{(1-5)}$ according to their committor probability (Fig. 1.7). It is easy to see from this depiction that the dominant pathway from $\mathcal{S}_2$ to $\mathcal{S}_4$ proceeds through $\mathcal{S}_5$.

1.3.10 VAMPnets

The above presented workflow represents estimation, validation, and analysis of classical MSMs. With the advent of deep learning (DL) methods in recent years, efforts were taken to cast this multi-step procedure into an end-to-end DL framework. A key development to that end was the derivation of variational scores that yield loss functions for the underlying optimization problem, in particular the VAMP scores (cf. Sec. 1.3.2) that can be used to optimize for basis functions that represent the slow dynamics of the system. Given that these basis function often are highly nonlinear, artificial deep neural networks are a good candidate for the task, an idea that was implemented under the name VAMPnets [78].

Summarizing Ref. [78], VAMPnets define two deep neural network lobes that encode the molecular configurations $\mathbf{q}_t$ and $\mathbf{q}_{t+\tau}$ at times t and $t+\tau$, respectively. The neural network lobes represent the basis functions χ_t and $\chi_{t+\tau}$ (cf. Sec. 1.3.2). In practice, both lobes usually share weights, i.e., $\chi = \chi_t = \chi_{t+\tau}$. In that case, the variational principle (Sec. 1.3.2) simplifies to using the same function for both t and $t+\tau$. The output layer of the neural networks has a defined number of nodes i; it is defined such that $\sum_i \chi(\mathbf{q})_i = 1$ and $\chi(\mathbf{q})_i \geq 0 \;\forall\; i$. Though not being a requirement, this choice allows us to interpret the neural network outputs as coarse-grained representations of the state space, i.e., the probability of belonging to a certain metastable state i.

The deep neural network parameters are optimized by first evaluating the covariance matrices using Eqs. (1.31) and by computing the approximated Koopman matrix from them using Eq. (1.32). Different VAMP-scores can be estimated now (cf. Eq. (1.33)), including the VAMP-2 score that was used in the original VAMPnet paper [78] or the VAMP-E score proposed in Ref. [50]. Whichever score is chosen, its negative is subsequently used as a loss function. As all involved operations are differentiable, the score can be minimized to obtain the optimal neural network weights. Please compare Sec. 6.4 for more details on VAMPnets and how to estimate them from data.

It is worth noting the relationship between VAMPnets and MSMs at this point. MSMs, as introduced here, can be seen as approximations to the PF operator (Sec. 1.1.5), whereas VAMPnets work with Koopman operators (cf. Sec. 1.3.2). These operators, however, are adjoint to each other [53], i.e., they carry the same information content encoded in a different way. Therefore, we can obtain a transition operator from VAMPnets that, in the special case of equilibrium sampling and crisp state assignments, is equivalent to the transition matrix.

This transition operator can be viewed as the Koopman operator in a whitened space, i.e., is accessible by a whitening operation of the Koopman operator in the original space $\bar{\mathbf{K}}$:

$$\bar{\mathbf{K}} = \mathbf{C}_{00}^{1/2}\mathbf{C}_{00}^{-1}\mathbf{C}_{0\tau}\mathbf{C}_{\tau\tau}^{-1/2} \tag{1.36}$$

$$\rightarrow \widetilde{\mathbf{P}} = \mathbf{C}_{00}^{-1}\mathbf{C}_{0\tau} \tag{1.37}$$

In the case of an MSM, this definition encodes the maximum likelihood estimator of the transition matrix [71]: With indicator basis functions (Eq. (1.17)), $\mathbf{C}_{0\tau}$ becomes the so-called *count matrix* and $\mathbf{C}_{00}^{-1}$ is a diagonal matrix with elements of inverse total counts per state. However, we note that in a usual non-equilibrium-sampling situation, $\widetilde{\mathbf{P}}$ only approximately satisfies the properties of a transition matrix. We can still apply the analysis tools shown in Sec. 1.3.9 and, in particular, follow slow processes along the eigenvectors of that matrix.

1.4 Modeling molecular machines

In the previous Chapters, theory and methods have been discussed for the purpose of modeling the kinetics of biomolecular machines. However, we have not mentioned what biomolecular machines (or, more accurately, proteins) actually are, or why they are important. As the application cases presented in this thesis (Chapters 4 and 5) are studies of protein or protein-drug binding dynamics, we give a brief general overview of the targets of these studies below.

Proteins are the major constituents of cells and fulfill most of a cell's function. They perform tasks ranging from enzymatic reactions through cellular signaling to immune responses [79]. In other words, proteins enable life in all forms that we know, and are therefore a fascinating and abundant research field. However, dynamics play a crucial role for proteins to function, making them hard to fully describe by experimental methods. Obtaining an appropriate description of protein dynamics necessitates spatial and temporal resolutions of Ångstroms and nanoseconds, respectively, to capture the underlying motions of single atoms. Achieving such resolutions experimentally is difficult, if not impossible. However, computer simulations such as MD simulations (Sec. 1.2) can aid in this task, providing mechanistic insights at theoretically arbitrary spatio-temporal resolutions. Their predictions can be corroborated with experimental studies or help to explain them [24, 36, 80].

1.4.1 Protein structure and dynamics

Proteins are macromolecules that consist of 20 different building blocks that are the amino acids. These building blocks are chained up to an unbranched polypeptide chain that can range from a few to thousands of amino acids in length. The sequence information is read from the DNA in the cell nucleus, transcribed to an RNA sequence, transported from the nucleus to the cytosol, and finally translated into an amino acid chain

by ribosomes. In brief, the translation of the RNA sequence into a protein requires formation of peptide bonds between adjacent amino acids that is catalyzed by ribosomes, and subsequent folding of the polypetid chain into a complex, three-dimensional protein structure [79].

The folding and structure of a protein is determined by different non-covalent forces such as Coulomb interactions between charged amino acid side chains. Furthermore, an important contribution to protein shape and stability is made by hydrophobicity of certain amino acid residues, which causes hydrophobic side chains to cluster in the protein core [79].

An interesting thought experiment, known as Levinthal's paradox [81, 82], is to note that each of the amino acids in a protein could be in one of multiple states. We could conclude that the number of protein folds must be very large and that protein folding must therefore be extremely slow (as in, too many conformations have to be sampled in order to find the right one). However, it is paradoxically observed that most proteins fold within seconds, which can be resolved by assuming that locally unfavorable conformations have a higher energy than favorable ones [82].

In many cases, all information relevant for folding a protein into its native state (in which the protein fulfills its function) is encoded in its DNA sequence, a postulate known as Anfinsen's dogma [83, 84]. A whole field of computational research has been dedicated to predicting a three-dimensional protein structure from sequence alone, an endeavor that culminated only recently in the development of a deep learning method called AlphaFold 2 [85]. However, it should be noted that not all folding processes follow this principle, as protein folding is not always self-catalized and, e.g., may require molecular chaperones [79].

The details of protein folding are beyond the scope of this thesis. Most proteins fold into a native state, which we here assume is measured experimentally (e.g., by X-ray crystallography) and available to us. It should be noted, though, that in general, the state of a protein under experimental conditions (e.g., crystallized proteins for X-ray) may differ from the native state.

Under physiological conditions, proteins are subject to thermal fluctuations and therefore do not rest in a single static structure. The high macromolecular complexity fosters a complex dynamics that is expressed on a broad range of timescales [36]. Protein dynamics plays an existential role for function, for example in neurotransmitter exocytosis [86], protein synthesis [87], the movement of molecular motors [88] or active transmembrane transporters [89].

To assess the statistical nature of protein dynamics, we often refer to a protein's energy landscape. The energy landscape of a protein, in brief, is defined by the energy of a given protein conformation, i.e., it usually lives in a very high-dimensional space. Protein dynamics had already been described by energy landscapes in the mid-70s [90] and gained popularity with the computational study of protein folding in the 90s [91]. Energy landscapes are therefore useful for understanding different kinds of protein dynamics. To achieve a human-readable model, energy landscapes are often projected into a low-dimensional space such as a 2D map of a Ramachandran plot [92, 93]. We

note that the choice of a meaningful space is crucial in this process, as the resulting *landscape* model is fundamentally determined by this choice.

In this thesis, protein dynamics will instead be described by Markov models. They do not necessitate projections into a low-dimensional space but instead describe exchange kinetics between metastable (protein) states. Compare Sec. 1.3 for a comprehensive discussion of Markov models.

1.5 Conceptual problems with existing kinetic modeling approaches

In the Sections 1.2 and 1.3, we have reviewed the basics of MD simulations and MSM analyses, and discussed the sampling problem that arises from the gap between integration time steps and biological timescales (Sec. 1.2.5). It has additionally become clear (Sec. 1.4) that not only time plays a role, but also space: Proteins are macromolecules and therefore are intrinsically high-dimensional in their Cartesian description. For example, our model of the TMPRSS2 serine protease domain (Chapter 5) has 3611 atoms (or 33,990 atoms with explicit water molecules), yielding $3 \times 3611 = 10,833$ Cartesian dimensions for the molecule alone. This poses computational and, especially, memory burdens not only for MD simulations but also for kinetic modeling methods.

1.5.1 The scaling problem of global states

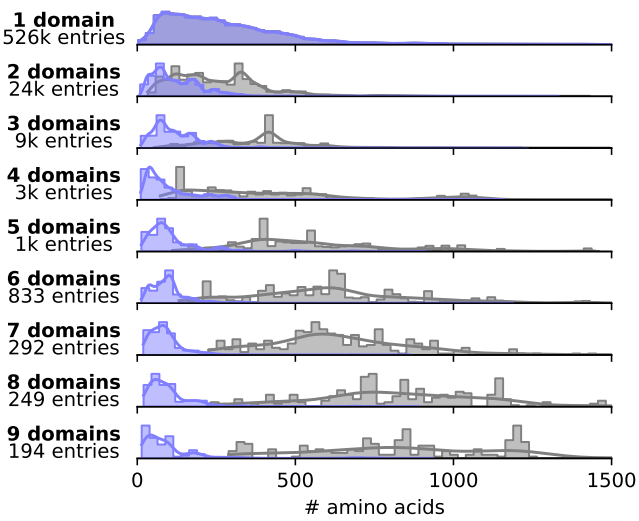


Figure 1.8: Size scaling of all proteins found in UniProt [94]. Number of amino acids per domain is depicted by blue histograms, number of amino acids of the full protein in grey. The number of proteins in a specific group, i.e.,with a fixed number of domains, is annotated.

The paradigm of global models is directly mirrored in the descriptors that have been used to model biomolecules: Simple one-dimensional molecular features are used, e.g.,

in protein folding, such as RMSD to a reference structure (usually to the folded state), the fraction of formed native contacts [95, 96], or solvent accessible surface area [97].

Low-dimensional descriptors, often denoted as *collective variables* (CVs), have been applied to great success in metadynamics [98–100] and related enhanced sampling methods. Although enhanced sampling is out of scope of this thesis, we note that the choice of CVs entails a definition of a basis function that represents the global phase space of a protein. CV-based methods are subject to the same intrinsic scaling problem as described above, although biasing approaches mitigate the problem to some extent. This becomes particularly clear when considering methods such as infrequent metadynamics [101] that aim to extract transition rates between (global) metastable states.

Furthermore, high-dimensional feature functions such as backbone dihedral angles or minimal residue-residue distances have been used [97, 102]. As kinetic modeling often requires to work in a low-dimensional space, methods such as principal component analysis (PCA) [103, 104] or TICA [69] are often applied to map hundreds or thousands of feature dimensions into a meaningful lower-dimensional subspace. In general, this mapping is still a descriptor of a global state space.

A special case is hierarchical TICA (hTICA) [105] that, motivated by the high memory demands of TICA, estimates TICA approximations to sub-sets of input features (termed level 1 TICA) to subsequently combine their dominant eigenfunctions into a level 2 TICA. As level 1 TICA can in principle be estimated in a spatially distributed fashion, hTICA can methodologically be considered a precursor method of Independent Markov Decomposition (IMD, Sec. 3). However, to our knowledge, hTICA has not been used to model spatially distributed *local* features in the sense of IMD.

Put in a broader context, time series analysis methods are often developed to work on global states. Among others, this includes MSM related methods such as HMMs [57, 59, 106] (Sec. 1.3.3) or augmented Markov models [107]. The concept of a global state description is hidden in linear dimension reduction methods such as the ones cited above, as well as in deep learning frameworks such as VAMPNets [78] (Sec. 1.3.10) or auto-encoder based methods [108, 109]. Furthermore, global state descriptions are not limited to the field of computational molecular biology but can be found in other transfer operator derived techniques such as the ones presented in Ref. [15].

However, the sampling demands for global kinetic models increase drastically for large systems, or in other words: The bigger the system, the more data we need to estimate a kinetic model such as an MSM. Driven by the curse of dimensionality [110, p. 7], the number of global states of a molecular system scales exponentially with system size [111], imposing a practical limit to any method that directly attempts to discover all these states and quantify exchange kinetics between them. In fact, larger systems require more data but are simultaneously also more computationally expensive to simulate–the vicious cycle of molecular dynamics (MD) simulations. Decomposing systems into smaller, weakly connected subsystems (or local domains) does not suffer from the same scaling problem, given that such a decomposition can be found. It can be viewed as a dynamical extension to describing molecular machines as consisting of multiple domains (Fig. 1.8) that, structurally, are often treated independently of each

other. As most existing methods rely on a global description of state space, the main objective of the methods developed here is to deal with the described scaling problem by decomposing the underlying dynamics into local domains.

Bibliography

[1] G. E. P. Box. “Science and Statistics”. *Journal of the American Statistical Association* 71.356 (1976), pp. 791–799.

[2] R. Netz. *Non-Equilibrium Statistical Mechanics*. Lecture Series on Advanced Statistical Physics. Deparment of Physics, Freie Universität Berlin, Summer Term 2015. https://www.physik.fu-berlin.de/en/einrichtungen/ag/ag-netz/lehre/Scripts/index.html.

[3] D. Tong. *Lectures on Kinetic Theory*. University of Cambridge Graduate Course. University of Cambridge, 2012. https://www.damtp.cam.ac.uk/user/tong/kinetic.html.

[4] C. Schütte. *Conformational Dynamics: Modelling, Theory, Algorithm, and Application to Biomolecules*. Tech. rep. 1999.

[5] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard. “A Direct Approach to Conformational Dynamics Based on Hybrid Monte Carlo”. *J. Comput. Phys.* 151.1 (1999), pp. 146–168.

[6] 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.* 134.17 (2011), p. 174105.

[7] G. R. Bowman, V. S. Pande, and F. Noé, eds. *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*. Vol. 797. Advances in Experimental Medicine and Biology. Dordrecht: Springer Netherlands, 2014.

[8] 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”. *The Journal of Chemical Physics* 126.15 (2007), p. 155101.

[9] D. S. Lemons and A. Gythiel. “Paul Langevin’s 1908 Paper “On the Theory of Brownian Motion” [“Sur La Théorie Du Mouvement Brownien,” C. R. Acad. Sci. (Paris) **146** , 530–533 (1908)]”. *American Journal of Physics* 65.11 (1997), pp. 1079–1081.

[10] P. Langevin. “Sur La Théorie Du Mouvement Brownien”. *C R Acad Sci Paris* 146 (1908), pp. 530–533.

[11] M. Tuckerman. *Statistical Mechanics: Theory and Molecular Simulation*. Oxford Graduate Texts. OUP Oxford, 2010.

[12] F. Schwabl. *Statistische Mechanik*. 3., aktualisierte Aufl. Springer-Lehrbuch. Berlin Heidelberg: Springer, 2006.

[13] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, and V. S. Pande. “OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics”. *PLOS Comput. Biol.* 13.7 (2017), e1005659.

[14] G. Froyland, N. Santitissadeekorn, and A. Monahan. “Transport in Time-Dependent Dynamical Systems: Finite-time Coherent Sets”. *Chaos* 20.4 (2010), p. 043116.

[15] M. Hoffmann, M. Scherer, T. Hempel, A. Mardt, B. de Silva, B. E. Husic, S. Klus, H. Wu, N. Kutz, S. L. Brunton, and F. Noé. “Deeptime: A Python Library for Machine Learning Dynamical Models from Time Series Data”. *Mach. Learn.: Sci. Technol.* 3.1 (2022), p. 015009.

[16] D. Frenkel and B. Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Vol. 1. San Diego etc: Elsevier (formerly published by Academic Press), 2002.

[17] F. Noé and S. Fischer. “Transition Networks for Modeling the Kinetics of Conformational Change in Macromolecules”. *Curr. Opin. Struct. Biol.* 18.2 (2008), pp. 154–162.

[18] F. Noé. “Machine Learning for Molecular Dynamics on Long Timescales”. *Machine Learning Meets Quantum Physics*. Ed. by K. T. Schütt, S. Chmiela, O. A. von Lilienfeld, A. Tkatchenko, K. Tsuda, and K.-R. Müller. Vol. 968. Cham: Springer International Publishing, 2020, pp. 331–372.

[19] N. G. van Kampen. *Stochastic Processes in Physics and Chemistry*. 3rd ed. North-Holland Personal Library. Amsterdam: Elsevier Science & Technology, 2007.

[20] J. A. McCammon, B. R. Gelin, and M. Karplus. "Dynamics of Folded Proteins". *Nature* 267.5612 (1977), pp. 585–590.

[21] M. I. Zimmerman, J. R. Porter, M. D. Ward, S. Singh, N. Vithani, A. Meller, U. L. Mallimadugula, C. E. Kuhn, J. H. Borowsky, R. P. Wiewiora, M. F. D. Hurley, A. M. Harbison, C. A. Fogarty, J. E. Coffland, E. Fadda, V. A. Voelz, J. D. Chodera, and G. R. Bowman. "SARS-CoV-2 Simulations Go Exascale to Predict Dramatic Spike Opening and Cryptic Pockets across the Proteome". *Nat. Chem.* 13.7 (2021), pp. 651–659.

[22] M. P. Allen. "Introduction to Molecular Dynamics Simulation". *Comput. Soft Matter Synth. Polym. Proteins* 23 (2004), pp. 1–28.

[23] R. O. Dror, R. M. Dirks, J. Grossman, H. Xu, and D. E. Shaw. "Biomolecular Simulation: A Computational Microscope for Molecular Biology". *Annu. Rev. Biophys.* 41.1 (2012), pp. 429–452.

[24] S. A. Hollingsworth and R. O. Dror. "Molecular Dynamics Simulation for All". *Neuron* 99.6 (2018), pp. 1129–1143.

[25] R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig, and A. D. MacKerell. "Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ_1 and χ_2 Dihedral Angles". *J. Chem. Theory Comput.* 8.9 (2012), pp. 3257–3273.

[26] A. D. MacKerell et al. "All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins". *J. Phys. Chem. B* 102.18 (1998), pp. 3586–3616.

[27] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl. "GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers". *SoftwareX* 1–2 (2015), pp. 19–25.

[28] K.-H. Chow and D. M. Ferguson. "Isothermal-Isobaric Molecular Dynamics Simulations with Monte Carlo Volume Sampling". *Computer Physics Communications* 91.1-3 (1995), pp. 283–289.

[29] J. Åqvist, P. Wennerström, M. Nervall, S. Bjelic, and B. O. Brandsdal. "Molecular Dynamics Simulations of Water and Biomolecules with a Monte Carlo Constant Pressure Algorithm". *Chemical Physics Letters* 384.4-6 (2004), pp. 288–294.

[30] L. Verlet. "Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules". *Phys. Rev.* 159.1 (1967), p. 98.

[31] T. Darden, D. York, and L. Pedersen. "Particle Mesh Ewald: An $N \cdot \log(N)$ Method for Ewald Sums in Large Systems". *The Journal of Chemical Physics* 98.12 (1993), pp. 10089–10092.

[32] J.-P. Ryckaert, G. Ciccotti, and H. J. Berendsen. "Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes". *Journal of Computational Physics* 23.3 (1977), pp. 327–341.

[33] H. C. Andersen. "Rattle: A "Velocity" Version of the Shake Algorithm for Molecular Dynamics Calculations". *Journal of Computational Physics* 52.1 (1983), pp. 24–34.

[34] C. W. Hopkins, S. Le Grand, R. C. Walker, and A. E. Roitberg. "Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning". *J. Chem. Theory Comput.* 11.4 (2015), pp. 1864–1874.

[35] J. C. Phillips, D. J. Hardy, J. D. C. Maia, J. E. Stone, J. V. Ribeiro, R. C. Bernardi, R. Buch, G. Fiorin, J. Hénin, W. Jiang, R. McGreevy, M. C. R. Melo, B. K. Radak, R. D. Skeel, A. Singharoy, Y. Wang, B. Roux, A. Aksimentiev, Z. Luthey-Schulten, L. V. Kalé, K. Schulten, C. Chipot, and E. Tajkhorshid. "Scalable Molecular Dynamics on CPU and GPU Architectures with NAMD". *J. Chem. Phys.* 153.4 (2020), p. 044130.

[36] K. Henzler-Wildman and D. Kern. “Dynamic Personalities of Proteins”. *Nature* 450.7172 (2007), pp. 964–972.

[37] N. Plattner, S. Doerr, G. D. Fabritiis, and F. Noé. “Complete Protein–Protein Association Kinetics in Atomic Detail Revealed by Molecular Dynamics Simulations and Markov Modelling”. *Nat. Chem.* 9.10 (2017), p. 1005.

[38] F. Rao and M. Spichty. “Thermodynamics and Kinetics of Large-Time-Step Molecular Dynamics”. *J. Comput. Chem.* 33.5 (2012), pp. 475–483.

[39] J. D. Chodera and F. Noé. “Markov State Models of Biomolecular Conformational Dynamics”. *Current Opinion in Structural Biology* 25 (2014), pp. 135–144.

[40] T. J. Lane, D. Shukla, K. A. Beauchamp, and V. S. Pande. “To Milliseconds and beyond: Challenges in the Simulation of Protein Folding”. *Current Opinion in Structural Biology* 23.1 (2013), pp. 58–65.

[41] B. E. Husic and V. S. Pande. “Markov State Models: From an Art to a Science”. *J. Am. Chem. Soc.* 140.7 (2018), pp. 2386–2396.

[43] S. Klus, P. Koltai, and C. Schütte. “On the Numerical Approximation of the Perron-Frobenius and Koopman Operator”. *J. Comput. Dyn.* 3.1 (2016), pp. 1–12.

[44] F. Nüske, H. Wu, J.-H. Prinz, C. Wehmeyer, C. Clementi, and F. Noé. “Markov State Models from Short Non-Equilibrium Simulations—Analysis and Correction of Estimation Bias”. *The Journal of Chemical Physics* 146.9 (2017), p. 094104.

[45] W. C. Swope, J. W. Pitera, and F. Suits. “Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 1. Theory”. *J. Phys. Chem. B* 108.21 (2004), pp. 6571–6581.

[46] F. Noé, I. Horenko, C. Schütte, and J. C. Smith. “Hierarchical Analysis of Conformational Dynamics in Biomolecules: Transition Networks of Metastable States”. *J. Chem. Phys.* 126.15 (2007), p. 155102.

[47] P. Koltai, H. Wu, F. Noé, and C. Schütte. “Optimal Data-Driven Estimation of Generalized Markov State Models for Non-Equilibrium Dynamics”. *Computation* 6.1 (2018), p. 22.

[48] F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl. “Constructing the Equilibrium Ensemble of Folding Pathways from Short Off-Equilibrium Simulations”. *Proc. Natl. Acad. Sci.* 106.45 (2009), pp. 19011–19016.

[49] F. Noé and F. Nüske. “A Variational Approach to Modeling Slow Processes in Stochastic Dynamical Systems”. *Multiscale Model. Simul.* 11.2 (2013), pp. 635–655.

[50] H. Wu and F. Noé. “Variational Approach for Learning Markov Processes from Time Series Data”. *J Nonlinear Sci* (2019).

[51] B. O. Koopman. “Hamiltonian Systems and Transformation in Hilbert Space”. *Proc. Natl. Acad. Sci. U.S.A.* 17.5 (1931), pp. 315–318.

[52] I. Mezić. “Spectral Properties of Dynamical Systems, Model Reduction and Decompositions”. *Nonlinear Dyn* 41.1-3 (2005), pp. 309–325.

[53] S. Klus, F. Nüske, P. Koltai, H. Wu, I. Kevrekidis, C. Schütte, and F. Noé. “Data-Driven Model Reduction and Transfer Operator Approximation”. *J Nonlinear Sci* 28.3 (2018), pp. 985–1010.

[54] I. Mezić. “Analysis of Fluid Flows via Spectral Properties of the Koopman Operator”. *Annu. Rev. Fluid Mech.* 45.1 (2013), pp. 357–378.

[55] C. Wehmeyer, M. K. Scherer, T. Hempel, B. E. Husic, S. Olsson, and F. Noé. “Introduction to Markov State Modeling with the PyEMMA Software [Article v1.0]”. *LiveCoMS* 1.1 (2019), p. 5965.

[56] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. “A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains”. *Ann. Math. Stat.* 41.1 (1970), pp. 164–171.

[57] L. R. Rabiner. “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”. *Proc. IEEE* 77.2 (1989), pp. 257–286.

[58] J.-H. Prinz, J. D. Chodera, and F. Noé. “Spectral Rate Theory for Two-State Kinetics”. *Phys. Rev. X* 4.1 (2014).

[59] 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.* 139.18 (2013), p. 184114.

[60] J. D. Chodera, P. Elms, F. Noé, B. Keller, C. M. Kaiser, A. Ewall-Wice, S. Marqusee, C. Bustamante, and N. S. Hinrichs. “Bayesian Hidden Markov Model Analysis of Single-Molecule Force Spectroscopy: Characterizing Kinetics under Measurement Uncertainty”. *ArXiv11081430 Cond-Mat Physicsphysics Q-Bio* (2011). arXiv: 1108.1430 [cond-mat, physics:physics, q-bio].

[61] M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz, and F. Noé. “PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models”. *J. Chem. Theory Comput.* 11.11 (2015), pp. 5525–5542.

[62] B. E. Husic, R. T. McGibbon, M. M. Sultan, and V. S. Pande. “Optimized Parameter Selection Reveals Trends in Markov State Models for Protein Folding”. *J. Chem. Phys.* 145.19 (2016), p. 194103.

[63] W. Humphrey, A. Dalke, and K. Schulten. “VMD: Visual Molecular Dynamics”. *J. Mol. Graph.* 14.1 (1996), pp. 33–38.

[64] F. Noé and C. Clementi. “Collective Variables for the Study of Long-Time Kinetics from Molecular Trajectories: Theory and Methods”. *Curr. Opin. Struct. Biol.* 43 (2017), pp. 141–147.

[65] F. Nüske, B. G. Keller, G. Pérez-Hernández, A. S. J. S. Mey, and F. Noé. “Variational Approach to Molecular Kinetics”. *J. Chem. Theory Comput.* 10.4 (2014), pp. 1739–1752.

[66] F. Noé and C. Clementi. “Kinetic Distance and Kinetic Maps from Molecular Dynamics Simulation”. *J. Chem. Theory Comput.* 11.10 (2015), pp. 5002–5011.

[67] R. T. McGibbon and V. S. Pande. “Variational Cross-Validation of Slow Dynamical Modes in Molecular Kinetics”. *The Journal of Chemical Physics* 142.12 (2015), p. 124105.

[68] B. E. Husic and V. S. Pande. “Note: MSM Lag Time Cannot Be Used for Variational Model Selection”. *J. Chem. Phys.* 147.17 (2017), p. 176101.

[69] G. Pérez-Hernández, F. Paul, T. Giorgino, G. D. Fabritiis, and F. Noé. “Identification of Slow Molecular Order Parameters for Markov Model Construction”. *J. Chem. Phys.* 139.1 (2013), p. 015102.

[70] C. R. Schwantes and V. S. Pande. “Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9”. *J. Chem. Theory Comput.* 9.4 (2013), pp. 2000–2009.

[71] B. Trendelkamp-Schroer, H. Wu, F. Paul, and F. Noé. “Estimation and Uncertainty of Reversible Markov Models”. *J. Chem. Phys.* 143.17 (2015), p. 174101.

[72] F. Noé. “Probability Distributions of Molecular Observables Computed from Markov Models”. *J. Chem. Phys.* 128.24 (2008), p. 244103.

[73] S. Röblitz and M. Weber. “Fuzzy Spectral Clustering by PCCA+: Application to Markov State Models and Data Classification”. *Adv. Data Anal. Classif.* 7.2 (2013), pp. 147–179.

[74] P. Deuflhard and M. Weber. “Robust Perron Cluster Analysis in Conformation Dynamics”. *Linear Algebra Appl.* 398 (2005), pp. 161–184.

[75] S. Kube and M. Weber. “A Coarse Graining Method for the Identification of Transition Rates between Molecular Conformations”. *J. Chem. Phys.* 126.2 (2007), p. 024103.

[76] W. E. and E. Vanden-Eijnden. “Towards a Theory of Transition Paths”. *J. Stat. Phys.* 123.3 (2006), pp. 503–523.

[77] P. Metzner, C. Schütte, and E. Vanden-Eijnden. “Transition Path Theory for Markov Jump Processes”. *Multiscale Model. Simul.* 7.3 (2009), pp. 1192–1219.

[78] A. Mardt, L. Pasquali, H. Wu, and F. Noé. “VAMPnets for Deep Learning of Molecular Kinetics”. *Nat. Commun.* 9.1 (2018), pp. 1–11.

[79] B. Alberts. “Molecular Biology of the Cell / Bruce Alberts, Alexander Johnson, Julian Lewis, David Morgan, Martin Raff, Keith Roberts Und Peter Walter.” *Molecular Biology of the Cell*. Sixth edition, international student edition. New York, NY [u.a: Garland Science, Taylor & Francis Group, 2015.

[80] M. Karplus and J. A. McCammon. “Molecular Dynamics Simulations of Biomolecules”. *Nat. Struct Biol.* 9.9 (2002), pp. 646–652.

[81] C. Levinthal. “How to Fold Graciously”. *Mossbauer Spectrosc. Biol. Syst.* 67 (1969), pp. 22–24.

[82] R. Zwanzig, A. Szabo, and B. Bagchi. “Levinthal’s Paradox.” *Proceedings of the National Academy of Sciences* 89.1 (1992), pp. 20–22.

[83] C. B. Anfinsen, E. Haber, M. Sela, and F. H. White. “The Kinetics of Formation of Native Ribonuclease during Oxidation of the Reduced Polypeptide Chain”. *Proc. Natl. Acad. Sci. U.S.A.* 47.9 (1961), pp. 1309–1314.

[84] C. B. Anfinsen. “Principles That Govern the Folding of Protein Chains”. *Science* 181.4096 (1973), pp. 223–230.

[85] J. Jumper et al. “Highly Accurate Protein Structure Prediction with AlphaFold”. *Nature* 596.7873 (2021), pp. 583–589.

[86] T. C. Südhof. “Neurotransmitter Release: The Last Millisecond in the Life of a Synaptic Vesicle”. *Neuron* 80.3 (2013), pp. 675–690.

[87] H. Berchtold, L. Reshetnikova, C. O. A. Reiser, N. K. Schirmer, M. Sprinzl, and R. Hilgenfeld. “Crystal Structure of Active Elongation Factor Tu Reveals Major Domain Rearrangements”. *Nature* 365.6442 (1993), pp. 126–132.

[88] K. Kinbara and T. Aida. “Toward Intelligent Molecular Machines: Directed Motions of Biological and Artificial Molecules and Assemblies”. *Chem. Rev.* 105.4 (2005), pp. 1377–1400.

[89] R. J. P. Dawson and K. P. Locher. “Structure of a Bacterial Multidrug ABC Transporter”. *Nature* 443.7108 (2006), pp. 180–185.

[90] R. H. Austin, K. W. Beeson, L. Eisenstein, H. Frauenfelder, and I. C. Gunsalus. “Dynamics of Ligand Binding to Myoglobin”. *Biochemistry* 14.24 (1975), pp. 5355–5373.

[91] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes. “Theory of Protein Folding: The Energy Landscape Perspective”. *Annu. Rev. Phys. Chem.* 48.1 (1997), pp. 545–600.

[92] G. N. Ramakrishnan, C. Ramakrishnan, and V. Sasisekharan. “Stereochemistry of Polypeptide Chain Configurations”. *J. mol. Biol* 7 (1963), pp. 95–99.

[93] C. Ramakrishnan and G. Ramachandran. “Stereochemical Criteria for Polypeptide and Protein Chain Conformations”. *Biophysical Journal* 5.6 (1965), pp. 909–933.

[94] The UniProt Consortium et al. “UniProt: The Universal Protein Knowledgebase in 2021”. *Nucleic Acids Res.* 49.D1 (2021), pp. D480–D489.

[95] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw. “How Fast-Folding Proteins Fold”. *Science* 334.6055 (2011), pp. 517–520.

[96] H. A. Scheraga, M. Khalili, and A. Liwo. “Protein-Folding Dynamics: Overview of Molecular Simulation Techniques”. *Annu. Rev. Phys. Chem.* 58.1 (2007), pp. 57–83.

[97] M. K. Scherer, B. E. Husic, M. Hoffmann, F. Paul, H. Wu, and F. Noé. “Variational Selection of Features for Molecular Kinetics”. *J. Chem. Phys.* 150.19 (2019), p. 194108.

[98] A. Laio and M. Parrinello. “Escaping Free-Energy Minima”. *Proceedings of the National Academy of Sciences* 99.20 (2002), pp. 12562–12566.

[99] O. Valsson, P. Tiwary, and M. Parrinello. “Enhancing Important Fluctuations: Rare Events and Metadynamics from a Conceptual Viewpoint”. *Annu. Rev. Phys. Chem.* 67.1 (2016), pp. 159–184.

[100] P. Tiwary and B. J. Berne. “Spectral Gap Optimization of Order Parameters for Sampling Complex Molecular Systems”. *Proc Natl Acad Sci USA* 113.11 (2016), pp. 2839–2844.

[101] P. Tiwary and M. Parrinello. “From Metadynamics to Dynamics”. *Phys. Rev. Lett.* 111.23 (2013), p. 230602.

[102] N. Plattner and F. Noé. “Protein Conformational Plasticity and Complex Ligand-Binding Kinetics Explored by Atomistic Simulations and Markov Models”. *Nat. Commun.* 6 (2015), p. 7653.

[103] K. Pearson. “LIII. On Lines and Planes of Closest Fit to Systems of Points in Space”. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* 2.11 (1901), pp. 559–572.

[104] H. Hotelling. “Analysis of a Complex of Statistical Variables into Principal Components.” *J. Educ. Psychol.* 24.6 (1933), p. 417.

[105] G. Pérez-Hernández and F. Noé. “Hierarchical Time-Lagged Independent Component Analysis: Computing Slow Modes and Reaction Coordinates for Large Molecular Systems”. *J. Chem. Theory Comput.* 12.12 (2016), pp. 6118–6129.

[106] L. E. Baum and T. Petrie. “Statistical Inference for Probabilistic Functions of Finite State Markov Chains”. *Ann. Math. Stat.* 37.6 (1966), pp. 1554–1563.

[107] S. Olsson, H. Wu, F. Paul, C. Clementi, and F. Noé. “Combining Experimental and Simulation Data of Molecular Processes via Augmented Markov Models”. *Proc. Natl. Acad. Sci.* 114.31 (2017), pp. 8265–8270.

[108] C. Wehmeyer and F. Noé. “Time-Lagged Autoencoders: Deep Learning of Slow Collective Variables for Molecular Kinetics”. *The Journal of Chemical Physics* 148.24 (2018), p. 241703.

[109] C. X. Hernández, H. K. Wayment-Steele, M. M. Sultan, B. E. Husic, and V. S. Pande. “Variational Encoding of Complex Dynamics”. *Phys. Rev. E* 97.6 (2018), p. 062412.

[110] W. K. Grassmann and F. S. Hillier, eds. *Computational Probability*. Vol. 24. International Series in Operations Research & Management Science. Boston, MA: Springer US, 2000.

[111] S. Olsson and F. Noé. “Dynamic Graphical Models of Molecular Kinetics”. *Proc. Natl. Acad. Sci.* 116.30 (2019), pp. 15001–15006.

2

Overarching interpretation, evaluation, and discussion

2.1 Contributions of this thesis

2.1.1 Decomposing systems into smaller domains

In this thesis, we provide a possible remedy to the explosion of global state space numbers in kinetic modeling. To this end, we advance the MSM method (cf. Sec. 1.3) to spatially decompose global systems into local, independent Markovian subsystems or domains. Instead of a global state description, e.g., expressed by a simple integer, we use a set of local states describing individual, independent domains akin to an Ising model (cf. Chapter 7). In Chapter 3, we focus on developing the mathematical framework to describe such decomposed dynamics. We derive the underlying transfer operator decomposition for truly independent systems, which, for MSMs, is given by a Kronecker product: A global MSM transition matrix $\mathbf{P}$ decomposes into transition matrices $\mathbf{P}_i$ and $\mathbf{P}_j$ that describe independent domains i and j, respectively,

$$\mathbf{P} = \mathbf{P}_i \otimes \mathbf{P}_j. \tag{2.1}$$

As we show, this decomposition can be generalized for other dynamical operators and arbitrary numbers of independent systems.

The first key contribution of this thesis are two methods to estimate such a dynamical decomposition from data. First, we implement independent Markov decomposition (IMD), which decomposes MSM transition matrices as shown in Eq. (2.1). We exploit properties of the Kronecker product [1] to find a domain decomposition with a network analysis (Chapter 3). To this end, we use the discrepancy between pairwise and the product of two local VAMP-n scores $\mathcal{R}_n$ [2] (cf. Sec 1.3.2) to measure *dependency* between these domains,

$$d(i,j) = \left|\mathcal{R}_n(\mathbf{P}) - \mathcal{R}_n(\mathbf{P}_i) \cdot \mathcal{R}_n(\mathbf{P}_j)\right|. \tag{2.2}$$

The *dependency* equals to zero for truly independent systems, which represents a necessary but not a sufficient condition for independence.

In Chapter 3, we demonstrate that IMD can save three orders of magnitude of sampling data compared to MSMs when applied to a truly independent numerical benchmark system of a tetrameric ion channel [3]. Using both analytical and numerical analyses, we find that IMD can be applied and is robust even in the presence of weak couplings (for analytical results, compare Appendix A.4).

Second, we develop a deep learning method to estimate a dynamical decomposition from data, called iVAMPnets (Chapter 6). To this end, we derive a loss function L that combines the *dependency* score (Eq. (2.2)) with the variational principle for Markov processes (VAMP) (cf. Sec. 1.3.2). Making use of the VAMP-E score $\mathcal{R}_E$ developed by Wu & Noé [2], its functional form is given by

$$L = -\sum_{i<j} \mathcal{R}_E^{ij} + \xi \sum_{i<j} \frac{|\mathcal{R}_E^{ij} - \mathcal{R}_E^i \cdot \mathcal{R}_E^j|}{\mathcal{R}_E^{ij}}. \tag{2.3}$$

The first sum accounts for the VAMP, i.e., is needed to find *slow* basis functions of domains i and j. The second term penalizes *dependency* between these domains akin to the *dependency* score. The parameter ξ is a scaling factor that balances these objectives. The loss function is evaluated over all pairs of domains to secure the sampling advantage of IMD, i.e., a global model of the kinetics is estimated at no point. The VAMP-E score, in general, measures deviations between estimated and true dynamics and is constructed such that it maps local features (using the independence assumptions) into joint spaces (not using independence assumptions). Additional to the loss function, iVAMPnets come with a neural network architecture featuring a trainable mask that identifies independent domains and assigns them to protein residues.

From a practical perspective, IMD and iVAMPnets add two contributions to Markov modeling, given that a system indeed consists of weakly or uncoupled domains. a) Sampling advantages, i.e., local models can be estimated with much less data compared to global state MSMs. This advantage is analyzed in detail in Chapter 3 and particularly used in Chapter 4. b) Models can be estimated at regions of interest (ROIs), if these are dynamically independent, therefore simplifying estimation as well as analysis tasks. E.g., it is much easier to model a single protein ROI than a full protein – less degrees of freedom have to be accounted for, discretization can be conducted in a lower-dimensional space, and less (metastable) states need to be interpreted. Advantage b) becomes especially evident when considering the conformational switches discussed in Chapter 4 or modeling drug-binding kinetics in Chapter 5.

2.1.2 Applying decomposition methods to molecular biology

The second key contribution of this thesis are application studies that tackle computational molecular biology questions, using the methods sketched above. In Chapter 4, we apply IMD to MD data of synaptotagmin-1 (syt) C2A, which is a calcium sensor im-

portant for neurotransmitter release. This protein system cannot be evaluated with the classical MSM approach as the available sampling is too sparse to capture all global state-to-state transitions. IMD therefore provides a means to analyze its mechanism of action.

Syt C2A is a member of the C2 domain family that has 146,000 sequences reported in pfam [4]*. C2 domains are proteins that bind to phospholipids, e.g., in a membrane, and are often dependent on Ca^{2+} ion binding [5]. They consist of around 130 amino acid residues and form a beta sandwich structure. C2 domains are functionally promiscuous protein modules used in signaling and membrane trafficking [5]; their function can generally be described as "attach[ing] their resident proteins to phospholipid membranes" [6], a process that can be regulated by Ca^{2+} concentration.

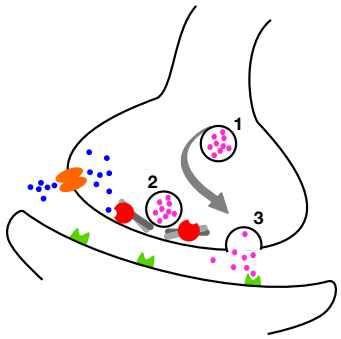


Figure 2.1: Symbolic depiction of a chemical synapse [7, 8, p. 1072]. Neurotransmitters are depicted as filled magenta polygons that start in a vesicle (1) and, in the pre-fusion state, are attached to the membrane with SNARE proteins (grey sticks, 2). Upon calcium influx (blue dots), the calcium sensor synaptotagmin-1 (red circles) binds the ions, the vesicle fuses with the membrane and neurotransmitters are released to the synaptic cleft (3).

A well-studied function of C2 domains is their role in the chemical synapse, where neurotransmitter release is triggered via syt [9]. Here, C2 domains play the role of a molecular switch that is activated by the local Ca^{2+} concentration at the release site [10] as sketched in Fig. 2.1. In short, to enable fast neural exocytosis upon calcium influx, the neural vesicles are primed, i.e., they are fixed close by the membrane by the soluble N-ethylmaleimide-sensitive factor attachment receptor (SNARE) complex. Membrane fusion, the energetically favorable next step, is clamped by the SNARE complex in concert with syt. Activated by the nerval membrane potential, calcium channels release ions into the synapse that bind to syt, causing conformational (cf. Chapter 4) and charge [11] changes that lead to syt-phospholipid binding and, ultimately, to membrane fusion [8, 9].

Neural exocytosis from the primed vesicle is extraordinarily fast and occurs on timescales below 100 µs [8]. As the details of conformational changes in syt C2A that are triggered by calcium binding are not well understood, we apply IMD (Chapter 3) to analyze these phenomena in atomistic detail. We approximate different protein loops as dynamically independent local Markovian domains (here termed *conformational*

*Accession number PF00168.

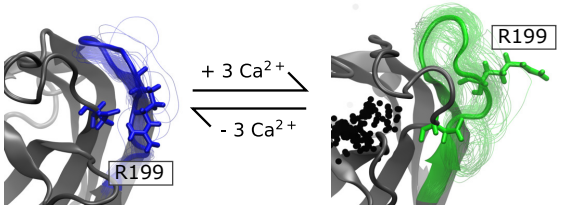


Figure 2.2: Metastable structures of an IMD subsystem model as a function of calcium binding, highlighting R199. Extracted from Fig. 4.3 [12].

switches) and assess how the dynamics in each of these domains change upon calcium binding. In particular, we find that calcium enables an arginine residue (R199) to be solvent-exposed (compare Fig. 2.2 for the structures). Given that R199 is crucial for membrane penetration [13], among other functions, this calcium-triggered conformational change may play a functional role for fast neurotransmitter release.

We augment our IMD model of syt C2A by assessing interactions between local domains in a subsequent step, an enhancement to IMD that goes beyond Chapter 3. We use a method connected to the estimated hidden Markov models to filter the local dynamics into simplified time-series to conduct information theoretic analyses. Again, we find that calcium binding has an impact, increasing directional information exchange between the domains. Our analysis sheds light on the dynamics of syt C2A and provides a quantitative, mechanistic model of protein kinetics as a function of calcium binding.

In Chapter 5, IMD is used to study the drug-binding kinetics of a serine protease. Serine proteases are enzymes that cleave peptides. Being present in organisms from viruses through prokaryotes to eukaryotes, serine-proteases are often characterized by a so-called catalytic triad that is usually formed by a nucleophilic serine, a histidine, and an aspartate [14]. For example, trypsins are a protein family of digestive enzymes that fall into the super-family of serine proteases, featuring about 62,000 family members [4]*. Trypsin (the enzyme that gave its name to the family) is produced in the human pancreas and is secreted to the duodenum where it cleaves polypeptides into smaller ones [7].

However, there are a variety of other functions that mammalian trypsins perform, such as in the immune response, blood coagulation, and fibrinolysis [14]. The trypsin family member studied in this chapter is the transmembrane protease serine 2 (TMPRSS2). It is highly expressed in the prostate epithelium compared to other human cells and has been studied in the context of prostate cancer [17, 18]. The function of TMPRSS2 in healthy humans is unknown, but it has been shown to be important for, e.g., viral entry of influenza A and corona viruses [19–24]. Its domain structure is shown in Fig. 2.3. In particular, TMPRSS2 gained attention during the SARS-CoV-2 pandemic as a factor that facilitates viral entry [25]. It is therefore a possible drug target to mitigate infection by the corona virus. It has been found to be inhibited by synthetic drugs such as camostat and nafamostat [25–27] as well as protein inhibitors [28, 29].

*As by the number of reported sequences in pfam [4], accession number PF00089.

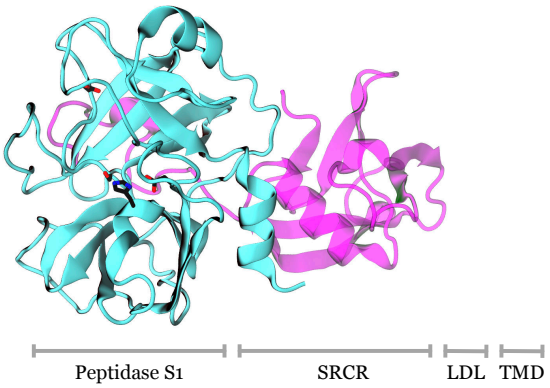


Figure 2.3: Domain structure of TMPRSS2 (from UniProt [15] accession number O15393) alongside a homology model of the Scavenger receptor cysteine-rich (SRCR) protein with the catalytically active peptidase S1 domain. Low-density lipoprotein (LDL) receptor and transmembrane domain (TMD) are not structurally resolved [16].

In Chapter 5, we shed light upon the mechanism of inhibition with nafamostat and camostat by modeling the active site of TMPRSS2 with the drugs as a dynamically independent local domain, i.e., as a single IMD domain. Our main contribution is a model of the mechanism of action of these drugs, which includes the formation of a reactive Michaelis complex (MC) and a subsequent covalent bond formation between enzyme and drug. Even though the covalent bond formation cannot be observed with MD, we reason that the differences in drug potency between camostat and nafamostat arise from the differences in MC populations. As we show here next to our experimental results, MC populations indeed show a good agreement with drug potency:

drug	**IC50** (experimental)	**MC population** (computational)
nafamostat	55 nM	3.1%
camostat	142 nM	1.0%

The IC50 denotes the drug concentration that is sufficient to block 50% of the enzymatic activity in the assay presented in Chapter 5. A more effective drug needs a lower concentration, which corresponds to a higher MC population. This model is corroborated by the MC population of GBPA (0.6%), a metabolite of camostat that is less potent than camostat [27]. Our computational model therefore provides a heuristic to quantitatively describe drug potency. Furthermore, it yields structural information about the drug binding modes in the enzymatic active site.

2.1.3 Augmenting decomposition methods, understanding their context

In Chapters 4 and 5, we apply IMD to decompose large, macromolecular systems into smaller ones. However, in both application studies, we use handcrafted decompositions of proteins into dynamically independent local domains. To make this approach more systematic and less dependent on structural intuition, we develop an end-to-end

deep learning method that combines IMD with VAMPnets [30]. To this end, we use the loss function presented in Eq. (2.3) (Chapter 6). The resulting method, iVAMPnets, provides an automated procedure for finding an optimal domain decomposition while simultaneously modeling the slow exchange kinetics at local ROIs.

iVAMPnets is tested and validated using numerical benchmark systems with known ground truth. It successfully decomposes these into domains and models their local dynamics independently. For example, a system consisting of ten independent 2-state subsystems can be solved, i.e., each subsystem feature is successfully identified from the data and the implied timescales are recovered. For comparison, a global model would need to describe all $2^{10} \times 2^{10}$ state-to-state transitions. To validate the method with a high-dimensional MD application case, we revisit syt C2A (cf. Chapter 4). We show that iVAMPnets identifies domains that are congruent to the ones identified by the network analysis presented in Chapter 3 (Sec. 3.3.3). Furthermore, local loop dynamics are described by high-resolution models (cf. Appendix D.4) that are comparable to the in-depth analysis conducted in Chapter 4. Therefore, iVAMPnets provide a systematic approach for modeling local dynamics of high-dimensional MD systems.

Lastly, in Chapter 7, we generalize our decomposition approach and discuss IMD and iVAMPnets in the context of other methods that use spatial decomposition techniques for kinetic modeling. To this end, IMD and iVAMPnets are interpreted as a truncated series expansion and related to the way we formally describe transfer operators, which is explained below using IMD and the transition matrix decomposition Eq. (2.1) as an example. Transition matrices (as well as any other matrix or, more generally, tensor) can be written in different formats, e.g., the *canonical format* [31]: A matrix $\mathbf{P} \in \mathbb{R}^{N \times N}$ can be expressed as a sum over Kronecker products of smaller matrices $\mathbf{P}_i^k \in \mathbb{R}^{n \times n}$ with $N = n^d$,

$$\mathbf{P} = \sum_{k=1}^{r} \bigotimes_{i=1}^{d} \mathbf{P}_i^k. \tag{2.4}$$

The *canonical rank* r is intrinsic to the matrix, and $d \in \mathbb{N}_+$ is a model parameter that describes the number of product splits. Although not required, we assume that all $\mathbf{P}_i^k$ have the same size for simplicity. This reformulation, in general, does not change the information content, i.e., can be understood as a lossless compression. However, it has advantages over a dense matrix description if the rank r is low, as the memory consumption of a dense matrix scales with $\mathcal{O}(N^2) = \mathcal{O}(n^{2d})$. In contrast, matrices stored in the canonical format do not scale exponentially with system size but with $\mathcal{O}(r \cdot n^2 \cdot d)$, paving the way to escape the curse of dimensionality [31].

In particular motivated by such memory demands are advances in the applied mathematics community that build upon rewriting transition matrices (or in general, tensors) into more favorable formats. An example is the tensor-train format [31, 32] which decomposes a given tensor into a chain-like network of low-dimensional tensors. Tensor-trains were shown to be appropriate for systems with nearest-neighbor interactions such as Ising models [33], which are also discussed in Chapter 7. IMD is a rank 1 decomposition of a transition matrix, i.e., it truncates the series expansion after the first

term. As detailed in Chapter 3, IMD exploits weak couplings between (local) domains of macromolecules to find this decomposition, i.e., the series expansion is exact only in the case of fully independent domains. In general, IMD represents a first-order approximation (or, in other words, a lossy compression) that neglects coupling terms.

2.2 Discussion and conclusions

2.2.1 Successes and caveats of decomposition methods

In this work, we have identified a fundamental scaling problem with the kinetic modeling of large macromolecules, which arises from a negative synergy between the curse of dimensionality and the classical MD sampling problem: Larger systems require more data for kinetic modeling – at the same time, generating this data becomes computationally more expensive (cf. Sec. 1.5). We have discussed a possible remedy that challenges the current paradigm of global protein descriptors (cf. Sec. 1.5) by decomposing global dynamics into a coupling of local ones. To this end, we have proposed a local approach to Markov modeling that decomposes a protein (or any other complex) into smaller local domains, given that these domains are sufficiently weakly coupled. This idea was cast into independent Markov decomposition (IMD, Chapter 3) and iVAMPnets (Chapter 6), two methods that can be used to a) find a domain decomposition and b) model single domains independently of each other. IMD and iVAMPnets have been derived by decomposing the underlying transfer operator (assuming that the domains are truly independent), validated with appropriate benchmark models, and shown to have practical applicability to high-dimensional MD data. In particular, both method implementations have been proved to be stable even for high-dimensional systems with weak couplings, and to be a reasonable approximation of the underlying dynamics.

However, the applied decompositions of transfer operators presented in Chapter 3 (e.g., Eq. (2.1) for transition matrices) strictly applies only to the case of truly uncoupled dynamical systems. Therefore, the downstream methodological results, including IMD and iVAMPnets, are strictly true only in the case of fully uncoupled domains, a case that may be rarely applicable to real biological systems. To prove the applicability of this approximation, both IMD and iVAMPnets are evaluated with systems of various degrees of coupling, showing that the modeling error made in these cases is within tolerance and that the results indeed mirror ground truth. As the used benchmark systems range from simple toy models to high-dimensional MD simulation data, we believe that IMD and iVAMPnets are good approximations even if domains are only approximately independent.

2.2.2 How local descriptions affect application studies

We applied IMD to two problems of computational molecular biology: First, we have provided an IMD model of syt C2A, which is a small protein in the neurotransmitter release machinery (Chapter 4). This model relies on the sampling advantage of IMD over MSMs. As we show later using VAMPnets (Appendix D.3), the available MD data indeed

does not sample all possible transitions in the global space, i.e., would not be sufficient to estimate a global state MSM. Therefore, Chapter 4 successfully demonstrates that decomposition approaches such as IMD can mitigate the scaling problem of molecular kinetics modeling.

However, IMD has been carried out manually in Chapter 4 as the dependency score (Sec. 3.2.3) had not been developed at the time. Therefore, the decomposition into domains is approximate and guided by structural intuition. Though it can be mostly reconciled when using systematic approaches such as network-based pairwise dependency score analysis (Sec. 3.3.3) or iVAMPnets (Sec. 6.2.5), a new modeling error is introduced here. Even if slow orthogonal degrees of freedom are ruled out by careful MSM validation tests (Appendix B.2.4, also compare Sec. 1.3.8), a more systematic way of partitioning systems is desirable. This issue has been addressed in Chapters 3 and 6.

Second, we used IMD to analyze drug binding to the catalytic pocket of the serine protease TMPRSS2 (Chapter 5). We provide a local Markov model of the drug binding kinetics, which, in concert with our experiments, successfully establishes a computational surrogate for the potency of TMPRSS2 inhibitors. As we show, the drug binding process is governed by the dynamics of the binding pocket, and therefore IMD offers a pragmatic way of modeling it. By removing other dynamical processes from the picture, a high-resolution model of drug binding can be achieved and validated experimentally. Potentially interesting, orthogonal processes are excluded from this analysis, i.e., focussing on one specific domain (active site and drug) is a modelers choice. Other domains of TMPRSS2 could hypothetically be modeled as well but were not of interest for the underlying goal to understand the inhibition mechanisms of a potential SARS-CoV-2 drug.

Due to the lack of a crystal structure, Chapter 5 is based on MD data seeded from a homology model which was chosen based on the properties and stability of its enzymatic active site. Therefore, a comprehensive study of the whole protease domain and subsequent domain decomposition, e.g., using iVAMPnets (Chapter 6), though desirable, would likely not have succeeded using the same data. As a consequence, the identification of the analyzed IMD domain has not been carried out using the *dependency* score (Sec. 3.2.3) or related methods. Instead, the domain identification is based on two factors: For one, we draw on established biochemical knowledge (cf., e.g., Ref. [14]) to identify the enzymatic active site and substrate recognition pocket. For the other, local domain MSMs are validated carefully using implied timescales and Chapman-Kolmogorov tests (cf. 1.3.8) as it had been done in Chapter 4. Thus, slow orthogonal degrees of freedom such as influences from other parts of the protein are ruled out. We note that converged MSM validation measures are a necessary but not a sufficient condition for independence. However, given the good agreement to our experimental results, we believe that our model is an appropriate representation of the drug binding process. This assessment is backed by the crystal structure of TMPRSS2 in complex with nafamostat that was published about one year after Chapter 5 [34], validating our model of the drug-binding mode of nafamostat. Even though an approximation, IMD can therefore

be used as a pragmatic approach to dealing with time-critical questions and helps prioritizing dynamical features that are important to a given problem.

2.2.3 Manual decomposition vs. end-to-end learning

In the previous chapters, in particular Chapter 3, we perform decomposition and kinetic modeling in two sequential steps, creating a separation between system decomposition and kinetics that may contribute to model inaccuracies. This issue is tackled by deriving a loss function (Eq. (2.3)) that combines both, *dependency* score (Eq. (2.2)) and the variational principle for Markov processes (VAMP) [2] in Chapter 6.

The local approach to Markov modeling is complemented by an automated end-to-end deep learning framework, called iVAMPnets (Chapter 6). Domain decomposition and local kinetic modeling can now be conducted simultaneously. In comparison to the manual decompositions performed in Chapters 4 and 5, iVAMPnets rely on a quantitative measure of independence rather than structural intuition. Building upon the decomposition procedures developed with IMD (Chapter 3), iVAMPnets cast the whole idea of decomposed modeling into a single optimization problem, providing an easy-to-use modeling tool for the kinetic modeling of large macromolecular systems. Harvesting from deep learning and MSM decomposition methods, iVAMPnets have succeeded in automatizing domain decomposition and kinetic modeling of weakly coupled systems even in the case of highly complex and globally unconverged data samples.

However, optimizing for a representation of the slow dynamics while simultaneously decomposing a system into multiple domains is a hard optimization problem (also compare Chapter 7). Balancing these two objectives can be a non-trivial task. Even though iVAMPnets provide a much more systematic framework to decompose a system as compared to our previous work (Chapters 4 and 5), there is now less control over the intermediate steps. E.g., domain decomposition cannot be treated as an independent problem anymore. Additionally, an optimal representation of the independent Markovian dynamics may not directly give access to a human-readable domain assignment. This challenge is met, e.g., by regularization techniques that ensure crispness of the domain identification mask. We equipped iVAMPnets with a number of new validation measures to make sure that valid decompositions can be reached even for unknown systems. To guide future application studies, we additionally supplemented this chapter with a counter example of a non-decomposable system (Sec. 6.2.6). Therefore, we are confident that iVAMPnets is a useful tool that is capable to systematically decompose macromolecular systems and to study their local kinetics.

In comparison to IMD (Chapter 3), the transfer operator found by iVAMPnets is a Koopman operator, i.e., an operator that does not propagate probabilities in time but observables (cf. Sec. 1.3.2). Therefore, we cannot directly derive physical quantities such as rate models from the local VAMPnets as done, e.g., in Chapter 4. To that end, augmenting iVAMPnets with physical constraints, as done by Ref. [35], is an important task for future work.

2.2.4 Assessing couplings

Furthermore, not every system can be partitioned into smaller parts, and most real-world problems may range somewhere between strongly coupled and almost independent. In this thesis, physical coupling terms have been excluded almost entirely from the analyses and method developments, a first order approximation that enabled the derivation of a simple formalism and the exploitation of vast sampling advantages. Analyses of couplings, such as performed by the *dependency* score (Chapter 3) or information theoretic quantities (Chapter 4), are heuristics that do not represent physical coupling terms such as a coupling tensor for Ising models (cf. Sec. 7.6). They were instead developed to work without the need to estimate a global transfer operator, a route that was taken to not spoil the sampling advantages gained by IMD. Even though the *dependency* score (Sec. 3.2.3) allows for a systematic and quantitative assessment whether or not a system can be decomposed, it only measures deviations from the no-coupling assumption and may not be regarded a quantitative *measure* of the degree of coupling. Instead, it is a heuristic to separate strongly coupled from weakly coupled domains, the latter being amenable to IMD. It remains a task for future research to derive a connection to a physical coupling model or to add coupling terms to the first order approximations suggested here, without fully sacrificing the gained sampling advantages.

To understand better how coupling terms relate to the presented methods, we have linked various decomposition methods by rewriting the underlying decomposition as a truncated series expansion Eq. (2.4) (cf. Chapter 7). We speculate that the coupling terms may be hidden in higher order terms of the series expansion. However, we did not find a direct connection between the latter and the *dependency* score, hampering efforts to extract (approximate) coupling terms from IMD or iVAMPnets. Providing a formal connection would be highly interesting as a means to quantify couplings, which however may be a challenging task without knowledge of the global transfer operator.

Lastly, IMD and iVAMPnets are neither the only nor the first methods to describe dynamics by decomposing systems into local domains. In Chapter 7, we put the contributions of this thesis into a broader context, coining the umbrella term Markov field models (MFMs) to subsume the different incarnations of this idea. For example, dynamic Ising models are an MFM that features explicit couplings. Tracing dynamical decomposition methods shows that the underlying problem, exploding state spaces, is common to different scientific communities and has been dealt with for a long time. Even though there are various models that describe decompositions into domains, approximate or exact, it is often not clear how to estimate such a model from data. IMD and iVAMPnets have contributed to this more general problem as their applications are not limited to computational molecular biology.

Being beneficial for independent or weakly coupled systems in terms of sampling efficiency and expressiveness, MFMs come with challenges and open questions, too: Estimating an optimal domain decomposition from data and modeling local kinetics in these domains is a non-trivial task that still requires significant expert knowledge about the studied system and the methods used to describe it. To make MFMs suitable for a wide range of computational molecular biology applications, future research may focus

on developing improved estimation procedures, in particular to further mitigate the domain decomposition problem. Understanding MFM convergence behavior in data-sparse situations may be crucial here, an aspect that was briefly touched in Chapter 3 using simple toy models (also compare Appendix A.6.2 for a system with coupling). Especially the important use case of sufficient local and insufficient global sampling is an issue of high interest as it governs most estimation procedures. Additionally, model uncertainties need to be further quantified to minimize systematic errors related to neglecting or sparsifying system-intrinsic couplings.

2.2.5 General outlook

On a broader note, MD simulations have just crossed the exascale barrier with distributed computing on the folding@home infrastructure [36]. That said, larger and more biologically relevant systems can be simulated with all-atom MD than ever, giving us the opportunity to observe, e.g., a whole SARS-CoV-2 virus in an aerosole particle [37] or a membrane model of the endoplasmic reticulum [38] in atomistic detail through the *computational microscope* [39]. In other words, the gap between the systems studied *in vitro* and *in silico* is closing, promising exciting new insights into molecular biology in the coming years. We hypothesize that kinetic modeling methods will follow that path, yielding new mechanistic insights and predictive models for future biomolecular applications.

Bibliography

[1] C. F. Van Loan. "The Ubiquitous Kronecker Product". *Journal of Computational and Applied Mathematics* 123.1-2 (2000), pp. 85–100.

[2] H. Wu and F. Noé. "Variational Approach for Learning Markov Processes from Time Series Data". *J Nonlinear Sci* (2019).

[3] Y. Rudy and J. R. Silva. "Computational Biology in the Study of Cardiac Ion Channels and Cell Electrophysiology". *Q. Rev. Biophys.* 39.1 (2006), pp. 57–116.

[4] J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G. A. Salazar, E. L. L. Sonnhammer, S. C. E. Tosatto, L. Paladin, S. Raj, L. J. Richardson, R. D. Finn, and A. Bateman. "Pfam: The Protein Families Database in 2021". *Nucleic Acids Res.* 49.D1 (2021), pp. D412–D419.

[5] J. Rizo and T. C. Südhof. "C2-Domains, Structure and Function of a Universal Ca2+-Binding Domain". *Journal of Biological Chemistry* 273.26 (1998), pp. 15879–15882.

[6] T. C. Südhof and J. Rizo. "Chapter 139 - C_2-Domains in Ca^{2+}-Signaling". *Handbook of Cell Signaling*. Ed. by R. A. Bradshaw and E. A. Dennis. Burlington: Academic Press, 2003, pp. 95–100.

[7] J. B. Reece, L. A. Urry, M. L. Cain, S. A. Wasserman, P. V. Minorsky, R. B. Jackson, and N. A. Campbell. *Campbell Biology*. 12th edition. Boston: Pearson, 2020.

[8] T. C. Südhof. "Neurotransmitter Release: The Last Millisecond in the Life of a Synaptic Vesicle". *Neuron* 80.3 (2013), pp. 675–690.

[9] E. R. Chapman. "How Does Synaptotagmin Trigger Neurotransmitter Release?" *Annu. Rev. Biochem.* 77.1 (2008), pp. 615–641.

[10] G. J. Augustine, F. Santamaria, and K. Tanaka. "Local Calcium Signaling in Neurons". *Neuron* 40.2 (2003), pp. 331–346.

[11] A. R. Striegel, L. M. Biela, C. S. Evans, Z. Wang, J. B. Delehoy, R. B. Sutton, E. R. Chapman, and N. E. Reist. "Calcium Binding by Synaptotagmin's C2A Domain Is an Essential Element of the Electrostatic Switch That Triggers Synchronous Synaptic Transmission". *J. Neurosci.* 32.4 (2012), pp. 1253–1260.

[12] T. Hempel, N. Plattner, and F. Noé. "Coupling of Conformational Switches in Calcium Sensor Unraveled with Local Markov Models and Transfer Entropy". *J. Chem. Theory Comput.* 16.4 (2020), pp. 2584–2593.

[13] J. L. Jiménez, G. R. Smith, B. Contreras-Moreira, J. G. Sgouros, F. A. Meunier, P. A. Bates, and G. Schiavo. "Functional Recycling of C2 Domains Throughout Evolution: A Comparative Study of Synaptotagmin, Protein Kinase C and Phospholipase C by Sequence, Structural and Modelling Approaches". *J. Mol. Biol.* 333.3 (2003), pp. 621–639.

[14] L. Hedstrom. "Serine Protease Mechanism and Specificity". *Chem. Rev.* 102.12 (2002), pp. 4501–4524.

[15] The UniProt Consortium et al. "UniProt: The Universal Protein Knowledgebase in 2021". *Nucleic Acids Res.* 49.D1 (2021), pp. D480–D489.

[16] W. Humphrey, A. Dalke, and K. Schulten. "VMD: Visual Molecular Dynamics". *J. Mol. Graph.* 14.1 (1996), pp. 33–38.

[17] B. Lin, C. Ferguson, J. T. White, S. Wang, R. Vessella, L. D. True, L. Hood, and P. S. Nelson. "Prostate-Localized and Androgen-Regulated Expression of the Membrane-Bound Serine Protease TMPRSS2". *Cancer Res.* 59.17 (1999), pp. 4180–4184.

[18] J. M. Lucas, C. Heinlein, T. Kim, S. A. Hernandez, M. S. Malik, L. D. True, C. Morrissey, E. Corey, B. Montgomery, E. Mostaghel, N. Clegg, I. Coleman, C. M. Brown, E. L. Schneider, C. Craik, J. A. Simon, A. Bedalov, and P. S. Nelson. "The Androgen-Regulated Protease TMPRSS2 Activates a Proteolytic Cascade Involving Components of the Tumor Microenvironment and Promotes Prostate Cancer Metastasis". *Cancer Discovery* 4.11 (2014), pp. 1310–1325.

[19] N. Iwata-Yoshikawa, T. Okamura, Y. Shimizu, H. Hasegawa, M. Takeda, and N. Nagata. “TMPRSS2 Contributes to Virus Spread and Immunopathology in the Airways of Murine Models after Coronavirus Infection”. *J. Virol.* 93.6 (2019).

[20] B. Hatesuer, S. Bertram, N. Mehnert, M. M. Bahgat, P. S. Nelson, S. Pöhlmann, S. Pöhlman, and K. Schughart. “Tmprss2 Is Essential for Influenza H1N1 Virus Pathogenesis in Mice”. *PLoS Pathog.* 9.12 (2013), e1003774.

[21] C. Tarnow, G. Engels, A. Arendt, F. Schwalm, H. Sediri, A. Preuss, P. S. Nelson, W. Garten, H.-D. Klenk, G. Gabriel, and E. Böttcher-Friebertshäuser. “TMPRSS2 Is a Host Factor That Is Essential for Pneumotropism and Pathogenicity of H7N9 Influenza A Virus in Mice”. *J. Virol.* 88.9 (2014), pp. 4744–4751.

[22] K. Sakai, Y. Ami, M. Tahara, T. Kubota, M. Anraku, M. Abe, N. Nakajima, T. Sekizuka, K. Shirato, Y. Suzaki, A. Ainai, Y. Nakatsu, K. Kanou, K. Nakamura, T. Suzuki, K. Komase, E. Nobusawa, K. Maenaka, M. Kuroda, H. Hasegawa, Y. Kawaoka, M. Tashiro, and M. Takeda. “The Host Protease TMPRSS2 Plays a Major Role in in Vivo Replication of Emerging H7N9 and Seasonal Influenza Viruses”. *J. Virol.* 88.10 (2014), pp. 5608–5616.

[23] R. L. O. Lambertz, I. Gerhauser, I. Nehlmeier, S. R. Leist, H. Kollmus, S. Pöhlmann, and K. Schughart. “Tmprss2 Knock-out Mice Are Resistant to H10 Influenza A Virus Pathogenesis”. *J. Gen. Virol.* 100.7 (2019), pp. 1073–1078.

[24] R. L. O. Lambertz, I. Gerhauser, I. Nehlmeier, S. Gärtner, M. Winkler, S. R. Leist, H. Kollmus, S. Pöhlmann, and K. Schughart. “H2 Influenza A Virus Is Not Pathogenic in Tmprss2 Knock-out Mice”. *Virol. J.* 17.1 (2020), p. 56.

[25] M. Hoffmann, H. Kleine-Weber, S. Schroeder, N. Krüger, T. Herrler, S. Erichsen, T. S. Schiergens, G. Herrler, N.-H. Wu, A. Nitsche, M. A. Müller, C. Drosten, and S. Pöhlmann. “SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor”. *Cell* 181.2 (2020), 271–280.e8.

[26] M. Hoffmann, S. Schroeder, H. Kleine-Weber, M. A. Müller, C. Drosten, and S. Pöhlmann. “Nafamostat Mesylate Blocks Activation of SARS-CoV-2: New Treatment Option for COVID-19”. *Antimicrob. Agents Chemother.* 64.6 (2020).

[27] M. Hoffmann, H. Hofmann-Winkler, J. C. Smith, N. Krüger, P. Arora, L. K. Sørensen, O. S. Søgaard, J. B. Hasselstrøm, M. Winkler, T. Hempel, L. Raich, S. Olsson, O. Danov, D. Jonigk, T. Yamazoe, K. Yamatsuta, H. Mizuno, S. Ludwig, F. Noé, M. Kjolby, A. Braun, J. M. Sheltzer, and S. Pöhlmann. “Camostat Mesylate Inhibits SARS-CoV-2 Activation by TMPRSS2-related Proteases and Its Metabolite GBPA Exerts Antiviral Activity”. *EBioMedicine* (2021), p. 103255.

[28] N. P. Azouz, A. M. Klingler, V. Callahan, I. V. Akhrymuk, K. Elez, L. Raich, B. M. Henry, J. L. Benoit, S. W. Benoit, F. Noé, K. Kehn-Hall, and M. E. Rothenberg. “Alpha 1 Antitrypsin Is an Inhibitor of the SARS-CoV-2-Priming Protease TMPRSS2”. *Pathog Immun* 6.1 (2021), pp. 55–74.

[29] F. Ritzmann, P. Chitirala, N. Krüger, M. Hoffmann, W. Zuo, F. Lammert, S. Smola, N. Tov, N. Alagem, P. M. Lepper, S. Pöhlmann, C. Beisswenger, C. Herr, R. Bals, and for the AAT-in-COVID-19 study group. “Therapeutic Application of Alpha-1-Antitrypsin in COVID-19”. *Am J Respir Crit Care Med* (2021), rccm.202104–0833LE.

[30] A. Mardt, L. Pasquali, H. Wu, and F. Noé. “VAMPnets for Deep Learning of Molecular Kinetics”. *Nat. Commun.* 9.1 (2018), pp. 1–11.

[31] P. Gelß. “The Tensor-Train Format and Its Applications. Modeling and Analysis of Chemical Reaction Networks, Catalytic Processes, Fluid Flows, and Brownian Dynamics”. PhD thesis. Freie Universität Berlin, 2017, xvii, 160 Seiten.

[32] I. V. Oseledets. “Tensor-Train Decomposition”. *SIAM J. Sci. Comput.* 33.5 (2011), pp. 2295–2317.

[33] P. Gelß, S. Klus, S. Matera, and C. Schütte. “Nearest-Neighbor Interaction Systems in the Tensor-Train Format”. *Journal of Computational Physics* 341 (2017), pp. 140–162.

[34] B. J. Fraser, S. Beldar, A. Seitova, A. Hutchinson, D. Mannar, Y. Li, D. Kwon, R. Tan, R. P. Wilson, K. Leopold, S. Subramaniam, L. Halabelian, C. H. Arrowsmith, and F. Bénard. "Structure and Activity of Human TMPRSS2 Protease Implicated in SARS-CoV-2 Activation". *Nat Chem Biol* 18.9 (2022), pp. 963–971.

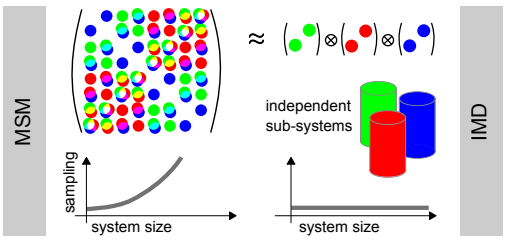
[35] A. Mardt, L. Pasquali, F. Noé, and H. Wu. "Deep Learning Markov and Koopman Models with Physical Constraints". *Proc. First Math. Sci. Mach. Learn. Conf.* Ed. by J. Lu and R. Ward. Vol. 107. Proceedings of Machine Learning Research. Princeton University, Princeton, NJ, USA: PMLR, 2020, pp. 451–475.

[36] M. I. Zimmerman, J. R. Porter, M. D. Ward, S. Singh, N. Vithani, A. Meller, U. L. Mallimadugula, C. E. Kuhn, J. H. Borowsky, R. P. Wiewiora, M. F. D. Hurley, A. M. Harbison, C. A. Fogarty, J. E. Coffland, E. Fadda, V. A. Voelz, J. D. Chodera, and G. R. Bowman. "SARS-CoV-2 Simulations Go Exascale to Predict Dramatic Spike Opening and Cryptic Pockets across the Proteome". *Nat. Chem.* 13.7 (2021), pp. 651–659.

[37] A. Dommer et al. *#COVIDisAirborne: AI-Enabled Multiscale Computational Microscopy of Delta SARS-CoV-2 in a Respiratory Aerosol.* Preprint. Biophysics, 2021. url: http://biorxiv.org/lookup/doi/10.1101/2021.11.12.468428.

[38] N. Trebesch and E. Tajkhorshid. "Embracing Biological Complexity in Atomistic Simulations of Cellular Membranes". *Biophysical Journal* 118.3 (2020), 88a.

[39] R. O. Dror, R. M. Dirks, J. Grossman, H. Xu, and D. E. Shaw. "Biomolecular Simulation: A Computational Microscope for Molecular Biology". *Annu. Rev. Biophys.* 41.1 (2012), pp. 429–452.



Visual summary.

3

Independent Markov Decomposition: Toward Modeling Kinetics of Biomolecular Complexes

This Chapter has been published as

Tim Hempel, Mauricio J. del Razo, Christopher T. Lee, Bryn C. Taylor, Rommie E. Amaro, and Frank Noé. "Independent Markov Decomposition: Toward Modeling Kinetics of Biomolecular Complexes". *Proceedings of the National Academy of Sciences* 118.31 (2021), e2105230118.
https://doi.org/10.1073/pnas.2105230118

Contributions TH was the lead author in this project. He has co-designed the research (with the other authors) and developed the mathematical formalism presented in this manuscript. MJR has helped formalizing the approach and contributed analysis of errors from couplings. TH has designed and implemented benchmark systems (CTL and BCT contributed the Hodgkin-Huxley model benchmark system) and evaluated the new method on them and on the synaptotagmin dataset. TH has created the figures (except Fig. 2 that was contributed by MJR), and was main author of the manuscript. All authors have contributed to writing the manuscript. (This paragraph summarizes TH's contributions alone, it is not an exhaustive list of other authors' contributions.)

Abstract

In order to advance the mission of *in silico* cell biology, modeling the interactions of large and complex biological systems becomes increasingly relevant. The combination of molecular dynamics (MD) simulations and Markov state models (MSMs) has enabled the construction of simplified models of molecular kinetics on long timescales. Despite its success, this approach is inherently limited by the size of the molecular system. With increasing size of macromolecular complexes, the number of independent or weakly coupled subsystems increases, and the number of global system states increases exponentially, making the sampling of all distinct global states unfeasible. In this work, we present a technique called Independent Markov Decomposition (IMD) that leverages weak coupling between subsystems in order to compute a global kinetic model without requiring to sample all combinatorial states of subsystems. We give a theoretical basis for IMD and propose an approach for finding and validating such a decomposition. Using empirical few-state MSMs of ion channel models that are well established in electrophysiology, we demonstrate that IMD models can reproduce experimental conductance measurements with a major reduction in sampling compared with a standard MSM approach. We further show how to find the optimal partition of all-atom protein simulations into weakly coupled subunits.

Significance statement

Molecular simulations of proteins are often interpreted using Markov state models (MSMs), in which each protein configuration is assigned to a global state. As we explore larger and more complex biological systems, the size of this global state space will face a combinatorial explosion, rendering it impossible to gather sufficient sampling data. In this work, we introduce an approach to decompose a system of interest into separable subsystems. We show that MSMs built for each subsystem can be later coupled to reproduce the behaviors of the global system. To aid in the choice of decomposition we also describe a score to quantify its goodness. This decomposition strategy has the promise to enable robust modeling of complex biomolecular systems.

3.1 Introduction

The dynamics of proteins and their functions are of key importance for biology. Molecular dynamics (MD) simulations are a popular method for interrogating the motions of proteins in various environments. A well-known limitation of MD is the timescale

mismatch between simulations and real life. Despite advances in computer hardware and algorithms, extreme timescale simulations remain orders of magnitude shorter than many relevant protein processes. Since one requires sufficient numbers of observations in order to obtain statistical confidence, various strategies have been developed to address this. One approach, building Markov state models (MSM), enables the construction of simple models of long-timescale molecular kinetics from many short off-equilibrium MD simulations [1–6] – see Refs. [7, 8] for thorough reviews. MSMs have successfully been built to obtain compact and yet accurate representations of the kinetics of full proteins [9–16], protein-ligand [17–22] and even protein-protein systems [23].

Although MSMs have significantly helped to reduce the MD sampling problem, the fundamental problem that arises from modeling increasingly large biomolecular systems remains. As protein complexes become larger, the number of uncoupled or weakly coupled subsystems increases. If each of these subsystems contain two or more substates, the number of global system states increases exponentially [24]. Therefore, any model treating the whole system by means of a global state poses requirements on the MD sampling that are fundamentally unscalable. This poses an inevitable problem as evolution tends to lead to increased biological complexity, including the optimization of processes through the formation of protein complexes and puncta [25–28].

In practice, many current models based on MD simulation of large biomolecular systems take the pragmatic approach of ignoring most of the system's dynamics. For example, if one is interested in how an ion channel conducts ions across a membrane, it may be sufficient to prepare the system in a state of interest and collect sufficient statistics of ion passages and perhaps local conformational changes of the selectivity filter residues, rather than trying to sample global conformational rearrangements of the protein complex on much longer timescales. However, our field has a collective interest in developing whole cell and systems modeling for *in silico* medicine, which will necessitate the eventual understanding of these large systems in a way that characterizes how all their components interact, undergo transitions, and can be influenced by, e.g., drug molecules, phosphorylation, and/or glycosylation states.

To this end, Noé & Olsson [24] have recently proposed dynamic graphical models which attempt to decompose protein systems in a way similar to Ising or Potts models – subsystems with states or "spins" that are coupled to one another. Dibak et al [29, 30] have developed a coupling of MSMs with reaction-diffusion dynamics in order to establish an infrastructure in which MSMs can be integrated into whole-cell models. Here we ask a more fundamental question, the answer of which is important to all these

integrative approaches: given a large biomolecular system, how should we decompose it into subsystems, such that these subsystems can be described by independent or weakly coupled MSMs?

Fragmenting proteins at the modeling stage is compatible with prior experience as macromolecules are often sub-divided into structural or functional subunits [31]. There is also evidence that proteins are decomposable into "quasi-independent groups of [spatially adjacent] amino acids" coined "protein sectors" [32]. Furthermore, experimental studies on drug binding or protein functional characterization often use isolated domains or monomers with great success [33].

Estimating an MSM on the decomposed protein can significantly reduce the total sampling necessary. From concepts in statistical physics, given a polymer of length N where each subunit exists in one of k states, the total conformational space is expressed as k^N (cf. Fig. 3.1). Modeling subsystems of a constant size effectively restricts the number of states that need to be sampled reversibly to a constant. Therefore exponentially less sampling is required for modeling smaller subsystems as compared to a global model [15, 24].

In this paper, we develop a mathematical framework of decomposing MSMs into local subsystem MSMs, termed independent Markov decomposition (IMD, Sec. Independent Markov decomposition), and propose a measure of decomposition quality, the *dependency* score (Sec. An MSM score of independence). In the following, we refer to IMD as the process of identifying subsystem MSMs and to an IMD *model* as a model that describes a system as a set of independent, local Markovian subunits.

We speculate that the IMD strategy can forge a new connection to other uses of MSMs such as those employed by the neuronal and cardiac modeling communities. There, phenomenological MSMs parameterized from electrophysiology data are used to predict the behavior of action potentials [34–39]. In Sec. Modeling a tetrameric ion channel using IMD we describe how a decomposed MSM can be connected to a phenomenological MSM. This new connection between fields brings us closer to our goals of understanding these large systems and their behaviors, advancing *in silico* medicine. We further showcase how the *dependency* score can be used to find an optimal partition of a system that does not come with clearly defined independent subunits (Sec. Optimal independent Markov partitions for tetrameric ion channels). We validate our approach with a toy model, showing that the decomposition approximation is high quality and that the proposed validation score works even with limited data (SI Appendix, Toy models). Finally, we demonstrate its applicability to an all-atom MD dataset of the Synaptotagmin-

C2A domain (Sec. Optimal independent Markov partitions for all-atom simulations of Synaptotagmin-C2A) and derive the graph structure of inter-residue *dependencies*.

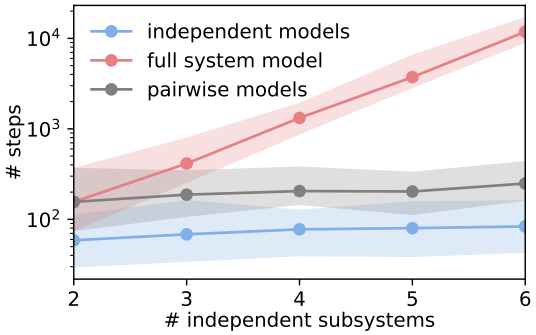


Figure 3.1: Scaling behavior of toy system consisting of n independent subsystems with 3 states each (SI Appendix, Toy models). Number of steps required to reversibly sample all transitions shown for proposed independent models (blue line), full system model (blue line) and pairwise models that are needed for computing the *dependency* score (gray line). Shadowed areas indicate 95% confidence intervals.

3.2 Independent Markov Decomposition

We first describe IMD for discrete-state MSMs before generalizing it to time series with continuous descriptors.

3.2.1 Markov State models

An MSM consists of a discretization of molecular state space into a disjoint set of states $\{S_1, \ldots, S_n\}$ and a Markov chain transition matrix $\mathbf{P}(\tau)$ modeling a memoryless jump process between these states. We can express whether we are in the ith state or not by using indicator functions:

$$\chi_i(\mathbf{x}) = \begin{cases} 1 & \mathbf{x} \in S_i \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

The vector $\chi = [\chi_1, \ldots, \chi_n]^\top$ is thus a "one-hot" (or binary) encoding that maps the continuous state $\mathbf{x}$ to the MSM discretization. For this or any other choice of features χ we can compute the instantaneous and time-lagged correlation matrices $\mathbf{C}_{00} = \sum_t \chi(\mathbf{x}_t)\chi^\top(\mathbf{x}_t)$ and $\mathbf{C}_{0\tau} = \sum_t \chi(\mathbf{x}_t)\chi^\top(\mathbf{x}_{t+\tau})$, respectively. For a fixed state discretization, the transition matrix that has maximum likelihood and also maximizes the variational approach of conformation dynamics (VAC) [40] is:

$$\mathbf{P}(\tau) = \mathbf{C}_{00}^{-1}\mathbf{C}_{0\tau}. \tag{3.2}$$

Let $\mathbf{p}_t$ denote the probability distribution of being in any of the n states at time t, for example $\mathbf{p}_0 = [1, 0, ..., 0]$ denotes that the system starts in state 0 at time 0. This vector can be evolved in time using the transition matrix, until it converges to the equilibrium distribution $\boldsymbol{\pi} = \lim_{t\to\infty} \mathbf{p}_t$:

$$\mathbf{p}_{t+\tau}^\top = \mathbf{p}_t^\top \mathbf{P}(\tau). \tag{3.3}$$

An important concept for optimizing the parameters or hyperparameters of MSMs and other Markovian kinetic models is the variational approach for Markov processes (VAMP) [41]. VAMP finds that a Markovian model that best approximates the high-dimensional continuous dynamics maximizes the VAMP-n score:

$$R_n(\mathbf{P}) = \left\| \mathbf{C}_{00}^{-1/2} \mathbf{C}_{0\tau} \mathbf{C}_{\tau\tau}^{-1/2} \right\|_n^n \tag{3.4}$$

where we can either use $n = 1$ for the trace norm or $n = 2$ for the Frobenius norm. If we run molecular dynamics at equilibrium conditions, we can employ correlation matrix estimators that provide $\mathbf{C}_{00} = \mathbf{C}_{\tau\tau}$ and symmetric $\mathbf{C}_{0\tau}$ (detailed balance). In this special case, VAMP becomes the VAC mentioned above, and the variational score simply becomes $R_n(\mathbf{P}) = \|\mathbf{P}(\tau)\|_n^n$. In other words, the optimal MSM is the one that maximizes the trace or the Frobenius norm of the transition matrix, which is equivalent to maximizing its eigenvalues. Since the eigenvalues equal the normalized time-autocorrelation of the slowest processes [1, 42], the VAC tries to find the Markovian model that best resolves the slowest processes of the molecular process under investigation [40, 43]. For a fixed state space discretization, optimizing the VAC results in the MSM estimator (3.2). If we also want to search over different state space discretizations, we can use VAC or VAMP as a score in a hyperparameter optimization problem [44] or optimize the VAMP score while representing χ with deep neural networks, leading us to VAMPnets [45].

3.2.2 Independent Markov decomposition

Now we move beyond the common concept of modeling the dynamics of the entire molecular system by a single MSM and instead try to decompose the system into almost independent MSMs. Let us start with the simple example shown in Fig. 3.2a, where a molecule consists of two domains, A and B, that are each described by a two-state MSM describing whether the domain is "closed" ($\alpha, \beta = 0°$) or "open" ($\alpha, \beta = 90°$). We assume that the kinetics of both domains are statistically independent, i.e. each domain switches states independent of the states of the other one – we simultaneously have $\mathbf{p}_{A,t+\tau} = \mathbf{P}_A(\tau)\mathbf{p}_{A,t}$ and $\mathbf{p}_{B,t+\tau} = \mathbf{P}_B(\tau)\mathbf{p}_{B,t}$ (Fig. 3.2b). As the MSMs A and B are statistically independent, the probability distribution of the entire system follows Eq. (3.3)

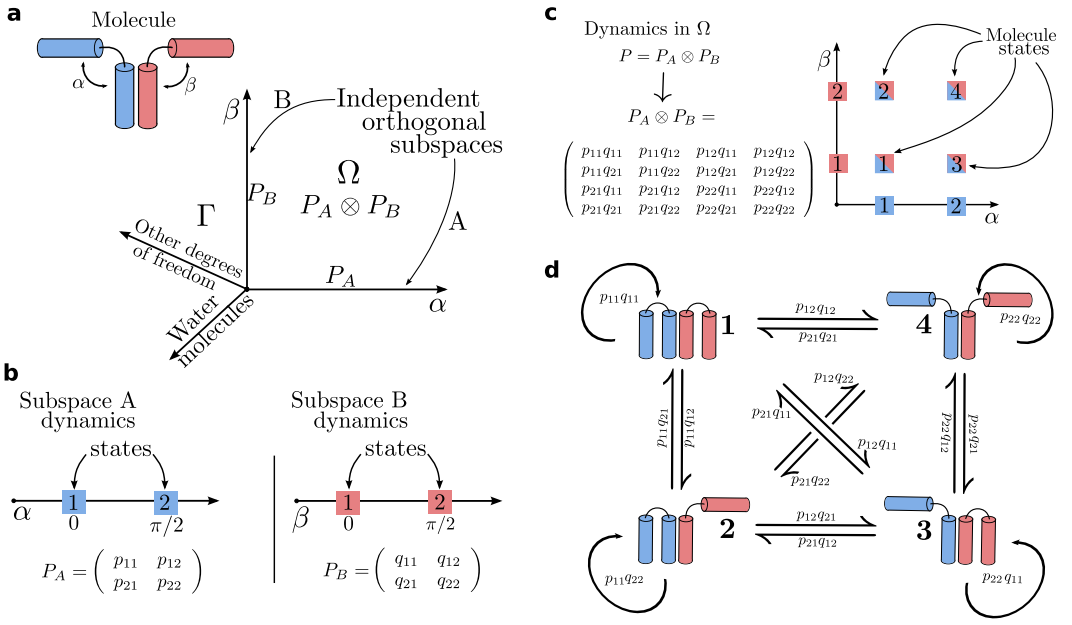


Figure 3.2: Operator decomposition and discretization on a test molecule. **(a)** A test molecule is decomposed into two subsystems (blue and red). The two angles α and β span subspaces A and B corresponding to the two subsystems, respectively. The space Γ is composed of all system degrees of freedom. The space Ω is the Cartesian product of A and B and its dynamics are described by Perron-Frobenius operators P_A and P_B, respectively. The dynamics in Ω are given as the tensor product $P_A \otimes P_B$. **(b)** The molecule has metastable states at $\alpha = 0, \pi/2$ and $\beta = 0, \pi/2$; the subspaces A and B can be discretized into MSMs with transition probability matrices $\mathbf{P}_A$ and $\mathbf{P}_B$. The quantities p_{ij} and q_{ij} are the transition probabilities from state i to j of subspaces A and B, respectively. **(c)** The discretized dynamics in Ω are given by the tensor product $\mathbf{P}_A \otimes \mathbf{P}_B$, yielding the four states of the full molecule. **(d)** Illustration of the four possible states of the molecule and the transitions between them.

with

$$\begin{aligned} \mathbf{p}_t &= \mathbf{p}_{A,t} \otimes \mathbf{p}_{B,t} \\ \mathbf{P}(\tau) &= \mathbf{P}_A(\tau) \otimes \mathbf{P}_B(\tau), \end{aligned} \tag{3.5}$$

where $\otimes$ is the Kronecker product [46] (see SI Appendix, Markov operators). The vector $\mathbf{p}_t$ now contains the probabilities of being in the four combinatorial states (A and B open, A open and B closed, A closed and B open, A and B closed), and $\mathbf{P}(\tau)$ is the 4×4 transition matrix between these combinatorial states whose transition probabilities are simply products of the individual transition events in subsystems A and B (Fig. 3.2c, d). The power of this approach is apparent when comparing Figures 3.2b and c: If the dynamics in A and B are independent or almost independent, we can estimate the sixteen transition probabilities that parametrize the whole system using only the eight elements of the transition matrices of the subspaces. This advantage increases expo-

nentially in larger systems: if we have N (almost) independent domains with m states each, distinguishing all states would require us to sample and estimate an exponential number of order of m^{2N} transitions, whereas a decomposition into independent MSMs reduces this to a polynomial number of Nm^2 transitions that can be scaled to large systems. From another point of view, IMD is more efficient because it obtains a greater number of "effective" transition counts for the global model by applying the Kronecker product (cf. SI Appendix, Effective counts and sampling). The above example trivially generalizes to N systems with $\mathbf{P}(\tau) = \bigotimes_I^N \mathbf{P}_I(\tau)$. We note that it is customary to dismiss variables of the full state space Γ (Fig. 3.2a) that are assumed to average quickly, e.g., solvent degrees of freedom. Thus the modeled space Ω in practice only encompasses the variables of interest, e.g., internal coordinates of a protein system.

3.2.3 An MSM score of independence

In practice, subdomains of biomolecules or biomolecular complexes will not be exactly independent. Moreover, the identification of a domain decomposition into almost independent subdomains is a non-trivial task. To enable algorithmic determination of almost independent subdomains, we develop an independence score that quantifies decomposition validity. To this end we come back to the variational approach Eq. (3.4). Conveniently, matrix norms follow simple rules when applied to a Kronecker product (SI Appendix, VAMP score decomposition). In practice, we will apply the trace and Frobenius norms that correspond to the VAMP-1 and VAMP-2 scores of the Koopman operator. The VAMP-2 score has successfully been used in many practical applications [16, 45, 47, 48]. If our molecular system consists of N independent subdomains such that its global MSM is a Kronecker product of N subspace MSMs as described above, its VAMP score is the simple product of VAMP scores (SI Appendix, VAMP score decomposition):

$$R_n(\mathbf{P}) = \prod_{I=1}^{N} R_n(\mathbf{P}_I). \tag{3.6}$$

Here, $R_n(\cdot)$ denotes the VAMP-n score of the transition operator. It could be the trace norm (VAMP-1) or Frobenius norm (VAMP-2) of the associated transition matrix. In practical applications, the VAMP-n score could be rank-reduced, i.e., restricted to the highest $k < m$ singular values. Note that Eq. (3.6) is a necessary but not a sufficient condition for Markov independence. Significant deviations from equality in Eq. (3.6) indicate that the assumption of independence is invalid. However, if separate MSMs $\mathbf{P}_I$ can probe the same molecular features, it is possible to satisfy Eq. (3.6) even though the

subsystem MSMs are not statistically independent. Eq. (3.6) must therefore always be used in conjunction with appropriate constraints. Here, we choose between different ways to assign independent molecular features to different MSMs and check which of these assignments best satisfies Eq. (3.6). In practice, we want to estimate an IMD model because often we cannot compute the global MSM $\mathbf{P}$ due to limited sampling (Fig. 3.1), and we consequently do not know $R_n(\mathbf{P})$. Therefore, we choose to only check the equality of Eq. (3.6) on pairs of subsystems A, B, i.e., $R_n(\mathbf{P}_{A,B}) = R_n(\mathbf{P}_A) \cdot R_n(\mathbf{P}_B)$. We then search over possible partitions of the molecular system into subsystems by evaluating the graph of pairwise *dependencies* $d(A, B)$:

$$d(A, B) = \left| R_n(\mathbf{P}_{A,B}) - R_n(\mathbf{P}_A) \cdot R_n(\mathbf{P}_B) \right| \tag{3.7}$$

In practice, computing $\mathbf{P}_{A,B}$ involves a new estimate of the transition probability matrix in the joint space of two systems. We show that our measure scales well with respect to limited sampling (also compare SI Appendix, Toy models).

The product in Eqs. (3.6) and (3.7) is purely a result of the chosen basis set of MSMs (Eq. (3.1), SI Appendix, VAMP score decomposition). In practical situations, it is desirable to find a decomposition directly based on molecular features such as distances or contacts instead of performing an MSM discretization and estimation for each subsystem. When considering more general features χ, there are two main changes to discrete-state MSMs: (i) observables are propagated by a different operator, called Koopman operator [49, 50], (ii) the joint space of observables is most easiestly described by "stacking" observable feature vectors rather than by defining an MSM discretization on the combinatorial space. For example, if $\mathbf{\Psi}_A = (\psi_A^1, \psi_A^2, \ldots)$ and $\mathbf{\Psi}_B = (\psi_B^1, \psi_B^2, \ldots)$ are the one-dimensional time series of features $\psi \in \mathbb{R}$ of two systems A and B, the joint space would be spanned by $\mathbf{\Psi}_{AB} = ((\psi_A^1, \psi_B^1), (\psi_A^2, \psi_B^2), \ldots)$. The transfer operator describing the independent dynamics in joint space is thus a block matrix of its constituting independent sub-operators (also called a direct sum, see SI Appendix, Markov operators for details). This also means that independent subsystem features are not correlated. We note that "stacking" in the MSM formulation would produce probability vectors not normalized to 1 and yield invalid (i.e., not irreducible) MSM transition matrices in the joint space. The trace and Frobenius norm of the Koopman operator thus decompose as sums such that the dependency score reads

$$d(A, B) = \left| R_n(\mathbf{K}_A) + R_n(\mathbf{K}_B) - R_n(\mathbf{K}_{A,B}) \right| \tag{3.8}$$

where **K**, the Koopman operator, takes the place of the transition matrix **P**. See SI Appendix, VAMP score decomposition for the derivation. We note that even though discussing MSM artifacts is out of scope for this work, it is unclear how possible discretization errors might propagate to the MSM-based *dependency* (Eq. (3.7)). However, such artifacts are entirely ruled out when working in observable space (Eq. (3.8)).

3.3 Results

3.3.1 Modeling a tetrameric ion channel using IMD

In cardiac electrophysiology, Markov models have been used to model phenomenological data from ion channels [37–39]. Ion channels are transmembrane proteins that respond to physiological stimuli and selectively control the flow of ions in excitable cells. Upon a change in membrane potential, voltage-gated ion channels undergo conformational changes that modulate ionic conductance. The symphony of ion channels collectively facilitate the propagation of electrical signals in excitable tissues, such as the heart and brain, and are important drug targets [51, 52]. The plethora of experimental measurements of ion channel properties sets the stage for computational simulations to provide molecular details and mechanistic insights [53]. Though possible to fit a phenomenological MSM using data from electrophysiological experiments, atomistic modeling remains out of reach due to the long timescales of channel opening. This is because single gate activation events are rare, and many ion channels have multiple gates which need to activate concurrently. Reversible sampling will further be hampered by a combinatorial number of pathways that lead to a fully open channel. We propose that for cases of non-cooperative gates, IMD can help solve this problem, which we demonstrate in the following series of numerical experiments. We consider a voltage-gated tetrameric potassium ion channel with four identical subunits, each with a voltage sensor. To construct an IMD model, we exploit the independence of individual subunits or gates and partition accordingly (Fig. 3.3a1). This produces four matrices $\mathbf{P}_i \in \mathbb{R}^{2\times 2}, 1 \leq i \leq 4$ that describe individual gate opening and closing. As derived above, the Kronecker product of subsystem transition matrices yields a transition matrix $\mathbf{P} \in \mathbb{R}^{16\times 16}$ of the full ion channel (Fig. 3.3a2). The 16 states enumerate all possible combinations of open and closed gates of the full ion channel, a state space referred to as $\tilde{S}$ in the following. We note that this decomposition is only possible between non-cooperative domains.

We construct a mapping to assign the 16 states of the transition matrix **P** to those of a phenomenological MSM. Our reference empirical model is the one developed in Ref. [54] for this channel (Fig. 3.3b). In Ref. [54], channel symmetry is used to define

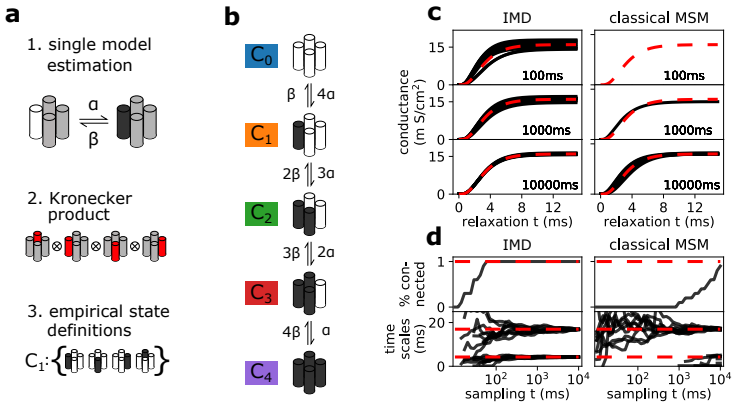


Figure 3.3: Reconstructing the Hodgkin-Huxley model from a simple discrete model. **(a)** Pipeline of steps required to assemble a full channel model from a single subunit model that describes opening and closing of a single subunit in the vicinity of the others (step 1). Kronecker product between all four sub-unit models assembles a model that still distinguishes between all combinatorial states (step 2). Empirical state definitions account for channel symmetries (step 3). Black denotes open, white closed, and gray undefined subunit. **(b)** Graphical depiction of full channel model in empirical state space. Note the symmetry of the channel, i.e. that at this stage only the number of open subsystems is known. **(c)** Relaxation from a closed state into the native state at 63 mV. We show conductance predicted by IMD model (left column) and classical MSM (right column) using different amounts of sampling. Note that the classical approach only yields results in the high sampling regime where all empirical states are connected. Results are compared to the original Hodgkin-Huxley model (red dashed line). **(d)** Sampling time necessary to estimate a decomposed MSM (left column) compared to a classical full system MSM (right column) for ten realizations of the Markov chain. We show the percentage of fully connected models in our ensemble of realizations (top row) and the 1st and 4th implied timescale computed from it (bottom row). Note that for the classical MSM, extreme amounts of sampling are necessary to even estimate all system-inherent implied timescales.

the full system states accordingly:

$$S = \begin{cases} C_0 & \text{all gates closed} \\ C_1 & \text{1 gate open} \\ C_2 & \text{2 gates open} \\ C_3 & \text{3 gates open} \\ C_4 & \text{all gates open} \end{cases}$$

Mapping of the transition matrix into the space of these empirical states can be obtained by converting the empirical state definitions into crisp membership vectors $\chi_s \in \{0,1\}^5$, with each element indicating which empirical configuration a full system configuration $s \in \tilde{S}$ belongs to. For example, the membership vector describing any state s_k with one open gate would be $\chi_{s_k} = (0,1,0,0,0)$, i.e., these states are associated to macroconfiguration C_1. The full membership matrix is constructed by stacking $\chi = [\chi_{s_1}, \chi_{s_2}, \ldots \chi_{s_{16}}] \in \{0,1\}^{5\times 16}$. Subsequently, the transition-matrix is coarse-

grained following [55, 56] $\mathbf{P}_{\text{empirical}} = \mathbf{\Pi}_c^{-1}\boldsymbol{\chi}^\top\mathbf{\Pi}\mathbf{P}\boldsymbol{\chi} \in \mathbb{R}^{5\times5}$ with $\mathbf{\Pi} = \text{diag}(\boldsymbol{\pi})$ the diagonal matrix of the stationary distribution π in full space and in empirical space $\mathbf{\Pi}_c = \text{diag}(\boldsymbol{\chi}^\top\boldsymbol{\pi})$.

Choosing rates α and β from the original work by Hodgkin-Huxley [34] at a voltage of 63 mV, we produce a simple discrete model. Using this model, we can generate sample trajectories from which to construct MSMs in accordance with Sec. Computational experiments. We estimate a model for the full system from this data by applying the aforementioned pipeline. Using this derived full system model, experimental observables from electrophysiology experiments can be assessed by relaxation of the Markov chain from a non-equilibrium distribution (e.g., a closed configuration) into the equilibrium at this particular voltage [57, 58]. We start from a configuration of fully closed states and further assume that the channel only conducts ions if it is open, i.e., our observable is only non-zero for the open state. This experiment is the computational analog to a voltage jump experiment from resting to +63mV in voltage clamp mode. Shown in Fig. 3.3c, the modeled conductance of the channel over time is reported. The predicted conductance time-series is compared with the numerically integrated ordinary differential equation for the potassium ion channel derived by Hodgkin and Huxley [34]. We find that the IMD model can accurately reproduce the full channel dynamics. IMD models were built by separately fitting four single gate trajectories (i.e. a full system trajectory split into its subsystems) and assembled using the aforementioned steps. For comparison, traditional MSMs were fit to sample trajectories computed from the full system transition matrix in its empirical state definition. We note that we compare the sampling necessary for IMD models to the empirical 5-state formulation (which does not resolve all 16 combinatorial states). In this way, we can rule out that the described sampling advantages of IMD are an artifact of exploited channel symmetry. The reduction in the amount of sampling needed due to the use of IMD can be quantified in terms of the length of simulation required to form a fully connected transition matrix. In Fig. 3.3d we present the percentage of connected IMD models estimated on an ensemble of ten realizations of the Markov chain and compare to a classical MSM. Note that even though a necessary condition for MSM estimation, connectivity is not a quality criterion - we discuss approximation quality below. Connectivity is computed as a function of simulated time (in ms), i.e., shows how probable a modeler can estimate a connected Markov model, IMD or classical, from a fixed amount of sampling. We note that the classical MSM approach can only estimate all system-inherent implied timescales when all empirical states are reversibly sampled, i.e., only for very large amounts of data. In terms of model approximation quality, the higher computational efficiency

of IMD is evident from the much faster convergence of implied timescales as a function of simulation length (Fig. 3.3d, also note root-mean-square error between estimated and ground truth eigenvalue spectra in SI Appendix, Fig. A.4). We find a reduction in sampling by three orders of magnitude, from tens of seconds to tens of milliseconds (Fig. 3.3d). For example, ionic conductance is reasonably approximated with 100 ms of sampling and the IMD approach (3.3c).

Here, we have presented an example where each gate operates independently. In practice, the gating behaviors of most ion channels are not completely independent, but are instead coupled. In this case, the decomposition yields an approximate model of the real dynamics, see SI Appendix, Weakly coupled systems for a discussion. The theoretical limit is posed by the assumption of stationarity that underlies MSM estimation. It is violated if external influences are strong and on similar timescales as the processes to be modeled. External influences that are much faster than the local dynamics are incorporated as an average over Markov states, similar to water molecules in regular MSMs. As demonstrated in the SI Appendix, Fig. A.1, modeling of weakly coupled systems is possible in a robust fashion.

3.3.2 Optimal independent Markov partitions for tetrameric ion channels

For our previous example, we prescribed a convenient partitioning scheme for the ion channel system. In contrast, in real-world situations a complex system may involve multiple independent subsystems but the coupling graph is unknown *a priori*. For instance, it might not be clear how to find independent protein segments of an unknown protein. A method is necessary to aid in the discovery of viable partitions which produce independent subsystems. In this section we demonstrate how the *dependency* defined in Sec. An MSM score of independence can be used as a score to bisect clusters of coupled subsystems from weakly coupled ones. The idea is to compute all possible pairwise *dependencies* between all subsystems and to use them as edge weights in a graph. If they exist, (almost) independent clusters of strongly coupled subsystems will be revealed by analyzing this graph. Once identified, these clusters might be modeled with single subsystem transition matrices within the IMD framework. For the purposes of demonstration, we zoom out from a single channel protein to a membrane patch (Fig. 3.4a). In our setup, this patch contains a dimer of channels which we model to be coupled by a weak, cooperative coupling. Individual channels are modeled using the same parameters as the above ion channel model but contain the additional element of an external deactivation switch (Fig. 3.4b). In a cellular environment, such a switch could, for example, be an inhibitory ligand that binds and unbinds at a certain rate. It is modeled

as a Markov process with probability 0.01 to change its state. The deactivation switch alters the conformational dynamics of each gate such that the probability to close or to stay closed is 95%. Thus, by construction, it is not possible to decompose a channel MSM into single gate MSMs because each gate is now coupled to the deactivation switch. Further, the strength of the intra-channel coupling can be controlled by a linear mixture parameter λ. The dynamics described above correspond to $\lambda = 1$, strong coupling. The coupling can be entirely deactivated by setting $\lambda = 0$. See SI Appendix, Dimer model for implementation details.

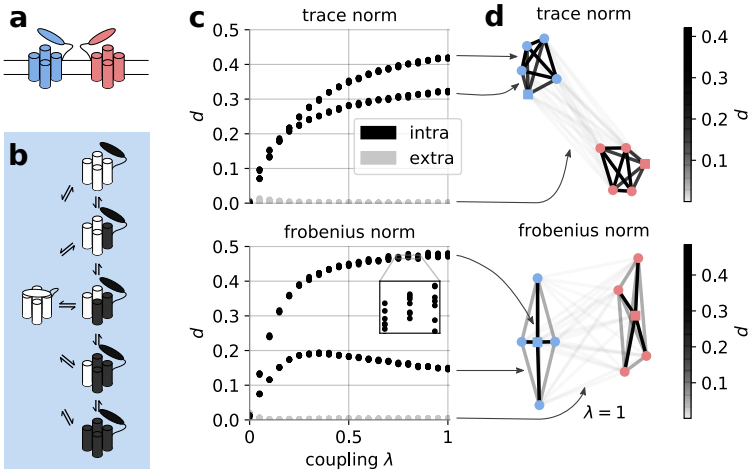


Figure 3.4: Visualization of channel dimer. **(a)** Two channels located in a membrane. Each channel consists of four gates (akin to Hodgkin-Huxley model, depicted by cylinders) and one desensitization switch (depicted as an additional oval domain). **(b)** States and possible transitions of individual channels (simplified, short lived switch-deactivated open states are ommitted in this figure). As both channels have the same dynamics, only one is shown as an example. **(c)** Dependency score as a function of coupling strength as defined by linear mixture parameter λ. Color code: Grey denotes scores between two molecules, black intra-channel pairs. **(d)** Graph of pairwise dependencies between all channel subunits for $\lambda = 1$. Edges are color coded according to *dependency* scores between two systems. Nodes belonging to a single channel are color-coded accordingly, squared nodes represent deactivation switches.

We generate discrete time series data from a transition matrix that models a dimer with these properties (SI Appendix, Dimer model and Computational experiments). From the data, the *dependency d* is computed for all possible pairs of subsystems. This involves the estimation of transition matrices for two isolated subsystems and comparing them with the transition matrix estimated in the joint space using Eq. (3.7). For example, one such pair could be the deactivation switch of one channel and a gate of the other channel. A natural representation of these pairwise norms between subsystems is a graph. It is formed by nodes (subsystems) and *dependency*-weighted edges; no assumption about its structure is made (e.g., that it is a fully connected graph). For

the numerical experiment described in this section, our analysis yields the graph shown in Fig. 3.4d. The graph is visualized by positioning the subsystems or graph nodes with the Fruchterman-Reingold algorithm [59, 60] which is sensitive to the edge weights. This means that subsystems with high *dependency* are grouped together. This helps us to visually identify clusters of coupled subsystems. Groups of subsystems that are far apart in this representation are coupled relatively weakly. We find that *dependencies* between subsystems of the same channel are significantly larger than zero while inter-channel interactions yield *dependencies* close to zero (see Fig. 3.4d). Further, reducing the coupling strength within a channel does not alter our qualitative results (Fig. 3.4c). The observed bifurcation of *dependencies* is due to the two types of coupling in the system (gate-gate vs. gate-deactivation switch) and is a feature of the dimer model system. In summary, our results show that we can learn the connectivity of a network of subsystems from discrete, simulated time series data. In particular, the *dependency* score provides an approach to find an optimal partition of a system with multiple types of coupling.

3.3.3 Optimal independent Markov partitions for all-atom simulations of Synaptotagmin-C2A

To showcase the applicability of the *dependency* score, we apply our method to a 180 µs molecular dynamics data set of the C2A domain of Synaptotagmin-1 (Syt). Syt is a crucial player in the neurotransmitter release machinery [61]. In our previous study we have found that single loops of its C2A domain can be described independently of each other using a hand crafted partition [15]. Here, we attempt to find an optimal partition by using the *dependency* score at the residue resolution (Sec. Application to MD dataset). Instead of working with MSM transition probabilities, we directly work in protein feature space in order to omit discretization artifacts. We find that indeed, Syt-C2A can be partitioned into defined subunits, or conformational switches, using a VAMP-2 based *dependency* score (Fig. 3.5). The *dependency* network spanned by Syt-C2A residues expresses defined subsystem clusters. Within each subsystem cluster, residues are embedded with high normalized *dependency* scores whereas between different subsystem clusters, these links are weaker (Fig. 3.5a). The boundary between what is considered a high and a low normalized *dependency* tends to be ~ 0.6, we however note that this value might be system-specific. The discovered partition contains the conformational switches defined in our last study [15]: In particular, the C78 switch (Fig. 3.5b, red) emerges as an independent cluster in the Fruchterman-Reingold projection, confirming our previous results. However, even though conformational switches in the

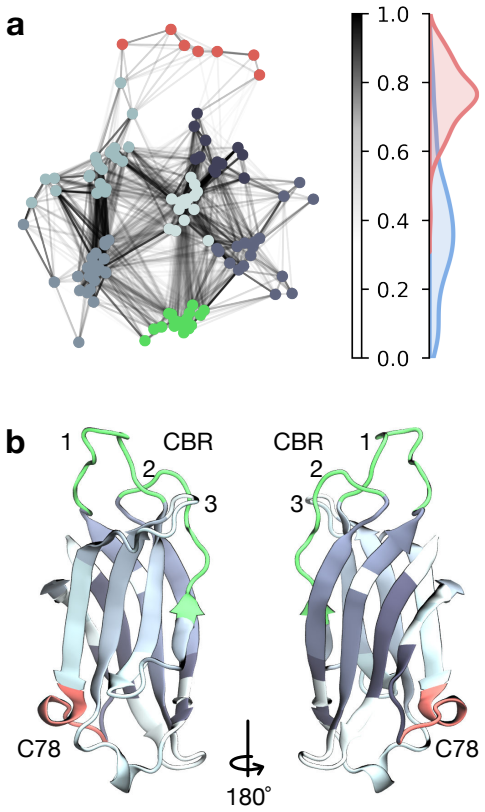


Figure 3.5: Dependency-network between residues of Syt-1 C2A depicted using a standard graph layout (Fruchterman-Reingold algorithm). **(a)** VAMP-2 normalized dependency network. Edge weights are indicated by colorbar. Nodes are colored according to an unsupervised classification by the k-means algorithm ($k = 7$). *Dependency* histograms depict coupling strength of residues within a subsystem cluster (red) and between different subsystem clusters (blue). **(b)** Visualization of protein structure with color coded segments from our VAMP-2 analysis (colors correspond to classification in panel a). VAMP-1 yields similar results (not shown here, see SI Appendix, Fig. A.2).

Calcium Binding Region (CBR), CBR-1 and 2 together (Fig. 3.5b, green), are connected to the other protein residues by a low dependency, describing these loops independently is an approximation that is only partially backed by this current study. Similar results are obtained when using a VAMP-1 based *dependency* (SI Appendix, Fig. A.2).

3.4 Discussion

Over the past several decades, MSM methodology has matured into a valuable tool for MD data analysis [1, 3, 4, 7, 8, 13, 20–23, 42]. For practitioners, modeling MD data with MSMs remains a non-trivial task, especially as researchers turn their focus towards the study of progressively larger biomolecular complexes. Larger systems generally come

with an increasing number of (metastable) states that demand vast amounts of sampling time and hamper attempts to rigorously model protein dynamics. In these scenarios, the classical MSM method reaches a point where the combinatorial explosion of states becomes a critical bottleneck. It is a fundamental problem that is inherent to any method which seeks to describe the global protein state [24]. One possible solution is to appreciate the notion of independent protein segments [32] and to split large systems into smaller, more manageable subsystems. In this spirit, we have proposed Independent Markov Decomposition. For practitioners, this means that, for example, an ion channel is modeled as a set of individual gates as opposed to a single protein. This approach approximates the system as a set of independent subsystems and is naturally agnostic to global system size. In this paper we have shown how the conceptual idea of IMD relates to the underlying transfer operator formulation, what sampling advantages can be expected, and how to use the proposed *dependency* score to find an optimal partition of an unknown system. Using the tetrameric potassium ion channel as a model system, we show that we can estimate a fully converged model with approximately three orders of magnitude less sampling when compared to a classical MSM. IMD therefore has the potential to leverage sampling efforts for large biological systems into a regime that is achievable with state-of-the-art simulation techniques and computer hardware. This effect is due to data being used more efficiently while small compromises are made by a mean-field-like approximation. For systems with potentially weak couplings, the validity of the approximation can be checked with our *dependency* score *a posteriori*. We further posit that due to the tremendous sampling advantages, the estimation errors introduced by weak couplings are likely to be smaller than the sampling error for classical global state MSMs. Our results suggest that IMD improves the assessment of sampling convergence for large systems. As real-world MD datasets are usually very high dimensional, in practice, it is a non-trivial task to assess whether the sampling is converged. Often, researchers can only speculate by using semi-empirical tests, i.e., matching of high-level experimental observables to model predictions. IMD offers a more rigorous way to tackle this problem. For example, when modeling a single protein loop, it is much easier to see if the process is sampled reversibly, a question that can be difficult to answer with a classical MSM on global states.

Furthermore, we have proposed a *dependency* score that quantifies the coupling between two subsystems. As there is no general rule how to define protein subsystems, the *dependency* score serves as an objective function to judge IMD model approximation quality and to find an optimal partition of unknown systems. In a numerical test system of a switched dimer model with weak cooperative coupling, the *dependency*

score has robustly bisected clusters of strongly coupled subsystems from weakly coupled ones. It thus enabled IMD model estimation without knowing the dependency graph structure *a priori*. In order to optimally partition a system in practical applications, a sufficiently large biomolecular system could be first partitioned into minimal subsystems such as residue side-chains. Scoring the dependency between these subsystems can reveal the structure of the dependency graph and thus give rise to a definition of (almost) independent protein segments. We note that IMD is designed for systems with time-constant, independent subunits, i.e., it is most probably not suitable for few-residue peptides or protein folding (for a counterexample using Chignolin [62], cf. SI Appendix, Fig. A.3). We have shown that for the C2A domain of Synaptotagmin-1, the *dependency* score can be used to identify clusters of subsystems that are linked relatively weakly between each other. These subsystems are similar to the conformational switches identified and independently modeled in Ref. [15]. We however note that the current, prototypical implementation of assigning residues to subsystems is subject to stochasticity. For future work, in particular for larger biomolecular complexes, it will be desirable to incorporate experimental knowledge about size and properties of "protein sectors" [32]. An aspect excluded in this conceptual study is the discretization of MD data, a step which can be crucial in practical MSM applications [4, 63]. We note that subsystem MSMs have smaller dimensionality and therefore discretization errors are smaller compared to the higher-dimensional full system. This implies that IMD models may reduce discretization artifacts compared to classical MSMs. However, further work should consider the implications of the discretization error as it is unclear how it propagates to joint space probability estimates and *dependency* score. Furthermore, the lag time τ has two-fold implications on IMD: First, when estimating local, independent subunit MSMs, the choice of lag time must be verified for each independent MSM as for classical MSMs (e.g., by implied timescales test). This might yield different lag times for different subunits, which is justified when working with independent models alone. However, if a global (or pairwise) model is desired, all constituting local models must strictly have the same lag time such that a global (or pairwise) operator is defined. This, second, is the reason why the *dependency* score can only be applied for a single global lag time. In practice, choosing a lag time for *dependency* network estimation might therefore be done as common practice with, e.g., time-lagged independent component analysis (TICA) analyses [63], i.e., starting with a lag time that most likely yields converged estimates. This choice should be validated by ensuring subsystem implied timescales convergence.

In this work, we propose that one way to keep pace with our interest in modeling large biological systems is by using a decomposition technique. For large systems, IMD models are more data efficient and might be easier to apply than classical global state MSMs. We believe that interrogating local features, e.g., ligand binding pockets, instead of global system states can be more informative and give better predictions at reduced computational cost. Because this approach comes with all the established methods and software of the MD MSM community, we anticipate that IMD will have a broad application basis for *in-silico* cell biology.

3.5 Methods

3.5.1 Computational experiments

Gate opening and closing rates of the toy potassium ion channel were obtained from the Hodgkin-Huxley model. Under voltage clamp conditions and neglecting the sodium and leak currents, we are left with the potassium ion channel contribution. The current is given as follows,

$$I_K = G_k(V_m - V_K) = \overline{g}_K n^4 (V_m - V_K),$$

where I_K is the current, G_K is the conductance, $\overline{g}_K$ is the maximal conductance, V_m and V_K are the total transmembrane potential and potassium ion reversal potential respectively. Here $n \in [0, 1]$ is a dimensionless quantity corresponding to channel activation. The time dependence of n is described using the following ordinary differential equation (ODE),

$$\frac{dn}{dt} = \alpha_n(V_m)(1 - n) - \beta_n(V_m)n,$$

where α_n and β_n are the kinetic rates (s^{-1}) of activation and deactivation respectively. In the original Hodgkin-Huxley model [34], the voltage sensitivity of the ion channel is modeled by the voltage dependence of the rates α_n and β_n,

$$\alpha_n(V_m) = \frac{0.01(10 - V_m)}{\exp\left(\frac{10 - V_m}{10}\right) - 1},$$
$$\beta_n(V_m) = 0.125 \exp\left(\frac{-V_m}{80}\right).$$

The term n^4 is the joint probability that the four independent subunits of the tetrameric potassium ion channel are concomitantly open. Thus α_n and β_n are the kinetic rates for an individual subunit to open and close respectively. This set of ODEs were integrated using the `odeint` function provided by scipy [64] to serve as the ground

truth for later comparison with IMD model and MSM results. We apply our framework to discrete time series data with known full system dynamics. For each system that we are using, details and generator matrix are given in the SI Appendix, Toy models and Dimer model. Generally, a transition matrix describing a (full) test system (possibly including couplings) is chosen, akin to $\mathbf{P}(\tau)$ in Eq. (3.5). Time series are generated using the Markov chain sampler implemented in pyEMMA/msmtools [65]. Subsequently, full system states are mapped to individual subsystem states, yielding subsystem trajectories which are parallel in time. Estimation of subsystem transition matrices ($\mathbf{P}_i(\tau)$ in Eq. (3.5)) is followed by assembly of a full system transition matrix. The latter is utilized to extract full system observables such as implied timescales.

3.5.2 Application to MD dataset

The protocol that was used to obtain MD simulation data and featurization of Syt-C2A is described in detail in Ref. [15]. In particular, as in the cited study, we use heavy atom coordinates of the superposed protein. We are aware that this could potentially yield spurious correlations, however a) no better descriptor of the slow dynamics could be found and b) we want to ensure compatibility to our previous study. Each residue is encoded as a vector of flattened coordinates Y_i and the *dependency* is computed on each pair of residues. The pairwise features are the stacked vectors $[Y_i, Y_j]$. Note that when directly working on coordinate features, unlike in the MSM examples, the *dependency* decomposes as a sum, not as a product (SI Appendix, VAMP score decomposition). Furthermore, the dependency is normalized to untangle the amount of kinetic variance from actual dependency, i.e.

$$d = \frac{|R_n(A) + R_n(B) - R_n(A,B)|}{\min(R_n(A), R_n(B))} \in [0,1] \tag{3.9}$$

with $R_n(x)$ being the VAMP-n-score of residue x. Note that in the case of high *dependency* scores, the two observable features might be proxies of the same process, however one of them could encode an additional one. Dividing by the min ensures we are only normalizing to the processes contained in both subsystem vectors. To not obfuscate the histogram analysis conducted for the *dependency* score network with weak links in otherwise strongly coupled clusters, we have taken into account only the strongest link connecting each residue. We thus extract the maximal normalized *dependency* score that connects a given residue to all other residues within a subsystem cluster (intra-subsystem) or to all residues of a different subsystem cluster (inter-subsystem), respectively. The VAMP-n-scores for Syt-C2A are computed with PyEMMA [65] at a lag time

of 50 ns. The lag time was chosen based on implied timescales convergence reported in Ref. [15].

Data availability

The code that implements our discrete models, generates the data, and reproduces the presented results can be found in our GitHub repository [66]. The molecular dynamics data set of Synaptotagmin C2A is available upon request. Some study data are available upon request.

Acknowledgements

TH thanks Moritz Hoffmann and Andreas Mardt (FU Berlin) for fruitful discussions. We acknowledge funding from Deutsche Forschungsgemeinschaft (SFB/TRR 186, Project A12; SFB1114, Project A04), the Berlin Mathematics Research Center MATH+ (AA1-6), the Bundesministerium für Bildung und Forschung, and the European Commission (ERC CoG 772230 "ScaleCell"). M.J.R. acknowledges support by the Dutch Institute for Emergent Phenomena at the University of Amsterdam. BCT and REA acknowledge support from the National Biomedical Computation Resource via NIH Grant P41-GM103426. CTL acknowledges support from a Hartwell Foundation Postdoctoral Fellowship. REA acknowledges funding from NIH R01 GM132826.

Author contributions

T.H., B.C.T., M.J.R., C.T.L., R.E.A., F.N. designed research; T.H., B.C.T., M.J.R., C.T.L. performed research; T.H., B.C.T., C.T.L., F.N. analyzed data; T.H., B.C.T., M.J.R., C.T.L., R.E.A., F.N. wrote the paper.

Bibliography

[1] W. C. Swope, J. W. Pitera, and F. Suits. “Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 1. Theory”. *J. Phys. Chem. B* 108.21 (2004), pp. 6571–6581.

[2] N. Singhal, C. D. Snow, and V. S. Pande. “Using Path Sampling to Build Better Markovian State Models: Predicting the Folding Rate and Mechanism of a Tryptophan Zipper Beta Hairpin”. *J. Chem. Phys.* 121.1 (2004), pp. 415–425.

[3] F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl. “Constructing the Equilibrium Ensemble of Folding Pathways from Short Off-Equilibrium Simulations”. *Proc. Natl. Acad. Sci.* 106.45 (2009), pp. 19011–19016.

[4] 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.* 134.17 (2011), p. 174105.

[5] 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”. *The Journal of Chemical Physics* 126.15 (2007), p. 155101.

[6] F. Noé, I. Horenko, C. Schütte, and J. C. Smith. “Hierarchical Analysis of Conformational Dynamics in Biomolecules: Transition Networks of Metastable States”. *J. Chem. Phys.* 126.15 (2007), p. 155102.

[7] J. D. Chodera and F. Noé. “Markov State Models of Biomolecular Conformational Dynamics”. *Current Opinion in Structural Biology* 25 (2014), pp. 135–144.

[8] B. E. Husic and V. S. Pande. “Markov State Models: From an Art to a Science”. *J. Am. Chem. Soc.* 140.7 (2018), pp. 2386–2396.

[9] V. A. Voelz, M. Jäger, S. Yao, Y. Chen, L. Zhu, S. A. Waldauer, G. R. Bowman, M. Friedrichs, O. Bakajin, L. J. Lapidus, S. Weiss, and V. S. Pande. “Slow Unfolded-State Structuring in Acyl-CoA Binding Protein Folding Revealed by Simulation and Experiment”. *J. Am. Chem. Soc.* 134.30 (2012), pp. 12565–12577.

[10] Q. Qiao, G. R. Bowman, and X. Huang. “Dynamics of an Intrinsically Disordered Protein Reveal Metastable Conformations That Potentially Seed Aggregation”. *J. Am. Chem. Soc.* 135.43 (2013), pp. 16092–16101.

[11] D. Shukla, Y. Meng, B. Roux, and V. S. Pande. “Activation Pathway of Src Kinase Reveals Intermediate States as Targets for Drug Design”. *Nat Commun* 5.1 (2014), p. 3397.

[12] M. M. Sultan, G. Kiss, and V. S. Pande. “Towards Simple Kinetic Models of Functional Dynamics for a Kinase Subfamily”. *Nature Chem* 10.9 (2018), pp. 903–909.

[13] S. M. Hanson, G. Georghiou, M. K. Thakur, W. T. Miller, J. S. Rest, J. D. Chodera, and M. A. Seeliger. “What Makes a Kinase Promiscuous for Inhibitors?” *Cell Chemical Biology* 26.3 (2019), 390–399.e5.

[14] F. Paul, Y. Meng, and B. Roux. “Identification of Druggable Kinase Target Conformations Using Markov Model Metastable States Analysis of Apo-Abl”. *J. Chem. Theory Comput.* 16.3 (2020), pp. 1896–1912.

[15] T. Hempel, N. Plattner, and F. Noé. *Coupling of Conformational Switches in Calcium Sensor Unraveled with Local Markov Models and Transfer Entropy*. Preprint. Biophysics, 2020. url: http://biorxiv.org/lookup/doi/10.1101/2020.02.25.964353.

[16] T. Löhr, K. Kohlhoff, G. T. Heller, C. Camilloni, and M. Vendruscolo. "A Kinetic Ensemble of the Alzheimer's Aβ Peptide". *Nat Comput Sci* 1.1 (2021), pp. 71–78.

[17] D.-A. Silva, G. R. Bowman, A. Sosa-Peinado, and X. Huang. "A Role for Both Conformational Selection and Induced Fit in Ligand Binding by the LAO Protein". *PLoS Comput Biol* 7.5 (2011). Ed. by R. Nussinov, e1002054.

[18] K. J. Kohlhoff, D. Shukla, M. Lawrenz, G. R. Bowman, D. E. Konerding, D. Belov, R. B. Altman, and V. S. Pande. "Cloud-Based Simulations on Google Exacycle Reveal Ligand Modulation of GPCR Activation Pathways". *Nature Chem* 6.1 (2014), pp. 15–21.

[19] P. Tiwary, V. Limongelli, M. Salvalaglio, and M. Parrinello. "Kinetics of Protein–Ligand Unbinding: Predicting Pathways, Rates, and Rate-Limiting Steps". *Proc Natl Acad Sci USA* 112.5 (2015), E386–E391.

[20] N. Plattner and F. Noé. "Protein Conformational Plasticity and Complex Ligand-Binding Kinetics Explored by Atomistic Simulations and Markov Models". *Nat. Commun.* 6 (2015), p. 7653.

[21] F. Paul, C. Wehmeyer, E. T. Abualrous, H. Wu, M. D. Crabtree, J. Schöneberg, J. Clarke, C. Freund, T. R. Weikl, and F. Noé. "Protein-Peptide Association Kinetics beyond the Seconds Timescale from Atomistic Simulations". *Nat. Commun.* 8.1 (2017), p. 1095.

[22] B. C. Taylor, C. T. Lee, and R. E. Amaro. "Structural Basis for Ligand Modulation of the CCR2 Conformational Landscape". *PNAS* 116.17 (2019), pp. 8131–8136.

[23] N. Plattner, S. Doerr, G. D. Fabritiis, and F. Noé. "Complete Protein–Protein Association Kinetics in Atomic Detail Revealed by Molecular Dynamics Simulations and Markov Modelling". *Nat. Chem.* 9.10 (2017), p. 1005.

[24] S. Olsson and F. Noé. "Dynamic Graphical Models of Molecular Kinetics". *Proc. Natl. Acad. Sci.* 116.30 (2019), pp. 15001–15006.

[25] C. Adami, C. Ofria, and T. C. Collier. "Evolution of Biological Complexity". *Proceedings of the National Academy of Sciences* 97.9 (2000), pp. 4463–4468.

[26] D. W. McShea and R. N. Brandon. *Biology's First Law: The Tendency for Diversity and Complexity to Increase in Evolutionary Systems*. Chicago; London: University of Chicago Press, 2010.

[27] Y. I. Wolf, M. I. Katsnelson, and E. V. Koonin. "Physical Foundations of Biological Complexity". *Proc Natl Acad Sci USA* 115.37 (2018), E8678–E8687.

[28] J. A. Marsh and S. A. Teichmann. "Structure, Dynamics, Assembly, and Evolution of Protein Complexes". *Annu. Rev. Biochem.* 84.1 (2015), pp. 551–575.

[29] M. Dibak, M. J. del Razo, D. De Sancho, C. Schütte, and F. Noé. "MSM/RD: Coupling Markov State Models of Molecular Kinetics with Reaction-Diffusion Simulations". *The Journal of Chemical Physics* 148.21 (2018), p. 214107.

[30] M. J. del Razo, M. Dibak, C. Schütte, and F. Noé. "Multiscale Molecular Kinetics by Coupling Markov State Models and Reaction-Diffusion Dynamics". *J. Chem. Phys.* 155.12 (2021), p. 124109.

[31] C. P. Ponting and R. R. Russell. “The Natural History of Protein Domains”. *Annu. Rev. Biophys. Biomol. Struct.* 31.1 (2002), pp. 45–71.

[32] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan. “Protein Sectors: Evolutionary Units of Three-Dimensional Structure”. *Cell* 138.4 (2009), pp. 774–786.

[33] Y. Tong, D. Hughes, L. Placanica, and M. Buck. “When Monomers Are Preferred: A Strategy for the Identification and Disruption of Weakly Oligomerized Proteins”. *Structure* 13.1 (2005), pp. 7–15.

[34] A. L. Hodgkin and A. F. Huxley. “A Quantitative Description of Membrane Current and Its Application to Conduction and Excitation in Nerve”. *The Journal of Physiology* 117.4 (1952), pp. 500–544.

[35] D. Noble. “Cardiac Action and Pacemaker Potentials Based on the Hodgkin-Huxley Equations”. *Nature* 188.4749 (1960), pp. 495–497.

[36] C. E. Clancy and Y. Rudy. “Linking a Genetic Defect to Its Cellular Phenotype in a Cardiac Arrhythmia”. *Nature* 400.6744 (1999), pp. 566–569.

[37] M. Fink and D. Noble. “Markov Models for Ion Channels: Versatility versus Identifiability and Speed”. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 367.1896 (2009), pp. 2161–2179.

[38] D. Sigg. “Modeling Ion Channels: Past, Present, and Future”. *J. Gen. Physiol.* 144.1 (2014), pp. 7–26.

[39] J. D. Moreno, T. J. Lewis, and C. E. Clancy. “Parameterization for In-Silico Modeling of Ion Channel Interactions with Drugs”. *PLOS ONE* 11.3 (2016), e0150761.

[40] F. Noé and F. Nüske. “A Variational Approach to Modeling Slow Processes in Stochastic Dynamical Systems”. *Multiscale Model. Simul.* 11.2 (2013), pp. 635–655.

[41] H. Wu and F. Noé. “Variational Approach for Learning Markov Processes from Time Series Data”. *J Nonlinear Sci* (2019).

[42] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard. “A Direct Approach to Conformational Dynamics Based on Hybrid Monte Carlo”. *J. Comput. Phys.* 151.1 (1999), pp. 146–168.

[43] F. Nüske, B. G. Keller, G. Pérez-Hernández, A. S. J. S. Mey, and F. Noé. “Variational Approach to Molecular Kinetics”. *J. Chem. Theory Comput.* 10.4 (2014), pp. 1739–1752.

[44] R. T. McGibbon and V. S. Pande. “Variational Cross-Validation of Slow Dynamical Modes in Molecular Kinetics”. *The Journal of Chemical Physics* 142.12 (2015), p. 124105.

[45] A. Mardt, L. Pasquali, H. Wu, and F. Noé. “VAMPnets for Deep Learning of Molecular Kinetics”. *Nat. Commun.* 9.1 (2018), pp. 1–11.

[46] I. Satake. *Linear Algebra*. Pure and Applied Mathematics. New York: Dekker, 1975.

[47] A. Mardt, L. Pasquali, F. Noé, and H. Wu. “Deep Learning Markov and Koopman Models with Physical Constraints”. *Proc. First Math. Sci. Mach. Learn. Conf.* Ed. by J. Lu and R. Ward. Vol. 107. Proceedings of Machine Learning Research. Princeton University, Princeton, NJ, USA: PMLR, 2020, pp. 451–475.

[48] T. Xie, A. France-Lanord, Y. Wang, Y. Shao-Horn, and J. C. Grossman. “Graph Dynamical Networks for Unsupervised Learning of Atomic Scale Dynamics in Materials”. *Nat Commun* 10.1 (2019), p. 2667.

[49] I. Mezić. “Analysis of Fluid Flows via Spectral Properties of the Koopman Operator”. *Annu. Rev. Fluid Mech.* 45.1 (2013), pp. 357–378.

[50] H. Wu, F. Nüske, F. Paul, S. Klus, P. Koltai, and F. Noé. “Variational Koopman Models: Slow Collective Variables and Molecular Kinetics from Short off-Equilibrium Simulations”. *J. Chem. Phys.* 146.15 (2017), p. 154104.

[51] J. J. Clare. “Targeting Voltage-Gated Sodium Channels for Pain Therapy”. *Expert Opinion on Investigational Drugs* 19.1 (2010), pp. 45–62.

[52] F. Ashcroft. *Ion Channels and Disease: Channelopathies*. 2000.

[53] E. Flood, C. Boiteux, B. Lev, I. Vorobyov, and T. W. Allen. “Atomistic Simulations of Membrane Ion Channel Conduction, Gating, and Modulation”. *Chem. Rev.* 119.13 (2019), pp. 7737–7832.

[54] Y. Rudy and J. R. Silva. “Computational Biology in the Study of Cardiac Ion Channels and Cell Electrophysiology”. *Q. Rev. Biophys.* 39.1 (2006), pp. 57–116.

[55] P. Deuflhard, W. Huisinga, A. Fischer, and C. Schütte. “Identification of Almost Invariant Aggregates in Reversible Nearly Uncoupled Markov Chains”. *Linear Algebra and its Applications* 315.1 (2000), pp. 39–59.

[56] S. Röblitz and M. Weber. “Fuzzy Spectral Clustering by PCCA+: Application to Markov State Models and Data Classification”. *Adv. Data Anal. Classif.* 7.2 (2013), pp. 147–179.

[57] F. Noé, S. Doose, I. Daidone, M. Löllmann, M. Sauer, J. D. Chodera, and J. C. Smith. “Dynamical Fingerprints for Probing Individual Relaxation Processes in Biomolecular Dynamics with Simulations and Kinetic Experiments”. *PNAS* 108.12 (2011), pp. 4822–4827.

[58] N.-V. Buchete and G. Hummer. “Coarse Master Equations for Peptide Folding Dynamics”. *J. Phys. Chem. B* 112.19 (2008), pp. 6057–6069.

[59] A. A. Hagberg, D. A. Schult, and P. J. Swart. “Exploring Network Structure, Dynamics, and Function Using NetworkX”. *Proc. 7th Python Sci. Conf.* Ed. by G. Varoquaux, T. Vaught, and J. Millman. Pasadena, CA USA, 2008, pp. 11–15.

[60] T. M. J. Fruchterman and E. M. Reingold. “Graph Drawing by Force-Directed Placement”. *Softw: Pract. Exper.* 21.11 (1991), pp. 1129–1164.

[61] T. C. Südhof. “Neurotransmitter Release: The Last Millisecond in the Life of a Synaptic Vesicle”. *Neuron* 80.3 (2013), pp. 675–690.

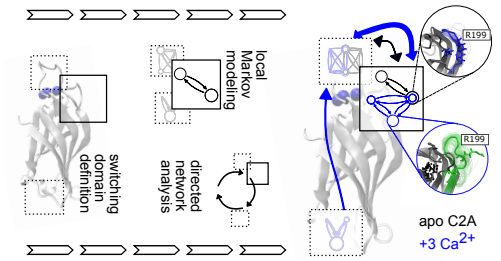
[62] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw. “How Fast-Folding Proteins Fold”. *Science* 334.6055 (2011), pp. 517–520.

[63] C. Wehmeyer, M. K. Scherer, T. Hempel, B. E. Husic, S. Olsson, and F. Noé. “Introduction to Markov State Modeling with the PyEMMA Software [Article v1.0]”. *LiveCoMS* 1.1 (2019), p. 5965.

[64] P. Virtanen et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. *Nat. Methods* 17 (2020), pp. 261–272.

[65] M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz, and F. Noé. “PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models”. *J. Chem. Theory Comput.* 11.11 (2015), pp. 5525–5542.

[66] T. Hempel, M. J. D. Razo, C. T. Lee, B. C. Taylor, R. E. Amaro, and F. Noé. *Independent Markov Decomposition*. Zenodo. 2021. `https://zenodo.org/record/5091726`.



Visual summary.

4

Coupling of Conformational Switches in Calcium Sensor Unraveled with Local Markov Models and Transfer Entropy

This Chapter has been published as

Tim Hempel, Nuria Plattner, and Frank Noé. "Coupling of Conformational Switches in Calcium Sensor Unraveled with Local Markov Models and Transfer Entropy". *Journal of Chemical Theory and Computation* 16.4 (2020), pp. 2584–2593. https://doi.org/10.1021/acs.jctc.0c00043

Reprinted with permission from T. Hempel, N. Plattner, and F. Noé. “Coupling of Conformational Switches in Calcium Sensor Unraveled with Local Markov Models and Transfer Entropy”. *J. Chem. Theory Comput.* 16.4 (2020), pp. 2584–2593. Copyright 2020 American Chemical Society. The online article can be obtained from the publisher via `http://pubs.acs.org/articlesonrequest/AOR-KEXGHKDKPVT463FTHUZN`.

Contributions TH was the lead author in this project. The research was designed by FN and NP. TH has conducted the research, including running large-scale MD simulations and hidden Markov state model estimation. The MD setup was contributed by NP. TH developed, implemented, and validated the new methods presented in the manuscript, including local hidden Markov state modeling and information theoretic coupling analyses. TH created the figures, and was main author of the manuscript. All authors contributed to writing the manuscript. (This paragraph summarizes TH’s contributions alone, it is not an exhaustive list of other authors’ contributions.)

Abstract

Proteins often have multiple switching domains that are coupled to each other and to the binding of ligands in order to realize signaling functions. Here we investigate the C2A domain of Synaptotagmin-1 (Syt-1), a calcium sensor in the neurotransmitter release machinery and a model system for the large family of C2 membrane binding domains. We combine extensive molecular dynamics (MD) simulations with Markov modeling in order to model conformational switching domains, their states, and their dependence on bound calcium ions. Then, we use transfer entropy to characterize how the switching domains are coupled via directed or allosteric mechanisms and give rise to the calcium sensing function of the protein. Our proposed switching mechanism contributes to the understanding of the neurotransmitter release machinery. Furthermore, the methodological approach we develop serves as a template to analyze conformational switching domains and the broad study of their coupling in macromolecular machines.

4.1 Introduction

Molecular modeling comes with challenges. Even though the advent of graphical processing units (GPUs) has delivered vast amounts of data from increasingly large molecular systems, modelers must still cope with two fundamental problems: the sampling problem and the curse of dimensionality. In particular, a limiting factor of molecular modeling is the (possibly exponential) growth of the number of metastable states with system size. Sampling all metastable states reversibly can thus be infeasible, especially if the described processes are extremely slow. Recently, dynamical graphical models have been proposed to overcome this problem by treating the evolution of configurations of many small molecular switches, e.g., dihedral rotamers, in an Ising-model like fashion [1].

Here we develop a complementary approach that, instead of working with very small spinlike molecular switches that can be readily identified from the structure, we identify conformational switching domains and model their kinetics with local Markov state models (MSMs). To this end, we partition the protein into subsystems that are modeled separately as conformational switching domains (henceforth simply "conformational switches"). We therefore do not need to parametrize an MSM of the full structure and thus require less extensive molecular dynamics (MD) sampling. This decomposition approach is supported by experimental evidence for dynamic protein segments that have undergone evolutionary development in an independent fashion [2]. Furthermore,

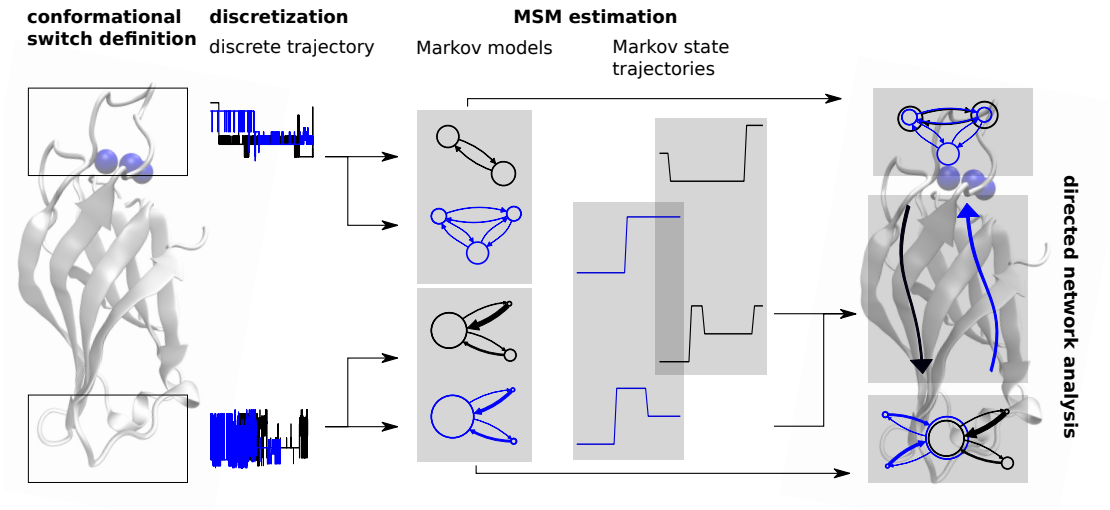


Figure 4.1: Method workflow at simplified example. From left to right: Two conformational switches are defined (top and bottom of protein), and raw MD trajectories of two ensembles (depicted by colors blue and black) are discretized into a (micro-) state space that is common to all ensembles of a particular conformational switch. Markov models are estimated separately, yielding transition matrices between metastable states and Markov state trajectories. In a following step, metastable states for each conformational switch can be identified between ensembles by comparing metastable state probability distributions, exploiting the common state space. Further, Markov state trajectories are used to estimate directional networks between conformational switches, exploiting the common time frame.

we quantify the coupling between the local conformational switches by using mutual information and transfer entropy in the time series of Markov states. Our approach is summarized in Fig. 4.1.

Our study is motivated by our interest in the role that protein conformational dynamics plays in determining spatiotemporal regulation of cellular processes and signaling. A class of proteins that play important roles in cellular signaling (among other processes) is the C2 family, which consists of 96,833 small β-sandwich structured protein domains [3]. C2 domains are membrane binders that in many cases require ion binding for activation.

One of the most exciting yet unresolved problems connected to C2 domains is neurotransmitter exocytosis. Here, C2 domains play roles in vesicle recruitment (MUNC13) as well as in the actual release mechanism (Synaptotagmin-1, -2, -7 and -9) [4]. More specifically, synchronous neurotransmitter release is triggered by the double-C2-motif Synaptotagmin-1 (Syt-1) as a reaction to the electrically driven calcium concentration change. Syt-1 is specifically known to cause fast, concerted neurotransmitter release [5] and hence is a good candidate for investigating calcium induced signaling with MD. Its structure and environment is summarized in Figure 4.2. Even though the chain of reac-

tions is well understood, it remains unclear how the ion concentration change triggers signaling on the atomistic level. Possible mechanisms include a passive reaction due to the change in electrostatic potential, changes of protein structure, or adapted conformational dynamics. The first mechanism has been studied extensively; our aim here, however, is to investigate the role of Syt-1 conformational dynamics for its signaling function. Short preliminary test simulations of both Syt-1 domains, C2A and C2B, with and without membrane revealed that even single domain conformational dynamics is complex. However, extensively sampling the whole system's rare transition events between all possible membrane binding modes and protein conformations is prohibitively expensive with direct MD. We thus break the problem down into obtaining kinetic models for individual components that will be coupled later [1]. As a first step, we analyze the Syt-1 C2A domain as it showed the most interesting behavior in our tests.

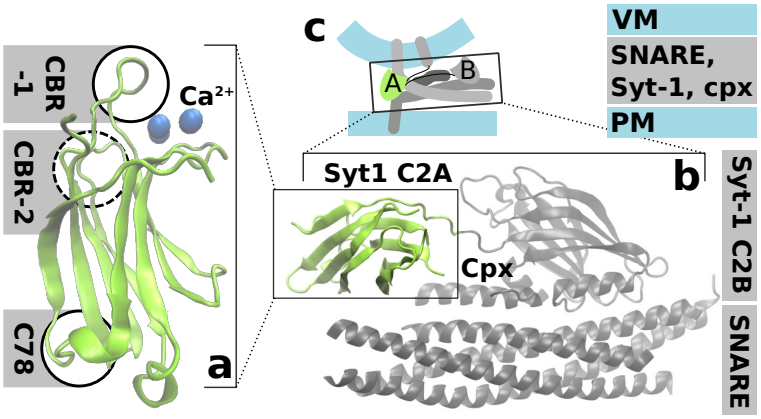


Figure 4.2: Syt-1 C2A with interaction partners for synaptic exocytosis. **(a)** Structure of Syt-1 C2A, active sites used in later analysis highlighted by circles; calcium ions are depicted as blue spheres. **(b)** Syt-1 C2A (green) in primed fusion complex with its sister domain, Syt-1 C2B, and SNARE/Cpx (gray) based on PDB 5W5C [6]. **(c)** cellular context with plasma membrane (PM) and vesicle membrane (VM) [6].

The difficulties in understanding the inherently complex signaling process arise partially due to a lack of experimental methods capable of capturing the functional processes at a high spatial resolution over sufficiently long time. Molecular dynamics (MD) simulations, on the other hand, enable us to analyze the C2 signaling function at atomistic resolution by assessing differences between the calcium bound and apo C2 domains, such as the Syt-1 C2A domain. We then use Markov modeling to analyze the equilibrium reweighted conformational dynamics of the system, its Ca^{2+} configuration, and the calcium dependent information exchange between its different conformational switches.

4.2 Determining Conformational Switches and Their Coupling

We employ two main methods in order to analyze protein dynamics as a coupling of conformational switching domains: (1) decomposing the protein structure into conformational switching domains and modeling their kinetics using local MSMs and (2) characterizing how these conformational switches are coupled by computing mutual information and transfer entropy between local MSMs. Here we outline the main concepts, see Figure 4.1 for a visual summary and Methods for details.

4.2.1 Local Markov State Models identify conformational switches

Conceptually, conformational switches are local protein regions that can exist in different discrete states, whose switching may be either spontaneous (i.e., due to thermal noise) or induced, e.g., by ligand binding or other changes of the equilibrium. Here we outline an approach to identify conformational switches that undergo spontaneous transitions in the simulation data.

We first decompose the protein or protein complex into conformational switching domains. When these are not obvious (as in the present application), one can choose from a variety of methods to find optimal decompositions of the full system into subsystems, e.g., refs [7] and [8].

Second, we discretize the conformational switch state space into metastable states using MSMs [9–15]. In the context of MD, MSMs model exchange between often fine-grained protein states by counting transitions between them. Based on this empirical count matrix, a transition probability matrix under reversibility constraints is estimated. Equilibrium probabilities of protein conformations follow from an eigendecomposition of this matrix. Not only does Markov modeling provide a means to combine arbitrary numbers of short off-equilibrium trajectories, but also it has turned out to be an efficient approach to learn about the slow and biologically relevant processes in vast MD data sets [16–18]. We further note that kinetic modeling is necessary because in the finite data regime, MD data sets are not equilibrium samples. Thus, simple time averages in general do not represent thermodynamic ensemble averages. These can, however, be estimated using MSMs.

While various MSM approaches can be employed in order to identify metastable sets, such as Robust Perron Cluster Cluster Analysis [19], VAMPnets [20], and others (e.g., refs [21] and [22]), here we estimate local Markov models via the hidden Markov model (HMM) approach described in ref [23]. HMMs yield a mapping of the complex dynamics into an easily interpretable set of a few “hidden” or metastable states of the local

conformational switches. Besides the kinetic model, metastable trajectories are used for the analysis. Choosing HMMs over MSMs has two advantages: (a) HMMs provide information also at short time scales due to their faster convergence and (b) HMMs can be used to generate metastable trajectories using the Viterbi algorithm [24]. In our experience, the Viterbi algorithm greatly outperforms metastable trajectory estimates of, e.g., MSMs/PCCA+, which is crucial for the follow-up analysis with mutual information and transfer entropy.

4.2.2 Coupling Conformational Switches via Transfer Entropy

To analyze pairwise directed influences between conformational switches, we use mutual information and transfer entropy. Mutual information describes undirected couplings akin to correlations which we estimate between conformational switches. A similar strategy as proposed here was pursued in ref [25], where MSMs were built using residue solvent-accessible surface area as features. Directed information, in contrast, is directional. To estimate both quantities, we exploit (a) the fact that metastable trajectories of different local protein features are time-synchronous and (b) that the systems in the HMM formulation are Markovian by definition. We can thus simplify Schreiber's original definition of transfer entropy [26] using the HMM transition matrix $p(x_{n+1}|x_n)$ and its stationary distribution $\pi_X(x_n)$ that are defined for states x_n in a time series X. From time series Y to time series X, transfer entropy $T(Y \to X)$ is hence defined

$$T(Y \to X) = \sum_{\substack{x_n, x_{n+1} \in X \\ y_n \in Y}} \pi_{X,Y}(x_n, y_n) \cdot p(x_{n+1}|x_n, y_n) \cdot \log_2 \left(\frac{p(x_{n+1}|x_n, y_n)}{p(x_{n+1}|x_n)} \right). \tag{4.1}$$

Joint probabilities in X and Y, $\pi_{X,Y}$, are obtained from a transition matrix estimate in the combinatorial state space. The transition probability $p(x_{n+1}|x_n, y_n)$ is computed from the combinatorial transition matrix by marginalization over y_{n+1}. This takes into account the common time frame and probes for causal effects of the current state onto the state that follows one Markov step in the future. Choosing the lag time of the underlying MSM ensures that our analysis indeed captures crosstalk of processes that our MSMs describe, even though other choices are theoretically possible.

Transfer entropy can be interpreted as a measure of the "incorrectness of [the] assumption" that "the state of [Y] has no influence on the transition probabilities on system [X]" [26]. It probes dependency in a direction-dependent fashion, i.e., is a heuristic

for the amount of influence the process Y has on the trajectory of a process X. According to this interpretation, we take $T(X \rightarrow Y) \gg T(Y \rightarrow X)$ as statistical evidence for directed influence from process X to process Y.

Our approach resembles the one applied by ref [27] in which Schreiber's transfer entropy was used to model allostery by describing entropy sources and sinks in a protein at residue resolution. However, whereas in ref [27] results are based on probability estimates from histogramed fluctuation vectors, here we model interactions between groups of residues in the space of Markov states and use probability estimates from HMMs. In comparison to approaches such as ref [28], we do not alter the underlying physical interactions to infer causal relationships.

4.3 Results

4.3.1 Syt-C2A Consists of Multiple Metastable Conformational Switches

The analysis of conformational dynamics is conducted with focus on potentially functional protein regions. The Syt-C2A protein body adopts a β-sandwich fold which is naturally very rigid, while the connections between the β-sheets are flexible. Furthermore, the largest flexible parts coincide with the calcium binding region (CBR), but there are several other flexible regions at various locations in the protein.

A simple analysis of root-mean-square fluctuations (SI Figure B.2) shows that only the CBR-1 loop becomes significantly stabilized with calcium binding. In the CBR-2, the opposite seems to happen, which hints toward an opening of the rigid body structure for this particular loop. CBR-2 is closely attached to the protein body in the crystal structure.

In order to obtain a human-readable model, we exploit our observation that Syt-C2A loops operate seemingly independently and subdivide Syt-C2A into three metastable regions of interest that are henceforth described as conformational switches: CBR-1, CBR-2, and a site opposite to the CBR (Figure 4.2). The latter region will be called C78 since it connects β-sheets 7 and 8 (according to nomenclature of ref [29]). According to our analysis, the remaining regions express no metastability and will thus not be analyzed. We employ a Markov state model analysis, here using Hidden Markov Models (HMMs) in order to identify the metastable states within these conformational switches (Methods).

4.3.2 Local Markov Models Reveal Calcium Mediated Loop Dynamics

Our models predict that CBR-1 can form an α-helix as depicted in Figure 4.3a. The free energy difference between the helical and other states is $\Delta G_\alpha - \Delta G_{\text{other}} = 1.08^{+0.26}_{-0.26}$ kcal mol^{-1} (without calcium) and slightly shifts to $\Delta G^*_\alpha - \Delta G^*_{\text{other}} = -0.02^{+0.8}_{-0.11}$ kcal mol^{-1} (with calcium). The lifetime of this helical state becomes higher, from about $2.34^{+1.15}_{-0.48}$ µs (without calcium) to $7.19^{+3.23}_{-1.46}$ µs (with calcium). This means that the calcium binding enables a more stable α-helical CBR-1 conformation.

This population change is mediated by the three calcium ions inside the CBR which interact with the CBR-1 loops by Coulomb interactions, attracting the polar acidic residue D172 (Figure 4.3a). This residue is located at the center of the α-helix sequence; Coulomb attraction toward the protein core stabilizes this structure. We corroborate this observation by noting that D172 has a very low solvent exposure in its α-helical conformation (see SI Figure B.3). The second most probable structure in the calcium bound protein is a β-hairpin-like structure (Figure 4.3a). In general, it is very rigid and gives high solvent exposure to M137.

However, even without calcium, stabilized configurations exist, as depicted in Figure 4.3a. One of them is characterized by M173 being buried within the CBR (red structure in Figure 4.3a). The stationary distribution is tilted toward the M173-burying state in the calcium unbound protein. Besides a disordered state (denoted by D in Figure 4.3a), CBR-1 can form another α-helix located toward the N-terminal end of this loop with D172 at its center (depicted in blue).

The CBR-2 loop in the crystal structures appears tightly bound to the protein body. This conformation is populated in both of our MD data sets, however, we find metastable states that are characterized by the rearrangement of CBR-2-stabilizing salt bridges (Figure 4.3b).

In particular we note that the salt bridge R199-D178, which is found in the crystal structure and in all of the calcium-unbound trajectories (Figure 4.3b, state A), can be opened by calcium binding. R199 becomes solvent exposed in this highest populated calcium bound state. This change goes along with a particular Ca^{2+} arrangement within the binding pocket (Figure 4.4b,c). Mechanistically, the CBR-2 residues are influenced by the calcium ion charges. This explains why positively charged residues such as R199 can loosen from the protein body in the calcium bound trajectories. The calcium binding thus enables R199 to be solvent accessible. We found that the conformational changes between the tightly bound state and the loosely bound ones is on the order of 12 µs with one exception: In the calcium bound case, it takes $50.65^{+61.45}_{-22.61}$ µs to attach the loop to the protein body again (result not shown in Figure 4.3b).

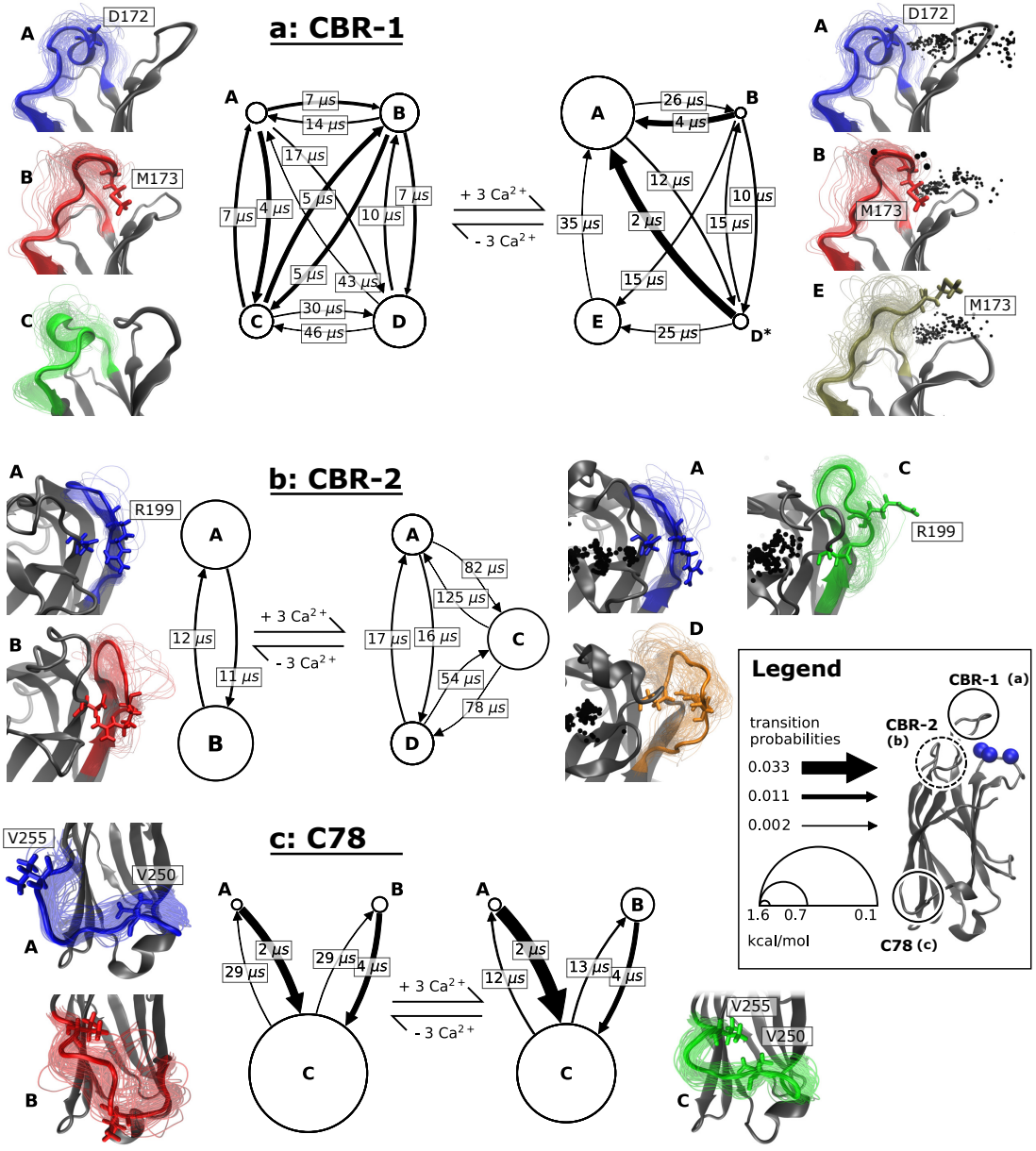


Figure 4.3: Markov models of Syt-1 C2A conformational switches in the apo state (left) and calcium-bound state (right). Capital letters denote metastable states. Representative structures are shown and color-coded to distinguish states within bound/apo models. Circle sizes are proportional to the state's equilibrium probabilities within each model. Arrows indicate transitions between metastable states, thickness proportional to the transition rate and annotated by pairwise inverse transition rate. **(a)** MSMs of CBR-1 including sampled Ca^{2+} positions (black dots). States D and D^* denote disordered structures that could not be assigned to unique structural elements. **(b)** MSMs of CBR-2 with its loop in closed (crystal structure configuration, blue) and open (red, green, orange) conformations. **(c)** MSMs of C78. Stick representations of single residues were drawn where instructive.

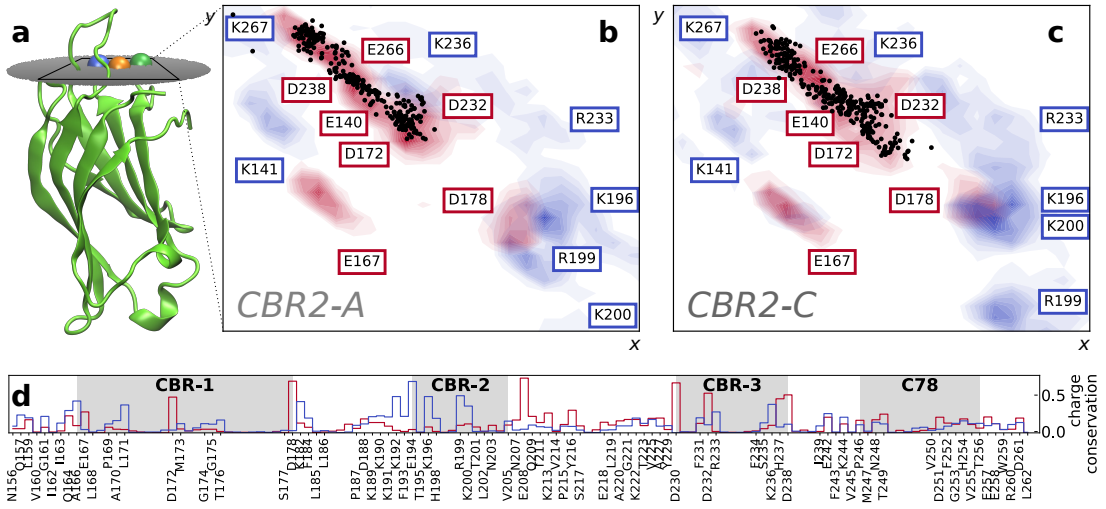


Figure 4.4: Ca^{2+} and binding pocket charge distribution. (a) projection plane drawn in Ca^{2+}-bound Syt-C2A structure. Panels **(b)** and **(c)**: Charge density averaged over 100 MSM samples and Ca^{2+} distribution in the calcium binding pocket. Ca^{2+} density is depicted by black points, positively charged amino acids as blue, negatively charged as red filled contour lines. MSM samples were drawn from CBR-2 MSM states A (panel **(b)**) and C (panel **(c)**), respectively. 2D projection (x, y) into the plane that is depicted in panel **(aO. (d)** Conservation of charges among C2 domain family sequences (PFAM entry PF00168, seed alignment); histogram annotated with Syt-1 C2A domain residue sequence.

The C78 loop undergoes slow processes between three metastable states which could be identified in both calcium bound and unbound data sets. With calcium binding, a shift of probability toward a V250-exposed state (cf. red structure in Figure 4.3c) can be observed. Figure 4.3c shows that two of the macrostates are characterized by releasing valine residues from the protein body. Specifically, V250 and V255 can be attached independently of calcium binding. In order to switch from the V250 to the V255 exposing state, both residues must be attached to the protein body as an intermediate state.

To conclude the discussion of our results, we would like to add that the applied Langevin integrator has been shown to have a dampening effect on the dynamics in the subnanosecond regime [30] which could possibly propagate to the high time scale dynamics described here.

4.3.3 Ca^{2+} Distribution Depends upon Protein Conformation

When analyzing the calcium configuration within the binding pocket we find that the Ca^{2+} distribution relaxes from the crystal structure coordinates into a broader distribution due to Coulomb repulsion between the ions. The crystal structure position of Ca1 acts as the inner edge of the binding pocket which is structured like a funnel along

which the ions can collectively move. Therefore, the Ca^{2+} distribution is dynamic and is concentrated across multiple possible binding sites that are more or less defined. The residues forming these binding sites are dynamic themselves, i.e. they occupy different conformations in different metastable states as exemplified in Figure 4.4b,c.

The importance of the charge distribution created by binding pocket residues becomes evident with a sequence alignment of all known members of the C2 family [3]. The histogram of conserved charges derived from the sequence alignment (Figure 4.4d) confirms that charges at Ca^{2+} coordinating residues (Syt-1 C2A residues D172, D178, D232, D230, and D238, also reported by ref [31]) are indeed conserved. Furthermore, we find that positive charges at position 199-200 (Syt-1 nomenclature) are conserved, underlining the significance of our findings. Another distinct feature in the charge conservation histogram is a conserved region of positive residues around K190, which corresponds to the lysine rich cluster described by ref [32] for the C2 domains of Syt-1 and rabphilin 3A.

We can further qualitatively generalize our results. Beyond the sequence conservation, C2 domains are highly conserved in structure as the resolved structures from the PFAM C2 family [3] have an RMSD of 1.80 ± 0.36 Å to the analyzed crystal structure (1BYN) [33]. It thus becomes very likely that conformational dynamics plays a role for other C2 domains, too, most likely within the CBR (or equivalent for non-calcium binders) or other non-β-sheet structural elements such as C78.

4.3.4 Distant Protein Features Show Allosteric Behavior

In Markov state modeling and other statistical analysis methods, there is a trade-off between the complexity of the analyzed system and the amount of detail one can resolve with statistical significance. The presented Syt-C2A example has distinct regions in which conformational switches occur. As commonly observed, obtaining a joint MSM including all combinations of conformational switches proved difficult due to insufficient statistics [1]. However, high-quality MSMs could be obtained for individual conformational switches, in the field of the remaining protein. Here we investigate how these different subsystems interact using an information theoretic approach to model the network.

A qualitative assessment of undirectional influences using Pearson's correlation coefficient (of coordinates) and mutual information (of MSMs) shows that there indeed is a weak coupling between conformational switches. As expected, coupling between spatially adjacent sites is higher than between distant sites. In particular, Ca^{2+} ions tighten the coupling between CBR-1 and CBR-2 (Figure 4.5).

It is only natural to ask if this coupling is symmetric or if any kind of unidirectional influence between conformational switches can be inferred. A crucial point for the analysis is that a common time frame can be defined from the Markov state trajectories of the individual conformational switches as they occur simultaneously. Transfer entropy $T(X \to Y)$ is a well established measure for assessing directional networks between time series X and Y [26]. It can be interpreted as a statistical measure for the amount of information transported between two systems and is thus direction dependent. We say that X influences Y if $T(X \to Y) \gg T(Y \to X)$ (Methods). This attempts to model statistical *causality* (in the Granger sense) between local protein features. In Granger's original definition [34] a random variable X "causes" Y if knowledge of the past of X reduces the uncertainty of predicting Y's future. Transfer entropy relates to this concept [35]; therefore we apply a tool corresponding to a very broad definition of allostery.

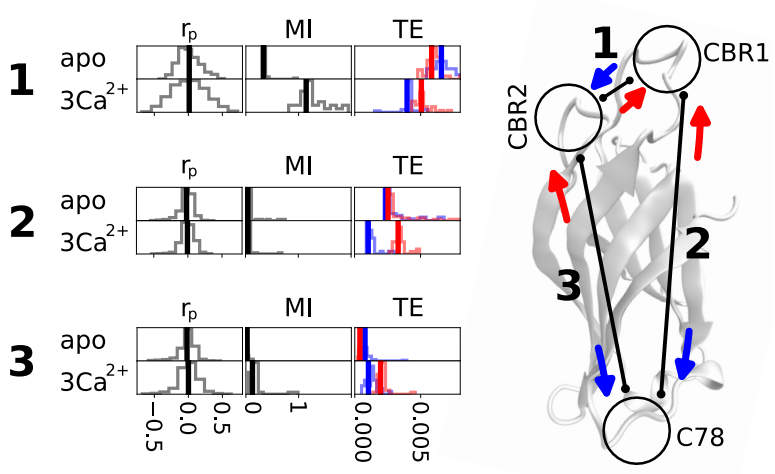


Figure 4.5: Coupling between conformational switching domains. Each link between conformational switches (enumerated 1-3) is quantified by naive computation of Pearson's correlation coefficient between plain coordinates (r_P), mutual information (MI) and probed for directionality with transfer entropy (TE). TE color code refers to different directions as shown in crystal structure. Bootstrapping histograms for MI and TE are shown to assess the estimation error (MI, TE columns). The correlation coefficients between all pairs of heavy atoms of two protein regions are depicted by histograms to qualitatively assess linear couplings (r_P column).

The transfer entropy estimates from our data are presented in Figure 4.5. We can identify a weak nonsymmetric coupling between C78 and CBR-1 in the calcium bound case. The transfer entropy between this pair of loops is significantly higher from C78 to CBR-1 than vice versa, i.e., the prediction error for CBR-1 dynamics can be decreased by adding information about C78. This can be interpreted as a causal influence from C78 onto CBR-1, suggesting an allosteric coupling that is induced by calcium binding. It accompanies a slight increase in mutual information between this pair of conformational

switches. We hypothesize that the effect is caused by long-range Coulomb interactions from the ions that is further mediated by the network of chemical bonds of the rigid protein body.

Please note that this result is purely based on statistical analysis, and however desirable, no mechanistic interpretation follows here.

4.4 Discussion

We have shown that challenges arising from the MD sampling problem and the curse of dimensionality can be met by combining local Markov state models of conformational switches with an information-theoretic analysis of their coupling. We find that the estimation of few-state local Markov models is very robust and avoids commonly experienced statistical problems with estimating Markov models of the global protein state. Furthermore, local models are easy to interpret. We therefore think that the present approach will be instrumental for the modeling of large proteins with loosely coupled local conformational switching domains.

Our results demonstrate that the effect of calcium binding to Syt-C2A is not a simple linear response as might be expected by increasing the charge in a specific part of the system but rather a complex change of the system kinetics. Even though experimental [36] and MD studies [37, 38] suggested that Syt-1 C2A does not switch between well-defined conformations upon calcium binding, sampling the system extensively shows that the C2A domain indeed undergoes subtle and rare conformational switching which is profoundly impacted by Ca^{2+}.

Furthermore, it is well established that upon calcium binding, the C2A domain effectively switches off its membrane repelling charge [39, 40] and is therefore attracted to the membrane while interacting with the SNARE-cpx complex (cf. Figure 4.2 and refs [5] and [6]). This model however neither specifies such an interaction nor does it explain how the signal (calcium ions) spreads to possibly distant interaction sites. Our model suggests that the conformational dynamics is altered upon calcium binding. This could be a potential mechanism for the calcium signal spreading to distant sites.

As in any study based on MD simulations, this work relies on the force field that was used to generate the data. In particular, we note that the applied calcium ion model is approximate and, e.g., does not account for electronic screening of charged moieties or environment-dependent changes in partial charges [41, 42]. Furthermore, it does not reflect multibody effects or even Ca^{2+}selectivity [43]. Even though this should be taken into account when basing experiments on our results, we are confident that charge-

mediated effects to the protein dynamics such as the ones laid out in this work can be described nevertheless but should be treated cautiously. Comparison to other force fields, in particular polarizable ones, or correction terms to classical force fields such as described in refs [44] and [45] would certainly improve the prediction quality at the binding site but are beyond the scope of this work.

At this point, our system model does not include elements of the cellular environment such as phospholipid membranes. The conformational dynamics is expected to be altered upon interactions with other proteins or membranes, e.g., membrane-inserted Ca^{2+} bound Syt-C2A will not obey the exact same dynamics as presented in this study. We note, however, that the presented methodological framework would be applicable and could be used to quantify effects of membrane binding to C2A conformational dynamics.

Our model predicts a calcium induced population change toward a CBR-1 α-helix (Figure 4.3a), burying negatively charged residues such as D172, thereby making the CBR more attractive to the membrane. This is consistent with the charge neutralization argument of ref [39]. Membrane attraction is further enhanced by hydrophobic residues that are not necessary for stabilizing the α-helix. The CBR-1 configurations that, according to our model, are highly populated in the calcium unbound protein (Figure 4.3a) might even reinforce the membrane repelling function by moving hydrophobic residue M173, which is exposed in the disordered CBR-1 configuration, into the binding pocket. CBR related polar acidic residues are solvent accessible in this case since they are not bound by calcium ions. This has the side effect that CBR attraction to solvent calcium ions rises.

In the CBR-2 conformational switch, we find that calcium binding enables release of R199 from the protein body (Figure 4.3b). As R199 has been reported to be of importance for membrane penetration [29], ionic interactions with SNARE [46], and formation of the C2A-C2B interface [47], our model might be key for understanding the interactions with the fusion machinery. In particular, Syt-1 interaction with SNARE might be triggered by calcium that, given its positive charge, induces the release of R199.

The C78 region generally contains many hydrophobic amino acids which opens the possibility of potential weak membrane interaction at this non-CBR site, a model previously proposed for Syt-1 C2B [48]. This interaction would allow the C2A domain to bind both, vesicle and plasma membrane, at the same time. Unfortunately, we are not aware of thorough studies that investigate C78 or adjacent sites. In order to verify our model, we thus propose experimental validation. Such a validation would include assessing the relevance of this protein site, e.g., by a mutation experiment that neutralizes all hy-

drophobic residues located there. If effects on membrane fusion can be determined, a second step involving substitution of single C78 residues might help to understand the mechanism and to validate our predictions. We expect especially valine residues 250 and 255 to alter functionality since their configurations are characteristic for metastable states in the local models.

Our study further suggests cooperativity between the loops. As expected, our correlation and mutual information analysis shows that CBR-1 and CBR-2 loops are weakly coupled. Our model further predicts that calcium binding induces a stronger coupling. Transfer entropy analysis suggests that calcium binding increases the influence that C78 exerts onto CBR-1, even though the absolute magnitude might be rather low. This weak allosteric coupling insinuates that perturbation of C78, e.g., through membrane contact or binding to another protein in the fusion complex, effects the conformational dynamics of CBR loops. Even though our model does not predict the properties of such an allosterically induced change, we believe that incorporating these findings into future studies will help to understand cryptic behavior of C2 domains.

Transfer entropy analysis as conducted in this work is limited to modeling the interplay of conformational switches. However, modeling a high resolution allosteric pathway with this technique is possible and the subject of current research.

We can further qualitatively generalize the mechanisms described here for Syt-1 C2A to a larger spectrum of proteins that are not necessarily connected to neurotransmitter secretion. As sequence alignment of the charges in the C2 family (Figure 4.4d) has shown, especially our newly described calcium dependent mechanism of CBR-2 regulation might indeed be a structural feature of C2 domains in general.

The present approach relies on identifying the conformational switching domains treated by local Markov models with expert knowledge. The development of machine learning approaches to select these conformational switches automatically is a topic of future research.

4.5 Methods

4.5.1 Molecular Dynamics Simulations

In order to observe all important dynamical processes of the system, we have carried out extensive sampling. Two hundred twenty trajectories of 2 µs individual length have been generated, adding up to a cumulated simulation time of 440 µs, which to the best of our knowledge is the largest atomistic MD data set generated for a C2 domain up to

now. Since synaptic vesicle fusion can be triggered by calcium in less than 100 µs [4] the assumption that all relevant processes have been sampled is justified.

Simulations have been carried out for the Syt-C2A domain in its Ca^{2+}-bound and Ca^{2+}-free state using the CHARMM36 force field [49]. The setup is based on the C2A structure contained in PDB 2R83 (which contains the double motif C2AB), and Ca^{2+} ion positions are initiated from PDB 1BYN which contains three calcium ions bound to the CBR. Starting structures for MD production runs were randomly drawn from a smaller precursive data set. For both setups a water box of side length 6 nm was generated with a KCl ion concentration of 0.1 mol L^{-1} at neutral total charge using the TIP3P [50] water model. The setups contain 8363 (calcium free) and 8359 water molecules (with calcium), respectively. The setup as well as equilibration in the NVT and NPT ensembles (100 ps each) were conducted with Gromacs [51], and production run simulations of 2 µs were carried out in the NPT ensemble at 300 K and 1 bar using the OpenMM software package [52]. We used 1 nm cutoff for nonbonded interactions, PME electrostatics, and rigid water molecules. We further exploited the heavy hydrogen approximation as described in the openMM documentation, allowing a 5 fs integration step with the openMM Langevin integrator with a friction coefficient of 1/ps. A simplified python script of our production run openMM configuration is given in SI Section B.2.1. An aggregated simulation time of 184 µs was obtained for the calcium free case, 256 µs for the calcium bound case, and in each case the accumulation of new data was continued until converged MSMs were obtained. We note that the Ca^{2+}binding positions are converged with time (see SI Section B.2.2). Visual representations of molecular structures were obtained with VMD [53].

4.5.2 Local MSMs

For modeling local loop dynamics, the full (bound and apo) trajectory data was directly clustered according to a minRMSD norm. This includes superposing the full protein (using MDTraj [54]) and defining discrete states according to the Euclidean norm between all atom coordinates of a specific region. This includes amino acid side chains and hydrogen atoms. We chose A170-T176 (CBR-1), R199-N203 (CBR-2) and T249-E258 (C78) to extract the local conformations, respectively. Discrete states (15-30) and the *k*-means algorithm [55] were used for each local model.

The selection of regions is inspired by a precursive full protein analysis using pairwise minimal residue distances between blocks of two residues and a state discretization in a low dimensional TICA projection [56]. Besides a relatively low spatial resolution at local sites, this analysis did not yield a valid MSM due to disconnected combinatorial

states. It however motivated us to conduct analyses on local sites and to select the ones with metastable dynamics. Please note that the distance feature described here was not used any further; local MSMs are instead built using the procedure described above.

Discretizing local protein features after global minRMSD superposition could potentially yield spurious correlations between distant protein sites [57]. We note that this is unlikely because a) the largest mass of the protein is concentrated in its rigid β-sandwich body and b) the comparably low number of discrete states at local sites is too small to capture such subtle influences. Microstate definitions for the example of CBR-2 are shown in SI Section B.2.3. Further and most importantly, the hidden metastable states that we discuss and use for mutual information and transfer entropy analysis reflect major internal loop rearrangements and are therefore most likely not affected by superposition artifacts.

HMMs were estimated separately for calcium bound and unbound data sets based on the common discretization described above. In contrast to Markov state models, HMMs are not based on the assumption that the microstates obtained by clustering are approximately Markovian. Instead, a number of hidden states (usually less than the observed microstates) are introduced as explained in detail in ref [24]. The number of hidden states was chosen such that it yields the best resolution of the slow processes in the protein and fulfills the validation criteria for Markov modeling (see below). All HMMs were estimated at lag time 50 ns using the PyEMMA software package [55]. We observed that only in relatively small configuration spaces (such as the ones defined by partitioning the protein into conformational switches), a robust, common discretization into Markov states could be obtained.

Corresponding HMM states between calcium bound and unbound HMMs were identified by computing KL-divergences between their observation probabilities per microstate (i.e. in the joint state space) and using a cutoff to identify hidden states across data sets. This operation is only possible because all data is discretized the same way i.e. microstate definitions do not differ. We use the KL-divergence between these distributions, and an example (identification of macrostates for C78) is shown in SI Figure B.4. Note that the KL-divergence becomes zero in the limit of equal distributions, thus one can define a cutoff of 2 below which states are identified for all models.

As our HMMs display high metastability, we can use the approximation $K = T - 1$ to compute the rate matrix K from the transition matrix T. The elements of K are pairwise inverse transition times.

Markov models were validated using two measures [15]. First, the implied time scales were tested for convergence. We further tested for stationary distribution convergence

to ensure that the presented results do not depend on the model lag time. Second, the Chapman-Kolmogorow equation was used to check for consistency between predictions of the presented models with direct estimates at higher lag times. Using Bayesian sampling of the posterior, the error was estimated for all the quantities computed by the MSMs/HMMs [58]. Results are presented in SI Section B.2.4.

4.5.3 Free Energy Validation of Ion Force Field

For model validation purposes, we derive binding free energies of each Ca^{2+} using alchemical free energy perturbation methods [59, 60] for the crystal structure conformation.

As a buffer solution contains monovalent ions, it can be expected that calcium binding sites are normally saturated by those ions. This is confirmed by our calcium-free simulations in which we observe potassium ions in the binding pocket. Alchemical free energy perturbation was hence simplified to a computational transformation of $Ca^{2+} \rightarrow K^{+}$ in the binding pocket and the opposite transformation in the solvent.

We further assume that in experiments, ions fill the binding pocket subsequently, i.e., we compute Ca3 in the presence of (Ca1, Ca2), Ca2 in the presence of Ca1, and Ca1 by itself. As we are interested in the binding free energy difference of a particular protein conformation, we apply harmonic constraints to the full protein. Relative free energies were computed using MBAR [60].

Our results were validated according to ref [61]. This includes the validation of alchemical intermediate overlap and convergence analysis. Further, we repeat the same calculation at least 20 times to make sure that our results are reproducible.

The resulting ion binding free energies from the equilibrated crystal structure are comparable to experiment within error; however they do not fully match the experimentally observed order (SI Section B.2.6).

4.5.4 Combinatorial Viterbi Path Model

Since a central concept of this work is hidden state space trajectories or Viterbi paths, the most important concepts [24] are summarized here. The Viterbi algorithm maximizes the probability P of the hidden state sequence $q_1 q_2 \dots q_n$ with the observed sequence $O_1 O_2 \dots O_n$ given our model λ

$$\max_{q_1 q_2 \dots q_n} P(q_1 q_2 \dots q_n, O_1 O_2 \dots O_n | \lambda). \tag{4.2}$$

To this end, the Viterbi path is a maximum likelihood path on the hidden states. It naturally comes with the time step of the underlying HMM. Its estimation algorithm is also part of PyEMMA [55].

The degree up to which the local models communicate was estimated by comparing the stationary distribution of independent models to a new estimate. By assuming independence, this property follows directly from the product of local model probabilities per state.

Without assuming independence, a new estimate for the dynamics in combinatorial state space was made based on the local model Viterbi paths. It yields the combinatorial transition matrix or conditional probabilities $p(x_{i+1}, y_{i+1}|x_i, y_i)$ and joint stationary distribution $\pi_{X,Y}(x, y)$ of two processes X and Y.

4.5.5 Mutual Information

Let X and Y be Markov processes that are localized at distinct spatial features of a protein and are well-described by (local, individual) HMMs. The mutual information M of X and Y is defined in terms of the joint probability $p_{X,Y}(x, y)$ and the independent probabilities $p_X(x),\ p_Y(y)$ of $x \in X$ and $y \in Y$. In this work, we identify these probabilities p as the stationary probabilities of Markov models, most commonly denoted by π. We can thus write

$$M(X, Y) = \sum_{\substack{x \in X \\ y \in Y}} \pi_{X,Y}(x, y)\ \log_2 \left(\frac{\pi_{X,Y}(x, y)}{\pi_X(x)\pi_Y(y)} \right). \tag{4.3}$$

We obtain the independent stationary distributions π_X and π_Y from the local hidden Markov models and the joint stationary distribution $\pi_{X,Y}$ from a combined Markov model estimated based on the Viterbi paths. We use eq (4.3) to compute mutual information between all pairs of conformational switches of Syt-1 C2A .

For interpreting mutual information, it can be rewritten in terms of the KL-divergence $\mathbf{D}(\cdot\|\cdot)$ as follows:

$$M(X, Y) = \mathbf{D}\left(\pi_{X,Y}(x, y) \| \pi_X(x)\pi_Y(y)\right) \tag{4.4}$$

Mutual information can thus be interpreted as a measure of the incorrectness of the assumption that both systems are independent [26].

Validation measures for both mutual information and transfer entropy are given in SI Section B.2.5.

Acknowledgement

We acknowledge funding from the Deutsche Forschungsgemeinschaft (SFB / TRR186, project A12 and NO825/3-2), the Einstein Foundation Berlin (SoOPic) and the European Commission (ERC CoG 772230 “ScaleCell”). We are grateful for discussions and feedback from Martin Scherer, Simon Bärfuss, Thomas Söllner, Fabian Paul, Shreyas Kaptan, Sebastian Stolzenberg, Guillermo Pérez-Hernández, Katarzyna Ziółkowska, and Andreas Mardt and thank Brooke E. Husic for proofreading the manuscript.

Bibliography

[1] S. Olsson and F. Noé. “Dynamic Graphical Models of Molecular Kinetics”. *Proc. Natl. Acad. Sci.* 116.30 (2019), pp. 15001–15006.

[2] P. Csermely, R. Palotai, and R. Nussinov. “Induced Fit, Conformational Selection and Independent Dynamic Segments: An Extended View of Binding Events”. *Trends Biochem. Sci.* 35.10 (2010), pp. 539–546.

[3] S. El-Gebali, J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, M. Qureshi, L. J. Richardson, G. A. Salazar, A. Smart, E. L. L. Sonnhammer, L. Hirsh, L. Paladin, D. Piovesan, S. C. E. Tosatto, and R. D. Finn. “The Pfam Protein Families Database in 2019”. *Nucleic Acids Res.* 47.D1 (2019), pp. D427–D432.

[4] T. C. Südhof. “Neurotransmitter Release: The Last Millisecond in the Life of a Synaptic Vesicle”. *Neuron* 80.3 (2013), pp. 675–690.

[5] E. R. Chapman. “How Does Synaptotagmin Trigger Neurotransmitter Release?” *Annu. Rev. Biochem.* 77.1 (2008), pp. 615–641.

[6] Q. Zhou, P. Zhou, A. L. Wang, D. Wu, M. Zhao, T. C. Südhof, and A. T. Brunger. “The Primed SNARE–Complexin–Synaptotagmin Complex for Neuronal Exocytosis”. *Nature* 548.7668 (2017), pp. 420–425.

[7] S. Bernhard and F. Noé. “Optimal Identification of Semi-Rigid Domains in Macromolecules from Molecular Dynamics Simulation”. *PLoS ONE* 5.5 (2010). Ed. by M. J. Buehler, e10491.

[8] L. Boninsegna, R. Banisch, and C. Clementi. “A Data-Driven Perspective on the Hierarchical Assembly of Molecular Structures”. *J. Chem. Theory Comput.* 14.1 (2018), pp. 453–460.

[9] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard. “A Direct Approach to Conformational Dynamics Based on Hybrid Monte Carlo”. *J. Comput. Phys.* 151.1 (1999), pp. 146–168.

[10] W. C. Swope, J. W. Pitera, and F. Suits. “Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 1. Theory”. *J. Phys. Chem. B* 108.21 (2004), pp. 6571–6581.

[11] N. Singhal, C. D. Snow, and V. S. Pande. “Using Path Sampling to Build Better Markovian State Models: Predicting the Folding Rate and Mechanism of a Tryptophan Zipper Beta Hairpin”. *J. Chem. Phys.* 121.1 (2004), pp. 415–425.

[12] F. Noé, I. Horenko, C. Schütte, and J. C. Smith. “Hierarchical Analysis of Conformational Dynamics in Biomolecules: Transition Networks of Metastable States”. *J. Chem. Phys.* 126.15 (2007), p. 155102.

[13] F. Noé. “Probability Distributions of Molecular Observables Computed from Markov Models”. *J. Chem. Phys.* 128.24 (2008), p. 244103.

[14] F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl. “Constructing the Equilibrium Ensemble of Folding Pathways from Short Off-Equilibrium Simulations”. *Proc. Natl. Acad. Sci.* 106.45 (2009), pp. 19011–19016.

[15] 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.* 134.17 (2011), p. 174105.

[16] N. Plattner, S. Doerr, G. D. Fabritiis, and F. Noé. “Complete Protein–Protein Association Kinetics in Atomic Detail Revealed by Molecular Dynamics Simulations and Markov Modelling”. *Nat. Chem.* 9.10 (2017), p. 1005.

[17] F. Paul, C. Wehmeyer, E. T. Abualrous, H. Wu, M. D. Crabtree, J. Schöneberg, J. Clarke, C. Freund, T. R. Weikl, and F. Noé. “Protein-Peptide Association Kinetics beyond the Seconds Timescale from Atomistic Simulations”. *Nat. Commun.* 8.1 (2017), p. 1095.

[18] B. E. Husic and V. S. Pande. “Markov State Models: From an Art to a Science”. *J. Am. Chem. Soc.* 140.7 (2018), pp. 2386–2396.

[19] S. Röblitz and M. Weber. “Fuzzy Spectral Clustering by PCCA+: Application to Markov State Models and Data Classification”. *Adv. Data Anal. Classif.* 7.2 (2013), pp. 147–179.

[20] A. Mardt, L. Pasquali, H. Wu, and F. Noé. “VAMPnets for Deep Learning of Molecular Kinetics”. *Nat. Commun.* 9.1 (2018), pp. 1–11.

[21] W. Wang, T. Liang, F. K. Sheong, X. Fan, and X. Huang. “An Efficient Bayesian Kinetic Lumping Algorithm to Identify Metastable Conformational States via Gibbs Sampling”. *J. Chem. Phys.* 149.7 (2018), p. 072337.

[22] L. Martini, A. Kells, R. Covino, G. Hummer, N.-V. Buchete, and E. Rosta. “Variational Identification of Markovian Transition States”. *Phys. Rev. X* 7.3 (2017), p. 031060.

[23] 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.* 139.18 (2013), p. 184114.

[24] L. R. Rabiner. “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”. *Proc. IEEE* 77.2 (1989), pp. 257–286.

[25] J. R. Porter, K. E. Moeder, C. A. Sibbald, M. I. Zimmerman, K. M. Hart, M. J. Greenberg, and G. R. Bowman. “Cooperative Changes in Solvent Exposure Identify Cryptic Pockets, Switches, and Allosteric Coupling”. *Biophys. J.* 116.5 (2019), pp. 818–830.

[26] T. Schreiber. “Measuring Information Transfer”. *Phys. Rev. Lett.* 85.2 (2000), pp. 461–464.

[27] A. Hacisuleyman and B. Erman. “Entropy Transfer between Residue Pairs and Allostery in Proteins: Quantifying Allosteric Communication in Ubiquitin”. *PLOS Comput. Biol.* 13.1 (2017), e1005319.

[28] H. Kamberaj and A. van der Vaart. “Correlated Motions and Interactions at the Onset of the DNA-Induced Partial Unfolding of Ets-1”. *Biophys. J.* 96.4 (2009), pp. 1307–1317.

[29] J. L. Jiménez, G. R. Smith, B. Contreras-Moreira, J. G. Sgouros, F. A. Meunier, P. A. Bates, and G. Schiavo. “Functional Recycling of C2 Domains Throughout Evolution: A Comparative Study of Synaptotagmin, Protein Kinase C and Phospholipase C by Sequence, Structural and Modelling Approaches”. *J. Mol. Biol.* 333.3 (2003), pp. 621–639.

[30] J. E. Basconi and M. R. Shirts. “Effects of Temperature Control Algorithms on Transport Properties and Kinetics in Molecular Dynamics Simulations”. *J. Chem. Theory Comput.* 9.7 (2013), pp. 2887–2899.

[31] R. Fernández-Chacón, A. Königstorfer, S. H. Gerber, J. García, M. F. Matos, C. F. Stevens, N. Brose, J. Rizo, C. Rosenmund, and T. C. Südhof. “Synaptotagmin I Functions as a Calcium Regulator of Release Probability”. *Nature* 410.6824 (2001), pp. 41–49.

[32] J. Guillén, C. Ferrer-Orta, M. Buxaderas, D. Pérez-Sánchez, M. Guerrero-Valero, G. Luengo-Gil, J. Pous, P. Guerra, J. C. Gómez-Fernández, N. Verdaguer, and S. Corbalán-García. "Structural Insights into the Ca2+ and PI(4,5)P2 Binding Modes of the C2 Domains of Rabphilin 3A and Synaptotagmin 1". *Proc. Natl. Acad. Sci.* 110.51 (2013), pp. 20503–20508.

[33] L. Holm and L. M. Laakso. "Dali Server Update". *Nucleic Acids Res.* 44.W1 (2016), W351–W355.

[34] C. W. J. Granger. "Investigating Causal Relations by Econometric Models and Cross-spectral Methods". *Econometrica* 37.3 (1969), pp. 424–438.

[35] P.-O. Amblard and O. J. J. Michel. "On Directed Information Theory and Granger Causality Graphs". *J. Comput. Neurosci.* 30.1 (2011), pp. 7–16.

[36] X. Shao, I. Fernandez, T. C. Südhof, and J. Rizo. "Solution Structures of the Ca2+-Free and Ca2+-Bound C2A Domain of Synaptotagmin I: Does Ca2+ Induce a Conformational Change?" *Biochemistry* 37.46 (1998), pp. 16106–16115.

[37] Z. Wu and K. Schulten. "Synaptotagmin's Role in Neurotransmitter Release Likely Involves Ca2+-Induced Conformational Transition". *Biophys. J.* 107.5 (2014), pp. 1156–1166.

[38] M. Bykhovskaia. "Calcium Binding Promotes Conformational Flexibility of the Neuronal Ca2+ Sensor Synaptotagmin". *Biophys. J.* 108.10 (2015), pp. 2507–2520.

[39] A. R. Striegel, L. M. Biela, C. S. Evans, Z. Wang, J. B. Delehoy, R. B. Sutton, E. R. Chapman, and N. E. Reist. "Calcium Binding by Synaptotagmin's C2A Domain Is an Essential Element of the Electrostatic Switch That Triggers Synchronous Synaptic Transmission". *J. Neurosci.* 32.4 (2012), pp. 1253–1260.

[40] S. Corbalan-Garcia and J. C. Gómez-Fernández. "Signaling through C2 Domains: More than One Lipid Target". *Biochim. Biophys. Acta - Biomembr.* Membrane Structure and Function: Relevance in the Cell's Physiology, Pathology and Therapy 1838.6 (2014), pp. 1536–1547.

[41] M. Kohagen, M. Lepšík, and P. Jungwirth. "Calcium Binding to Calmodulin by Molecular Dynamics with Effective Polarization". *J. Phys. Chem. Lett.* 5.22 (2014), pp. 3964–3969.

[42] I. Leontyev and A. Stuchebrukhov. "Accounting for Electronic Polarization in Non-Polarizable Force Fields". *Phys. Chem. Chem. Phys.* 13.7 (2011), pp. 2613–2626.

[43] Z. Jing, C. Liu, R. Qi, and P. Ren. "Many-Body Effect Determines the Selectivity for Ca^{2+} and Mg^{2+} in Proteins". *Proc. Natl. Acad. Sci.* 115.32 (2018), E7495–E7501.

[44] T. Martinek, E. Duboué-Dijon, Š. Timr, P. E. Mason, K. Baxová, H. E. Fischer, B. Schmidt, E. Pluhařová, and P. Jungwirth. "Calcium Ions in Aqueous Solutions: Accurate Force Field Description Aided by *Ab Initio* Molecular Dynamics and Neutron Scattering". *J. Chem. Phys.* 148.22 (2018), p. 222813.

[45] A. Saxena and D. Sept. "Multisite Ion Models That Improve Coordination and Free Energy Calculations in Molecular Dynamics Simulations". *J. Chem. Theory Comput.* 9.8 (2013), pp. 3538–3542.

[46] Q. Zhou, Y. Lai, T. Bacaj, M. Zhao, A. Y. Lyubimov, M. Uervirojnangkoorn, O. B. Zeldin, A. S. Brewster, N. K. Sauter, A. E. Cohen, S. M. Soltis, R. Alonso-Mori, M. Chollet, H. T. Lemke, R. A. Pfuetzner, U. B. Choi, W. I. Weis, J. Diao, T. C. Südhof, and A. T. Brunger. "Architecture of the Synaptotagmin-SNARE Machinery for Neuronal Exocytosis". *Nature* 525.7567 (2015), pp. 62–67.

[47] K. L. Fuson, M. Montes, J. J. Robert, and R. B. Sutton. “Structure of Human Synaptotagmin 1 C2AB in the Absence of Ca2+ Reveals a Novel Domain Association,” *Biochemistry* 46.45 (2007), pp. 13041–13048.

[48] J. Rizo. “Mechanism of Neurotransmitter Release Coming into Focus”. *Protein Sci.* 27.8 (2018), pp. 1364–1391.

[49] J. B. Klauda, R. M. Venable, J. A. Freites, J. W. O’Connor, D. J. Tobias, C. Mondragon-Ramirez, I. Vorobyov, A. D. MacKerell, and R. W. Pastor. “Update of the CHARMM All-Atom Additive Force Field for Lipids: Validation on Six Lipid Types”. *J. Phys. Chem. B* 114.23 (2010), pp. 7830–7843.

[50] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. “Comparison of Simple Potential Functions for Simulating Liquid Water”. *J. Chem. Phys.* 79.2 (1983), pp. 926–935.

[51] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl. “GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers”. *SoftwareX* 1–2 (2015), pp. 19–25.

[52] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, and V. S. Pande. “OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics”. *PLOS Comput. Biol.* 13.7 (2017), e1005659.

[53] W. Humphrey, A. Dalke, and K. Schulten. “VMD: Visual Molecular Dynamics”. *J. Mol. Graph.* 14.1 (1996), pp. 33–38.

[54] R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane, and V. S. Pande. “MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories”. *Biophys. J.* 109.8 (2015), pp. 1528–1532.

[55] M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz, and F. Noé. “PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models”. *J. Chem. Theory Comput.* 11.11 (2015), pp. 5525–5542.

[56] G. Pérez-Hernández, F. Paul, T. Giorgino, G. D. Fabritiis, and F. Noé. “Identification of Slow Molecular Order Parameters for Markov Model Construction”. *J. Chem. Phys.* 139.1 (2013), p. 015102.

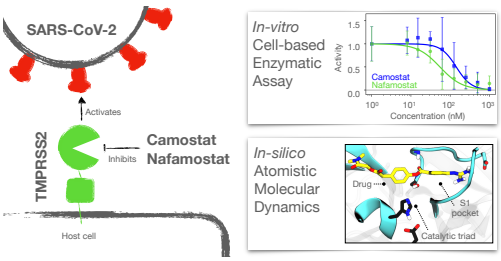
[57] P. H. Hünenberger, A. E. Mark, and W. F. van Gunsteren. “Fluctuation and Cross-correlation Analysis of Protein Motions Observed in Nanosecond Molecular Dynamics Simulations”. *J. Mol. Biol.* 252.4 (1995), pp. 492–503.

[58] B. Trendelkamp-Schroer, H. Wu, F. Paul, and F. Noé. “Estimation and Uncertainty of Reversible Markov Models”. *J. Chem. Phys.* 143.17 (2015), p. 174101.

[59] N. Hansen and W. F. van Gunsteren. “Practical Aspects of Free-Energy Calculations: A Review”. *J. Chem. Theory Comput.* 10.7 (2014), pp. 2632–2647.

[60] M. R. Shirts and J. D. Chodera. “Statistically Optimal Analysis of Samples from Multiple Equilibrium States”. *J. Chem. Phys.* 129.12 (2008), p. 124105.

[61] P. V. Klimovich, M. R. Shirts, and D. L. Mobley. “Guidelines for the Analysis of Free Energy Calculations”. *J. Comput. Aided Mol. Des.* 29.5 (2015), pp. 397–411.



Visual summary.

5

Molecular Mechanism of Inhibiting the SARS-CoV-2 Cell Entry Facilitator TMPRSS2 with Camostat and Nafamostat

This Chapter has been published as

Tim Hempel, Lluís Raich, Simon Olsson, Nurit P. Azouz, Andrea M. Klingler, Markus Hoffmann, Stefan Pöhlmann, Marc E. Rothenberg, and Frank Noé. "Molecular Mechanism of Inhibiting the SARS-CoV-2 Cell Entry Facilitator TMPRSS2 with Camostat and Nafamostat". *Chemical Science* (2021), 10.1039.D0SC05064D.
https://doi.org/10.1039/D0SC05064D

This Chapter is licensed under the Creative Commons Attribution Non-Commercial 3.0 Unported License. To view a copy of this license, visit https://creativecommons.org/licenses/by-nc/3.0/.

Contributions TH was lead author in this project. The research was designed by FN, MH, SP, and MER. TH has conducted the majority of the computational and theoretical part of the research presented in this manuscript. TH has created MD setups (with SO and LR), run large scale MD simulations, analyzed and interpreted MD simulations (with LR), designed and estimated Markov state models, and analytically derived rate model (with FN). *In-vitro* experiments were planned, conducted, analyzed, and contributed by NPA and AMK. TH created Fig. 2c, Fig. 3, and Fig. 4. TH was main author of the manuscript. LR, SO, and FN contributed to writing the manuscript. (This paragraph summarizes TH's contributions alone, it is not an exhaustive list of other authors' contributions.)

Abstract

The entry of the coronavirus SARS-CoV-2 into human lung cells can be inhibited by the approved drugs camostat and nafamostat. Here we elucidate the molecular mechanism of these drugs by combining experiments and simulations. *In vitro* assays confirm that both drugs inhibit the human protein TMPRSS2, a SARS-Cov-2 spike protein activator. As no experimental structure is available, we provide a model of the TMPRSS2 equilibrium structure and its fluctuations by relaxing an initial homology structure with extensive 330 microseconds of all-atom molecular dynamics (MD) and Markov modeling. Through Markov modeling, we describe the binding process of both drugs and a metabolic product of camostat (GBPA) to TMPRSS2, reaching a Michaelis complex (MC) state, which precedes the formation of a long-lived covalent inhibitory state. We find that nafamostat has a higher MC population than camostat and GBPA, suggesting that nafamostat is more readily available to form the stable covalent enzyme-substrate intermediate, effectively explaining its high potency. This model is backed by our *in vitro* experiments and consistent with previous virus cell entry assays. Our TMPRSS2-drug structures are made public to guide the design of more potent and specific inhibitors.

5.1 Introduction

In December 2019 several cases of unusual and severe pneumonia were reported in the city of Wuhan, China. These cases were traced back to a new coronavirus, SARS-CoV-2 (Severe acute respiratory syndrome coronavirus 2); the disease is called COVID-19 [1]. As of October 11, 2020 there are over 37 million confirmed COVID-19 cases and more than 1 million deaths [2], with both numbers likely to be severe underestimates. Given estimates of the infection mortality rate of 0.4 to 1.4 % [3–5] the virus has the potential to kill tens of millions of people unless efficient vaccines or drugs are available.

As other coronaviruses [6–9], SARS-CoV-2 exploits host proteins to initiate cell-entry, in particular TMPRSS2 and ACE2, two membrane-bound proteins expressed in the upper and lower respiratory tract [10–13]. TMPRSS2 contains an extracellular trypsin-like serine-protease domain that can proteolytically activate the spike (S) protein on the surface of SARS-CoV-2 viral particles [14] (Fig. 5.1). While in certain cell lines, the S-protein can also be activated by the endo/lysosomal pH-dependent cysteine protease cathepsin L [14, 15], virus entry into human airway cells [14, 16] seems to depend on TMPRSS2 but not cathepsin L. Consistently, epidemiological data

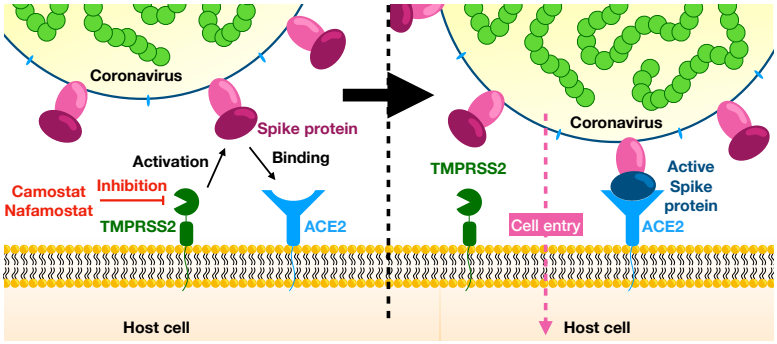


Figure 5.1: Overview of viral entry mechanism.

of prostate cancer patients undergoing androgen-deprivation therapies, which lowers TMPRSS2 levels, indicate a lower risk of contracting the SARS-CoV-2 infection [17]. We further note that low concentration levels of TMPRSS2 are observed in children and infants, possibly explaining lower risks of severe COVID-19 infections in younger age groups [18].

TMPRSS2 is also exploited by other coronaviruses and influenza A viruses for activation of surface glycoproteins, viral spread, and pathogenesis [19–25]. TMPRSS2 knock-out mice have no phenotype in the absence of infection [26], indicating that inhibiting TMPRSS2 function might not be associated with substantial unwanted side effects. As a result, TMPRSS2 is a promising therapeutic target in the context of influenza A and coronavirus infection, including SARS-CoV-2. Since TMPRSS2 is host encoded and thus genetically stable, treatment should be associated with a low risk of drug resistance

Here, we study the structural basis and molecular mechanism of TMPRSS2 inhibition by nafamostat, camostat, and its metabolic product 4-(4-guanidinobenzoyloxy)phenylacetic acid (GBPA). These guanidinobenzoyl-containing drugs are approved for human use in Japan and have been demonstrated to inhibit SARS-CoV-2 cell-entry [14, 27–29]. A recent survey of FDA approved drugs further found nafamostat to be an effective inhibitor of SARS-Cov-2 infection in human lung-cell cultures [30]. We report experimental measurements demonstrating that nafamostat and camostat inhibit TMPRSS2 activity by using our recently established cell-based assay [31], consistent with *in vitro* enzymatic TMPRSS2 activity assays [32].

Despite the hopes associated with TMPRSS2 inhibition, we are, as yet, lacking an experimental structure. We here go beyond the previous dependence on homology models by an extensive 330 microseconds of high-throughput all-atom molecular dynamics

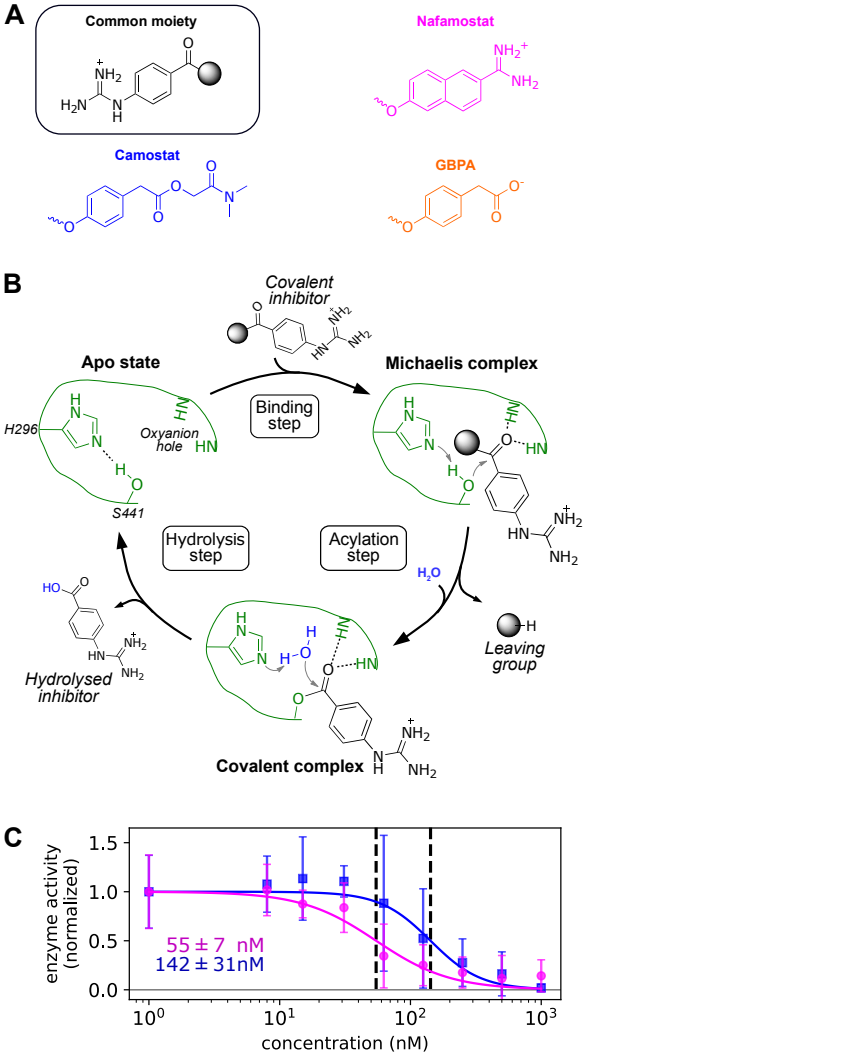


Figure 5.2: (A) Chemical structure of nafamostat (magenta), camostat (blue), and GBPA (orange), split in a common moiety (4-guanidinobenzoyl) and different leaving groups. Note that GBPA is the hydrolyzed version of camostat's leaving group ester. **(B)** General mechanism of serine proteases applied to the hydrolysis of 4-guanidinobenzoyl esters by TMPRSS2. Only H296 and S441 residues of the catalytic triad and the two backbone NH groups of the oxyanion hole are depicted for clarity (enzyme color coded in green). **(C)** Dose response behavior of TMPRSS2 inhibition by nafamostat (magenta) and camostat (blue) with IC50s (data normalized, background subtracted). Experimental enzyme activities are reported at different drug concentrations as mean and standard deviation across independent experiments, continuous lines depict fitted dose-response model used for IC50 computation.

(MD) simulations and Markov modeling. This approach provides an ensemble of equilibrium structures of the protein-drug complex and also drug binding kinetics. We show that nafamostat, camostat, and GBPA are covalent inhibitors with an identical covalent complex, but their different inhibitory activity can be explained by different populations of their Michaelis complex preceding the covalent complex. These findings, combined with the simulation structures that we make publicly available, provide an important basis for developing more potent and specific TMPRSS2 inhibitors.

5.2 Results

5.2.1 Camostat and Nafamostat inhibit the catalytic activity of TMPRSS2

First we confirm that camostat and nafamostat are TMPRSS2 inhibitors. To this end, we employ our recently reported activity assay [31] of the full-length TMPRSS2 protein on the surface of live cells with both inhibitors (Fig. 5.2A). Briefly, we transfected the human cell-line HEK-293T with a TMPRSS2 expression vector. We then measured the protease activity of the transfected cells using the fluorogenic peptide substrate BOC-QAR-AMC, following incubation of the cell with increasing inhibitor concentrations. Peptide-digestion induced a minimal increase in fluorescent signal in control cells without exogenous TMPRSS2 expression (un-normalized mean enzyme activity = 2.4), while TMPRSS2 over-expression resulted in a much faster peptide digestion (un-normalized mean enzyme activity = 12.8). Therefore, our assay is mostly specific for TMPRSS2 [31]. Significantly lower enzyme activity at higher drug concentrations can thus be attributed to TMPRSS2 inhibition.

For both camostat and nafamostat, we see a clear dose-dependent inhibition and estimate their respective IC50 values to 142 $\pm$ 31 nM and 55 $\pm$ 7 nM (Fig. 5.2C). Our results are consistent with the finding that both drugs inhibit cell entry of SARS-CoV-2 and other coronaviruses, and that nafamostat is the most potent inhibitor [27, 28, 32].

Note that in humans, camostat is rapidly processed to 4-(4-guanidinobenzoyloxy)-phenylacetic acid (GBPA) (Fig. 5.2A) [33]. It has been recently shown that GBPA also inhibits TMPRSS2 and cell entry of SARS-CoV-2 viruses, although slightly less efficiently than camostat [29]. Hence, we subsequently study the molecular interactions between TMPRSS2 and all three compounds: camostat, GBPA, and nafamostat.

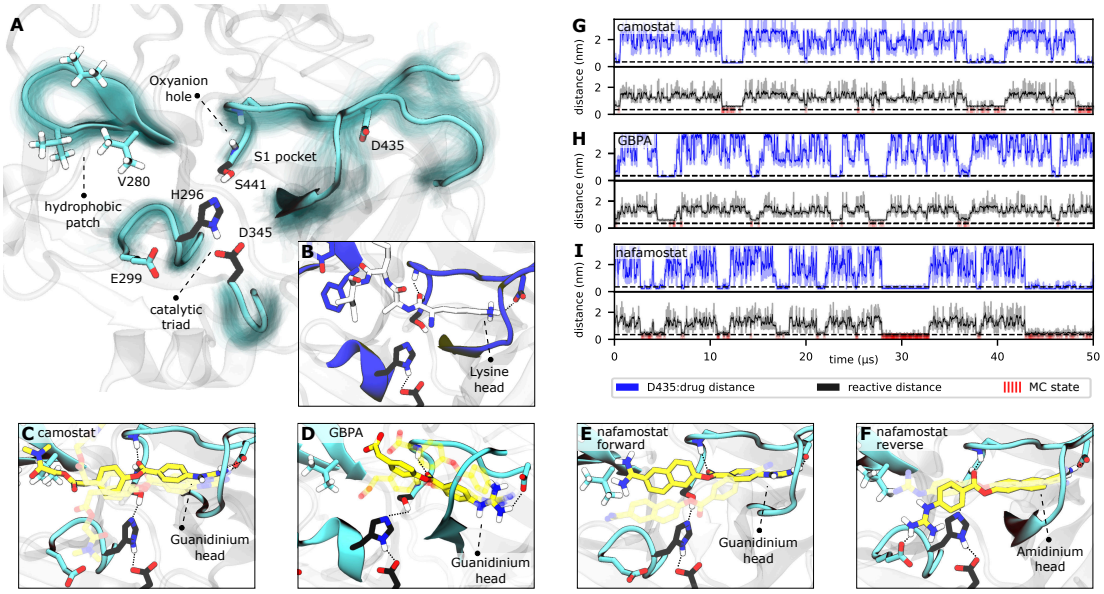


Figure 5.3: TMPRSS2 structure and Michaelis complex with camostat, its metabolic product GBPA, and nafamostat. **(A)** Active site overview of the catalytic domain of TMPRSS2. Protein flexibility is shown by cyan halo, catalyic triad is shown in black. **(B)** pre-catalytic binding mode shown at example of trypsin peptide recognition (PDB ID 4Y0Y [34], peptide displayed in white). **(C)-(F)** representative structures of camostat (**(C)**), GBPA (**(D)**), and nafamostat (**(E)-(F)**) in complex with TMPRSS2. All drugs (yellow licorice representation) bind into the S1 pocket of TMPRSS2, in **(C)-(E)** with their guanidinium heads interacting with D435, while **(F)** shows a reverse binding mode with nafamostat binding with its aminidinium head. **(G)-(I)**: Markov model simulations of minimal distance to D435 (at the S1 pocket, blue), reactivity coordinate (black), and reactivity state (i.e. when trajectory is in MC state, red) for camostat (**(G)**), GBPA (**(H)**), and nafamostat (**(I)**).

5.2.2 Equilibrium structures of TMPRSS2 in complex with Camostat and Nafamostat

We now set off to investigate the molecular mechanism of TMPRSS2 inhibition by nafamostat, camostat, and its metabolite GBPA. No TMPRSS2 crystal structure is available to date, however it has been shown that all-atom MD simulations can reliably model the equilibrium structures of proteins when (i) a reasonable model is available as starting structure, and (ii) simulations sample extensively, such that deficiencies of the starting structure can be overcome [35–39].

Here, we initialize our simulations with recent homology models of the TMPRSS2 protease domain and with camostat/nafamostat docked to them [40]. Trypsins adopt a common fold and share an active-site charge relay system whose structural requirements for catalytic activity are well understood [41]; we select our MD model consistent with these structural requirements. In particular, we focus on systems with Asp435 (substrate recognition) deprotonated and His296 (catalytic function) in a neutral form

(N_δ protonated), as well as on the interactions of a charged lysine nearby the catalytic Asp345 (Figs. C.1, C.2).

In order to avoid artifacts of the initial structural model and to simulate the equilibrium ensemble of the TMPRSS2-drug complexes, we collected a total of 100 µs of simulation data for TMPRSS2-camostat, 50 µs for TMPRSS2-GBPA, and 180 µs for TMPRSS2-nafamostat. Every drug dataset has converged RMSD distributions (Fig. C.5) and samples various drug poses and multiple association / dissociation events. Using Markov modeling [42–46] we derive the structures of the long-lived (metastable) states and characterize protein-drug binding kinetics and thermodynamics.

We find TMPRSS2 has flexible loops around the binding site but maintains stable structural features shared by other trypsin-like proteases (Figs. 5.3A and C.6). After formation of a non-covalent substrate-enzyme complex (binding step, Fig. 5.2B), trypsins cleave peptide-like bonds in two catalytic steps, assisted by a conserved catalytic triad (Asp345, His296, and Ser441 in TMPRSS2). The first step involves the formation of a covalent acyl-enzyme intermediate between the substrate and Ser441 [41]. During this step, His296 serves as a general base to deprotonate the nucleophilic Ser441, and subsequently as a general acid to protonate the leaving group of the substrate. The second step involves the hydrolysis of the acyl-enzyme intermediate, releasing the cleaved substrate and restoring the active form of the enzyme (Fig. 5.2B).

Along these two steps, the so called “oxyanion hole”, formed by the backbone NHs of Gly439 and Ser441, helps to activate and stabilize the carbonyl of the scissile bond. Another important structural feature is the S1 pocket, which contains a well conserved aspartate (Asp435) that is essential for substrate binding and recognition. At the opposite site of the S1 pocket, a loop containing a hydrophobic patch delimits the binding region of substrates within enzymatic active site. All these structural elements, known to play crucial roles in the function of serine proteases [41], are generally stable and preserved in our equilibrium structures (cf. Figs. C.2, C.6).

5.2.3 Structural basis of TMPRSS2 inhibition by Camostat and Nafamostat

Drugs with a guanidinobenzoyl moiety can inhibit trypsins by mimicking their natural substrates (Fig. 5.3B). Indeed, the ester group, resembling a peptide bond, can react with the catalytic serine with rates that are orders of magnitude faster [47], forming the acyl-enzyme intermediate. In contrast to peptide catalysis, the drug’s guanidinobenzoyl group stays covalently linked to the catalytic serine with a small off-rate, rendering it an effective chemical inhibitor [48]. Note that in its inhibited state, the TMPRSS2

active site is modified such that protease activity is disabled, preventing SARS-CoV-2 S-protein cleavage.

The present MD simulations sample different conformations of the complex formed by the enzyme with each of the drugs that precede the covalent substrate-enzyme complex. We can, therefore, elucidate their binding and how specific interactions stabilize different modes. However, please note that our simulations do not simulate the covalent complex's formation. All binding modes mimic interactions made between trypsins and their natural substrates, in which lysine heads interact with a conserved asparatate in the S1 pocket (Asp435, Fig. 5.3B). In camostat, its metabolic product GPBA, and nafamostat, the role of the lysine heads is taken by the guanidinium heads which bind in the S1 pocket and also interact with Asp435 (Fig. 5.3C-E). However, the guanidinium-Asp435 salt bridge is formed and broken transiently especially for camostat and GBPA (Fig. 5.3G-I), indicating that these drugs are not optimized for the TMPRSS2 pocket (Fig. C.4).

Nafamostat also binds in a "reverse" orientation where the amidinium head binds into the S1 pocket and interacts with Asp435 (Fig. 5.3F, [40]). In this orientation, the guanidinium head mainly interacts with Glu299, with the drug reactive center slightly displaced from the oxyanion hole, while the "forward" orientation (Fig. 5.3E) keeps the amidinium head mainly nearby Val280, with the ester center well positioned for the reaction (Fig. C.3). This observation is in agreement with several crystal structures of acyl-enzyme intermediates between different trypsins and guanidinobenzoyl molecules bound to the S1 pocket (e.g. PDBs 2AH4 [49], 3DFL [50], 1GBT [51]). There are also "inverse substrates" known to react with rates comparable to the ones of normal esters, suggesting that the inverted nafamostat orientation may also be reactive [41].

A fraction of the bound-state structures resembles a reactive Michaelis complex (MC) which fulfills the necessary criteria for catalysis of the inhibitory acyl-enzyme complex: small distances of (i) the drug ester carbon to catalytic serine oxygen, and (ii) the catalytic serine hydrogen to catalytic histidine nitrogen (Methods). We observe that besides Asp435 binding to the S1 pocket, drugs in the MC state are particularly stabilized by the oxyanion hole. Our model predicts that nafamostat has the highest MC state population followed by camostat and GBPA (Fig. 5.4), an order that coincides with the one of experimental drug binding affinities [29]. We note that the relative free energies of binding to the MC states are significantly different between nafamostat (2.1 ± 0.1 kcal/mol) and the other drugs (2.8 ± 0.1 kcal/mol and 3.1 ± 0.2 kcal/mol for camostat and GBPA, respectively), with the bootstrap sample distributions of camostat and its metabolite displaying a partial overlap.

Whereas the contact patterns of camostat and nafamostat associated states are similar, the leaving group in the inverted nafamostat conformation shows contacts predominantly with residues E299 and Tyr337 (Fig. C.3). GBPA, due to its shorter length, has less contacts to residues outside of the S1 pocket. In the reactive MC state, interestingly, all tested drugs display similar contact patterns overall, and their leaving groups bind in between Val280 and His296, with their ester group in contact with Ser441 (Fig. C.3).

5.2.4 Kinetic mechanism of TMPRSS2 inhibition by Camostat, GBPA, and Nafamostat

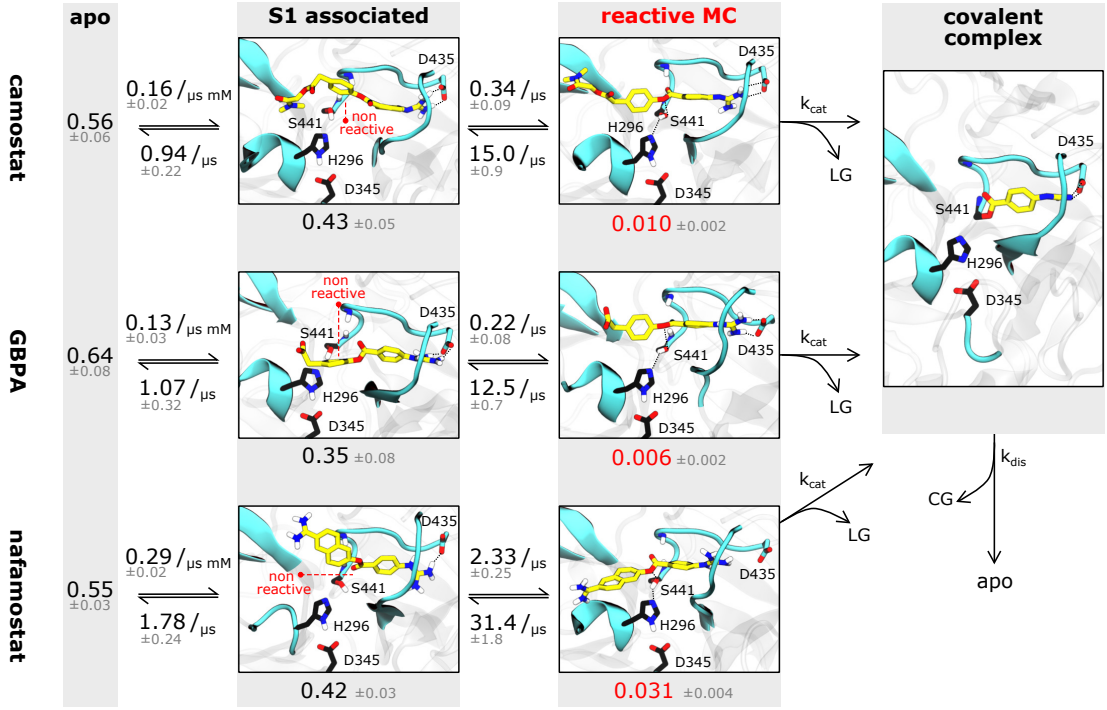


Figure 5.4: Binding kinetics model of camostat (top), camostat metabolic product GBPA (middle), and nafamostat (bottom). Inhibition process is depicted from left (apo state) to right (covalent complex). Single representative structures for each intermediate state are shown–note that all states have significant flexibility. Drugs are depicted in yellow, catalytic triad residues in black, leaving groups and covalent group are denoted by LG and CG, respectively. Rates and populations predicted by our model are annotated at reaction arrows and states, respectively. The covalent complex is illustrated using a structure with prostatin (PDB 3DFL [50]).

Finally, we investigate the molecular basis for the greater inhibition by nafamostat and formulate starting points for designing new and more efficient covalent TMPRSS2 inhibitors following these leads.

To illustrate the reversible binding of camostat, its product GBPA, and nafamostat to TMPRSS2, we used our Markov models to simulate long time-scale trajectories of 50 μs (Fig. 5.3G-I). We see a clear correlation between tight inhibitor - Asp435 interactions and contact formation between catalytic serine and the inhibitor ester group, potentially forming a reactive complex. In other words, the binding of reactive drugs in the S1 pocket favors the interactions necessary for a catalytically competent MC.

We estimate the dissociation constants for the non-covalent complex, i.e. the ratio of dissociated state and non-covalent complex populations, to be between 6 and 9 mM for the three drugs. Even though our IC50-measurements include other processes and thus are not straightforward to compare, IC50-values in the 10s-100s nanomolar range (i.e. 4-5 orders of magnitude smaller, Fig. 5.2C) are a strong indicator that the major source of inhibition cannot be the non-covalent complex, but is rather the longer-lived covalent acyl-enzyme complex. However, as all three drugs yield identical acyl-enzyme complexes, the differences in TMPRSS2 inhibition can only arise from either (1) the formation or population of their MCs, or (2) differences in the catalytic rate k_{cat} of acylation.

Interestingly, we observe that the MSM-predicted populations of the MCs in nafamostat, camostat, and GBPA have approximate ratios of 6:2:1, respectively, as well as a significantly higher on-rate for nafamostat (Fig. 5.4). A simple three-state kinetic model of dissociated state, MC and covalent complex shows that the overall association constant (K_a, ratio of inhibited versus apo protein states) directly scales with the association constant of the MC (K_a^M, ratio of MC versus dissociated states) by a constant factor (Methods):

$$K_a = K_a^M \frac{k_{\text{cat}} + k_{\text{dis}}}{k_{\text{dis}}} \tag{5.1}$$

Simply speaking, this indicates that nafamostat is a better inhibitor because it is more often found in the reactive MC state, and is therefore more likely to be attacked by the catalytic serine oxygen and enter the long-lived acyl-enzyme inhibitor complex.

Moreover, we note that the k_{cat} of acylation of these drugs may depend on their leaving group pKa's. Leaving groups with a low pKa will require less assistance from acid catalysis and will be easily displaced by the nucleophilic serine, favoring the formation of the acyl-enzyme intermediate. We expect the leaving group of nafamostat to have a lower pKa than the one of camostat, following the values of similar molecules such as naphtol (9.57 [52]) and 4-methylphenol (10.26 [53]), respectively. Indeed, these comparative insights are backed by computational pKa predictions for nafamostat (9.17), camostat (9.36), and GBPA (10.02) (Fig. C.4). We note that these predictions are made

in aqueous solution, which could differ slightly from the estimates in the enzyme due to the different environment. Nonetheless, we expect the pKa values to be in the same relative order given that the three compounds have similar contacts with the enzyme in the reactive state (Fig. C.3). This suggests that the k_{cat} of acylation will be slightly faster for nafamostat in particular compared to the camostat metabolic product GBPA, further contributing to nafamostat's superior inhibition of TMPRSS2.

5.3 Discussion

Camostat and nafamostat are promising drug candidates for a COVID-19 treatment strategy. Here we have combined cell-based assays, extensive molecular simulations, and Markov modeling to unravel the molecular action principle of these drugs and provide data that may help to improve them further.

Our binding assays provide evidence that both inhibitors directly act on TMPRSS2 and that nafamostat is more potent compared to camostat, and this qualitative difference is in agreement with complementary *in vitro* assays on purified protein construct [32] or cell-entry assays [27, 28]. We note that the absolute IC50 values differ between these three assay types, reflecting differences in experimental conditions and which function is being inhibited and measured.

While no crystallographic structure of TMPRSS2 is available, we provide extensive 330 microseconds of all-atom MD simulations starting from a homology model that generate stable equilibrium structure ensembles of the protein-drug complexes. These simulations sample multiple association / dissociation events and various drug poses in the protein active site. Our analyses show that the non-covalent complexes of nafamostat, camostat, and its metabolic product GBPA are relatively short-lived, suggesting that the main inhibitory effect is due to the formation of the long-lived covalent acyl-enzyme complex between the drug's guanidinobenzoyl moiety and the catalytic serine of TMPRSS2.

Although the MC state is not the main cause of inhibition, its population directly translates into the potency of the inhibitor, as higher MC population corresponds to a higher catalytic rate and therefore yields a larger population of inhibited enzyme. Consistently with the higher potency of nafamostat, it is found to have a threefold more stable MC compared to camostat, and sixfold compared to GBPA. A second contribution may be the pKa of drug leaving groups, affecting the rate of enzyme acylation.

Our detailed models of the thermodynamic and kinetics of inhibitor binding highlight the bound state's heterogeneity, with both drugs adopting multiple distinct poses.

We note the importance of residue Asp435 in the conserved S1-pocket, which stabilizes the MC state and helps to orient the reactive molecules in a conformation that is suited for catalysis. Nafamostat has two groups that can potentially bind into the S1 pocket, whereas camostat has only one. However, we find that the population of S1 associated states are similar between nafamostat, camostat, and GBPA, suggesting that non-covalent inhibition is likely a minor contribution to the overall inhibition of TMPRSS2.

We conclude that the design of future TMPRSS2 inhibitors with increased potency and specificity should incorporate the following points:

First, stabilizing the non-covalent complex with the TMPRSS2 active site is beneficial for both, covalent and non-covalent inhibitors. As S1 pocket binding is a major contribution to the stability of the non-covalent complex, effective drugs may contain hydrogen bond donors and positively charged moieties that could interact principally with Asp435, but also with different backbone carbonyls of the loops that compose the cavity (e.g. from Trp461 to Gly464).

Second, for covalent inhibitors, we must consider that the catalytic serine is at a distance of around 1.3 nm from Asp435. Thus, the reactive center of an effective drug and its S1-interacting moieties should be within that distance. We note that, even though all three molecules fit well in the overall active site, the guanidinobenzoyl moiety is slightly shorter than the ideal size of the TMPRSS2 cavity (Fig. C.4). We further suggest that a drug should be size-compatible to the hydrophobic patch on the S1 distal site (Figs. 5.3A and C.4). We speculate that drugs with a large end to end distance and high rigidity may not fit well in the described TMPRSS2 scaffold, and in particular, might be significantly less reactive.

Third, optimizing the pKa of the drug's leaving group might be beneficial for improving covalent TMPRSS2 inhibitors. The first step of the reaction would be faster, and the acetyl-enzyme intermediate would accumulate. We note that the deacetylation off-rate must be very low, ideally on the order of magnitude of guanidinobenzoyl moiety containing drugs.

Finally, we make our simulated equilibrium structures of TMPRSS2 in complex with the simulated drugs available, hoping they will be useful to guide future drug discovery efforts.

Acknowledgements

We acknowledge financial support from Deutsche Forschungsgemeinschaft DFG (SFB/TRR 186, Project A12), the European Commission (ERC CoG 772230 "ScaleCell"), the Berlin Mathematics center MATH+ (AA1-6) and the federal ministry of education and research BMBF (BIFOLD and RAPID Consortium, 01KI1723D). Stefan Pöhlmann was supported by DFG (PO 716/11-1) and BMBF (01K11723D). We are grateful for in-depth discussions with John D. Chodera (MSKCC New York), Matthew D. Hall (NIH), Katarina Elez, Robin Winter, Tuan Le, Moritz Hoffmann (FU Berlin), and the members of the JEDI COVID-19 grand challenge.

Software and data availability

Structural ensembles of camostat, GBPA, and nafamostat binding poses are published online at `https://github.com/noegroup/tmprss2_structures`.

Author contributions

M.H., S.P., M.E.R., and F.N. designed the study. T.H., L.R., S.O., N.P.A., and A.M.K. performed research. T.H., L.R., S.O., and N.P.A. analyzed the data. T.H., L.R., S.O., F.N. wrote the manuscript.

Competing interests

M.E.R. is a consultant for Pulm One, Spoon Guru, ClostraBio, Serpin Pharm, Allakos, Celgene, Astra Zeneca, Arena Pharmaceuticals, GlaxoSmith Kline, Guidepoint and Suvretta Capital Management, and has an equity interest in the first five listed, and royalties from reslizumab (Teva Pharmaceuticals), PEESSv2 (Mapi Research Trust) and UpToDate. M.E.R. is an inventor of patents owned by Cincinnati Children's Hospital.

5.4 Materials and Methods

5.4.1 TMPRSS2 activity assays

The TMPRSS2 activity assay was described previously [31]. Briefly, we transfected HEK-293T with a PLX304 plasmid containing the open reading frame (ORF) sequence of TMPRSS2 which encodes for the full length protein (492 amino acids). Control experiments are conducted with PLX304 plasmids.

Eighteen hours later, we replaced the media to either PBS alone or PBS in the presence of varying concentrations of candidate inhibitors camostat and nafamostat. Fifteen minutes later, we added the fluorogenic substrate BOC-QAR-AMC to the wells to induce a measurable signal of enzyme activity. We measured the fluorescent signal immediately after adding the substrate, in 15 minutes intervals for a total time of 150 minutes [31]. A baseline proteolytic activity of control cells was measured; we hypothesize that this is because of proteolytic cleavage of the substrate by endogenous transmembrane proteases. However, the TMPRSS2 overexpression cells have significantly increased proteolytic activity compared with control cells [31].

To validate the exogenous expression of TMPRSS2, we performed western-blot analysis of cell lysates from TMPRSS2 overexpressing cells and control cells. A 60 kDa band was observed in TMPRSS2 overexpressing cells but not in control cells, which is the expected molecular weight of TMPRSS2 protein after post transcriptional modifications, indicating that the target protein has been successfully expressed.

5.4.2 IC50 estimation

We used a generalized log-logistic dose-response model

$$f(x,(b,c,d,e)) = c + \frac{d-c}{1+e^{b(\ln(x)-\ln(e))}}$$

with the concentration x, c and d representing the lower and upper limits, b steepness of the curve, and e to estimate IC50 values [54].

Upper and lower limits were set to the means computed from control experiments with no drug (upper limit) and PLX plasmid (no TMPRSS2; background noise). We used scipy's [55] curve fitting algorithms to extract the IC50 with error estimates.

5.4.3 Molecular dynamics simulations

MD simulations were run with OpenMM 7.4.0 [56] using the CHARMM 36 force field (2019 version) [57]. Camostat and nafamostat structures were taken from PubChem [58] with PubChem CIDs 4413 (nafamostat) and 2536 (camostat), respectively, and modeled with the CHARMM general force field (CGenFF v. 4.3) [59]. We generated our MD setups with CharmmGUI [60]. We initiate a simulation box of side length 7.5 nm with a NaCl ion concentration of 0.1 mol/l at neutral charge and the TIP3P water model [61]. The setups contain 12038 (camostat), 12030 (GBPA), and 12039 (nafamostat) water molecules, respectively.

We run simulations in the NPT ensemble and keep the temperature at 310 K (physiological temperature) and the pressure at 1 bar. We use a Langevin integrator with 5 fs integration step and heavy hydrogen approximation (4 amu). PME electrostatics, rigid water molecules, and a 1 nm cutoff for non-bonded interactions are used. Simulation times vary between 100 and 500 ns and accumulate to 100 μs (camostat), 50 μs (GBPA), 180 μs (nafamostat), respectively. Structures were visualized using VMD [62].

Due to the lack of a crystal structure for TMPRSS2, MD simulations were seeded from a homology model. It is taken from Ref. [40], model 3W94 is chosen based on precursive MD analyses that showed that 3W94 has the most stable catalytic triad configuration (Figs. C.1, C.2). The construct includes amino acids 256 to 491 of the full sequence, corresponding to the catalytic chain except for a C-terminal Glycine missing due to homology modeling against a shorter sequence. MD simulations are seeded as follows: Equilibrated docking poses (highest scorers of Ref. [40]) of the ligand were generated in a precursive run using another homology model. We note that the used camostat docking pose resembles the one described by [63]. This data set was equilibrated with local energy minimization, 100 ps simulations with 2 fs time steps in NVT and NPT ensemble subsequently. Frames are selected based on a preliminary metastability analysis, protein conformation is constraint to 3W94 homology model using a constraint force minimizing minRMSD. Production run MD simulations are started from these poses, i.e. from the same protein configuration and with 77 (nafamostat) and 60 (camostat) ligand docking poses, respectively. To ensure convergence of sampling statistics, we ran multiple adaptive runs of simulations, seeding new simulations with coordinates associated with sparsely sampled states.

We later added the camostat metabolite GBPA by following the same setup procedure. Due to its similarity to camostat, we seeded production simulations from representative structures of the camostat stage 1 Markov model (described below) using 200 representative structures.

5.4.4 Markov modeling

We model the binding and unbinding rates in a two step procedure using Markov state type models [42–45, 64–66]. First, we describe drug unbound and associated states using a hidden Markov model (HMM) [67]. Second, we define a reactive state by using distance cutoffs.

In detail, in the first stage we define distance features between drug guanidinium group and TMPRSS2 Asp435 (minimal distance), drug amidinium group and TMPRSS2 Asp435 (minimal distance, nafamostat only). We further use a binary “reactive” dis-

tance feature defined by drug ester carbon to catalytic Ser441-OG, and catalytic serine (HG) to catalytic histidine (NE2) and a threshold of 0.35 nm. If both last mentioned distances are below the threshold, both nucleophilic attack of the serine-OG to the drug ester group and proton transfer from serine to histidin are possible, thus defining the reactive state.

We discretize this space into 243 (camostat), 240 (GBPA), and 490 (nafamostat) states using regular spatial clustering and use an HMM at lag time 5 ns with 5 (camostat, GBPA) or 8 (nafamostat) hidden states. Nafamostat yields two metastable S1 associated states encoding for both binding directions, camostat / GBPA a single one, that are defined by being at salt bridge distance to Asp435. We note no significant correlation between the hidden states and the reactive state, i.e. reactivity is not metastable. Also note that in contrast to later modeling stages, reactivity according to this HMM does not necessitate S1 pocket binding. The described HMMs are used to generate the (non-equilibrium) time series presented in Fig. 5.3G-I. Besides distance to D435, we also show a reactivity coordinate which we define as the mean of a) drug ester carbon to catalytic serine oxygen and b) catalytic serine hydrogen to catalytic histidine nitrogen. Reactivity, i.e. when both reactive distances are within range, is indicated with red markers (MC state).

In the second stage, we split the HMM bound states into reactive and non-reactive by combining HMM Viterbi paths [68] and the reactive state trajectories to one single discrete trajectory consisting of 3 states. We define the S1 associated states by filtering the Viterbi paths of the HMM according to S1-association. We use the reactivity trajectories to further bisect the S1 associated state into reactive and non-reactive states, yielding a three state discretization of the drug binding mode. Note that the S1-reactive state is a subset of the reactive state in the stage 1 HMM model.

We estimate a reversible maximum likelihood Markov state model (MSM) from the stage 2 trajectories as described in [45]. We report the stationary probability vector as well as transition rates. The latter are approximated using the matrix logarithm approximation of scipy [55] to compute the transition rate matrix R from the transition probability matrix T using the definition $T = \exp(R\tau)$ with the lag time τ. We found that all reported quantities are converged with respect to the lag time above $\tau = 500$ ns which was thus chosen as the model lag time. Errors are estimated by bootstrapping validation using a random sample (with replacement) of the stage 2 trajectory data. All MSM/HMM analyses were conducted using the PyEMMA 2 software (version 2.5.7) [69].

Dissociation constants $K_d = p_{\text{unbound}}/p_{\text{bound}}$ from the non-covalent state were estimated from this model and amount to $5.95\,\text{mM}$ $(4.60, 7.30)$ for camostat, $8.45\,\text{mM}$

$(5.81, 11.65)$ for GBPA, and $6.07\,\mathrm{mM}$ $(5.55, 6.93)$ for nafamostat (68% confidence intervals).

5.4.5 Kinetic model

Simplifying the binding kinetics into a three-state model describing the binding to / dissociation from the Michaelis complex (ligand concentration c and rates k_{on}, k_{off}), catalytic rate of entering the covalent complex (k_{cat}) and dissociation to the apo state (k_{dis}), the kinetics are described by the rate matrix:

$$K = \begin{bmatrix} -c\,k_{\mathrm{on}} & c\,k_{\mathrm{on}} & 0 \\ k_{\mathrm{off}} & -k_{\mathrm{off}} - k_{\mathrm{cat}} & k_{\mathrm{cat}} \\ k_{\mathrm{dis}} & 0 & -k_{\mathrm{dis}} \end{bmatrix} \tag{5.2}$$

with the (unnormalized) equilibrium distribution

$$\pi = \begin{bmatrix} \frac{k_{\mathrm{dis}}(k_{\mathrm{off}}+k_{\mathrm{cat}})}{c\,k_{\mathrm{on}}+k_{\mathrm{cat}}} \\ k_{\mathrm{dis}}/k_{\mathrm{cat}} \\ 1 \end{bmatrix} \tag{5.3}$$

The overall dissociation constant is then:

$$K_d = \frac{\pi_1}{\pi_2 + \pi_3} = \frac{k_{\mathrm{dis}}(k_{\mathrm{off}} + k_{\mathrm{cat}})}{k_{\mathrm{on}}(k_{\mathrm{dis}} + k_{\mathrm{cat}})} \tag{5.4}$$

The non-covalent dissociation constant of the Michaelis complex:

$$K_d^M = \frac{\pi_1}{\pi_2} = \frac{k_{\mathrm{off}}}{c\,k_{\mathrm{on}} + k_{\mathrm{cat}}} \tag{5.5}$$

The dissociation constant scales as:

$$K_d = K_d^M \frac{k_{\mathrm{dis}}}{k_{\mathrm{cat}} + k_{\mathrm{dis}}} \tag{5.6}$$

And thus the association constant scales with the stability of the Michaelis complex by a constant factor given by the rates of chemical catalysis and dissociation:

$$K_a = K_a^M \frac{k_{\mathrm{cat}} + k_{\mathrm{dis}}}{k_{\mathrm{dis}}} \tag{5.7}$$

Bibliography

[1] C. Sohrabi, Z. Alsafi, N. O'Neill, M. Khan, A. Kerwan, A. Al-Jabir, C. Iosifidis, and R. Agha. "World Health Organization Declares Global Emergency: A Review of the 2019 Novel Coronavirus (COVID-19)". *Int. J. Surg.* 76 (2020), pp. 71–76.

[2] The World Health Organization. *Coronavirus Disease (COVID-19) Situation Report 194*. Tech. rep. 2020.

[3] T. W. Russell, J. Hellewell, C. I. Jarvis, K. van Zandvoort, S. Abbott, R. Ratnayake, C. C.-1. working group, S. Flasche, R. M. Eggo, W. J. Edmunds, and A. J. Kucharski. "Estimating the Infection and Case Fatality Ratio for Coronavirus Disease (COVID-19) Using Age-Adjusted Data from the Outbreak on the Diamond Princess Cruise Ship, February 2020". *Eurosurveillance* 25.12, 2000256 (2020).

[4] H. Streeck, B. Schulte, B. Kuemmerer, E. Richter, T. Hoeller, C. Fuhrmann, E. Bartok, R. Dolscheid, M. Berger, L. Wessendorf, M. Eschbach-Bludau, A. Kellings, A. Schwaiger, M. Coenen, P. Hoffmann, M. Noethen, A.-M. Eis-Huebinger, M. Exner, R. Schmithausen, M. Schmid, and G. Hartmann. *Infection Fatality Rate of SARS-CoV-2 Infection in a German Community with a Super-Spreading Event*. Preprint. Infectious Diseases (except HIV/AIDS), 2020. url: `http://medrxiv.org/lookup/doi/10.1101/2020.05.04.20090076`.

[5] R. Verity et al. "Estimates of the Severity of Coronavirus Disease 2019: A Model-Based Analysis". *Lancet Infect. Dis.* 20.6 (2020), pp. 669–677.

[6] H. Hofmann and S. Pöhlmann. "Cellular Entry of the SARS Coronavirus". *Trends Microbiol.* 12.10 (2004), pp. 466–472.

[7] W. Li, M. J. Moore, N. Vasilieva, J. Sui, S. K. Wong, M. A. Berne, M. Somasundaran, J. L. Sullivan, K. Luzuriaga, T. C. Greenough, H. Choe, and M. Farzan. "Angiotensin-Converting Enzyme 2 Is a Functional Receptor for the SARS Coronavirus". *Nature* 426.6965 (2003), pp. 450–454.

[8] S. Matsuyama, N. Nagata, K. Shirato, M. Kawase, M. Takeda, and F. Taguchi. "Efficient Activation of the Severe Acute Respiratory Syndrome Coronavirus Spike Protein by the Transmembrane Protease TMPRSS2". *J. Virol.* 84.24 (2010), pp. 12658–12664.

[9] S. Belouzard, V. C. Chu, and G. R. Whittaker. "Activation of the SARS Coronavirus Spike Protein via Sequential Proteolytic Cleavage at Two Distinct Sites". *Proc. Natl. Acad. Sci.* 106.14 (2009), pp. 5871–5876.

[10] C. G. K. Ziegler et al. "SARS-CoV-2 Receptor ACE2 Is an Interferon-Stimulated Gene in Human Airway Epithelial Cells and Is Detected in Specific Cell Subsets across Tissues". *Cell* 181.5 (2020), 1016–1035.e19.

[11] S. Lukassen, R. L. Chua, T. Trefzer, N. C. Kahn, M. A. Schneider, T. Muley, H. Winter, M. Meister, C. Veith, A. W. Boots, B. P. Hennig, M. Kreuter, C. Conrad, and R. Eils. "SARS-CoV-2 Receptor ACE2 and TMPRSS2 Are Primarily Expressed in Bronchial Transient Secretory Cells". *EMBO J.* 39.10 (2020), e105114.

[12] K. Bilinska, P. Jakubowska, C. S. Von Bartheld, and R. Butowt. "Expression of the SARS-CoV-2 Entry Proteins, ACE2 and TMPRSS2, in Cells of the Olfactory Epithelium: Identification of Cell Types and Trends with Age". *ACS Chem Neurosci* 11.11 (2020), pp. 1555–1562.

[13] Y. J. Hou et al. “SARS-CoV-2 Reverse Genetics Reveals a Variable Infection Gradient in the Respiratory Tract”. *Cell* 182.2 (2020), 429–446.e14.

[14] M. Hoffmann, H. Kleine-Weber, S. Schroeder, N. Krüger, T. Herrler, S. Erichsen, T. S. Schiergens, G. Herrler, N.-H. Wu, A. Nitsche, M. A. Müller, C. Drosten, and S. Pöhlmann. “SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor”. *Cell* 181.2 (2020), 271–280.e8.

[15] T. Ou, H. Mou, L. Zhang, A. Ojha, H. Choe, and M. Farzan. “Hydroxychloroquine-Mediated Inhibition of SARS-CoV-2 Entry Is Attenuated by TMPRSS2”. *BioRxiv Doi101101-20200722216150* (2020).

[16] M. Hoffmann, K. Mösbauer, H. Hofmann-Winkler, A. Kaul, H. Kleine-Weber, N. Krüger, N. C. Gassen, M. A. Müller, C. Drosten, and S. Pöhlmann. “Chloroquine Does Not Inhibit Infection of Human Lung Cells with SARS-CoV-2”. *Nature* 585 (2020), pp. 588–590.

[17] M. Montopoli, S. Zumerle, R. Vettor, M. Rugge, M. Zorzi, C. Catapano, G. Carbone, A. Cavalli, F. Pagano, E. Ragazzi, T. Prayer-Galetti, and A. Alimonti. “Androgen-Deprivation Therapies for Prostate Cancer and Risk of Infection by SARS-CoV-2: A Population-Based Study (N = 4532)”. *Ann. Oncol.* 31.8 (2020), pp. 1040–1045.

[18] B. A. Schuler, A. C. Habermann, E. J. Plosa, C. J. Taylor, C. Jetter, M. E. Kapp, J. T. Benjamin, P. Gulleman, D. S. Nichols, L. Z. Braunstein, A. Hackett, M. Koval, S. H. Guttentag, T. S. Blackwell, Vanderbilt COVID-19 Consortium Cohort, S. A. Webber, N. E. Banovich, J. A. Kropski, J. M. S. Sucre, and HCA Lung Biological Network. *Age-Determined Expression of Priming Protease TMPRSS2 and Localization of SARS-CoV-2 Infection in the Lung Epithelium*. Preprint. Developmental Biology, 2020. url: `http://biorxiv.org/lookup/doi/10.1101/2020.05.22.111187`.

[19] Y. Zhou, P. Vedantham, K. Lu, J. Agudelo, R. Carrion, J. W. Nunneley, D. Barnard, S. Pöhlmann, J. H. McKerrow, A. R. Renslo, and G. Simmons. “Protease Inhibitors Targeting Coronavirus and Filovirus Entry”. *Antiviral Res.* 116 (2015), pp. 76–84.

[20] N. Iwata-Yoshikawa, T. Okamura, Y. Shimizu, H. Hasegawa, M. Takeda, and N. Nagata. “TMPRSS2 Contributes to Virus Spread and Immunopathology in the Airways of Murine Models after Coronavirus Infection”. *J. Virol.* 93.6 (2019).

[21] B. Hatesuer, S. Bertram, N. Mehnert, M. M. Bahgat, P. S. Nelson, S. Pöhlmann, S. Pöhlman, and K. Schughart. “Tmprss2 Is Essential for Influenza H1N1 Virus Pathogenesis in Mice”. *PLoS Pathog.* 9.12 (2013), e1003774.

[22] C. Tarnow, G. Engels, A. Arendt, F. Schwalm, H. Sediri, A. Preuss, P. S. Nelson, W. Garten, H.-D. Klenk, G. Gabriel, and E. Böttcher-Friebertshäuser. “TMPRSS2 Is a Host Factor That Is Essential for Pneumotropism and Pathogenicity of H7N9 Influenza A Virus in Mice”. *J. Virol.* 88.9 (2014), pp. 4744–4751.

[23] K. Sakai, Y. Ami, M. Tahara, T. Kubota, M. Anraku, M. Abe, N. Nakajima, T. Sekizuka, K. Shirato, Y. Suzaki, A. Ainai, Y. Nakatsu, K. Kanou, K. Nakamura, T. Suzuki, K. Komase, E. Nobusawa, K. Maenaka, M. Kuroda, H. Hasegawa, Y. Kawaoka, M. Tashiro, and M. Takeda. “The Host Protease TMPRSS2 Plays a Major Role in in Vivo Replication of Emerging H7N9 and Seasonal Influenza Viruses”. *J. Virol.* 88.10 (2014), pp. 5608–5616.

[24] R. L. O. Lambertz, I. Gerhauser, I. Nehlmeier, S. R. Leist, H. Kollmus, S. Pöhlmann, and K. Schughart. “Tmprss2 Knock-out Mice Are Resistant to H10 Influenza A Virus Pathogenesis”. *J. Gen. Virol.* 100.7 (2019), pp. 1073–1078.

[25] R. L. O. Lambertz, I. Gerhauser, I. Nehlmeier, S. Gärtner, M. Winkler, S. R. Leist, H. Kollmus, S. Pöhlmann, and K. Schughart. “H2 Influenza A Virus Is Not Pathogenic in Tmprss2 Knock-out Mice”. *Virol. J.* 17.1 (2020), p. 56.

[26] T. S. Kim, C. Heinlein, R. C. Hackman, and P. S. Nelson. “Phenotypic Analysis of Mice Lacking the Tmprss2-Encoded Protease”. *Mol. Cell. Biol.* 26.3 (2006), pp. 965–975.

[27] M. Hoffmann, S. Schroeder, H. Kleine-Weber, M. A. Müller, C. Drosten, and S. Pöhlmann. “Nafamostat Mesylate Blocks Activation of SARS-CoV-2: New Treatment Option for COVID-19”. *Antimicrob. Agents Chemother.* 64.6 (2020).

[28] M. Yamamoto, M. Kiso, Y. Sakai-Tagawa, K. Iwatsuki-Horimoto, M. Imai, M. Takeda, N. Kinoshita, N. Ohmagari, J. Gohda, K. Semba, Z. Matsuda, Y. Kawaguchi, Y. Kawaoka, and J.-i. Inoue. “The Anticoagulant Nafamostat Potently Inhibits SARS-CoV-2 s Protein-Mediated Fusion in a Cell Fusion Assay System and Viral Infection in Vitro in a Cell-Type-Dependent Manner”. *Viruses* 12.6 (2020), p. 629.

[29] M. Hoffmann, H. Hofmann-Winkler, J. C. Smith, N. Krueger, L. K. Sorensen, O. S. Sogaard, J. B. Hasselstrom, M. Winkler, T. Hempel, L. Raich, S. Olsson, T. Yamazoe, K. Yamatsuta, H. Mizuno, S. Ludwig, F. Noe, J. M. Sheltzer, M. Kjolby, and S. Poehlmann. *Camostat Mesylate Inhibits SARS-CoV-2 Activation by TMPRSS2-related Proteases and Its Metabolite GBPA Exerts Antiviral Activity*. Preprint. Molecular Biology, 2020. url: `http://biorxiv.org/lookup/doi/10.1101/2020.08.05.237651`.

[30] M. Ko, S. Jeon, W.-S. Ryu, and S. Kim. “Comparative Analysis of Antiviral Efficacy of FDA-approved Drugs against SARS-CoV-2 in Human Lung Cells”. *J. Med. Virol.* (2020).

[31] N. P. Azouz, A. M. Klingler, and M. E. Rothenberg. *Alpha 1 Antitrypsin Is an Inhibitor of the SARS-CoV2–Priming Protease TMPRSS2*. Preprint. Microbiology, 2020. url: `http://biorxiv.org/lookup/doi/10.1101/2020.05.04.077826`.

[32] J. H. Shrimp, S. C. Kales, P. E. Sanderson, A. Simeonov, M. Shen, and M. D. Hall. “An Enzymatic TMPRSS2 Assay for Assessment of Clinical Candidates and Discovery of Inhibitors as Potential Treatment of COVID-19” (2020).

[33] I. Midgley, A. J. Hood, P. Proctor, L. F. Chasseaud, S. R. Irons, K. N. Cheng, C. J. Brindley, and R. Bonn. “Metabolic Fate of ^{14} C-Camostat Mesylate in Man, Rat and Dog after Intravenous Administration”. *Xenobiotica* 24.1 (1994), pp. 79–92.

[34] S. Ye, B. Loll, A. A. Berger, U. Mülow, C. Alings, M. C. Wahl, and B. Koksch. “Fluorine Teams up with Water to Restore Inhibitor Activity to Mutant BPTI”. *Chem. Sci.* 6.9 (2015), pp. 5246–5254.

[35] N. Plattner and F. Noé. “Protein Conformational Plasticity and Complex Ligand-Binding Kinetics Explored by Atomistic Simulations and Markov Models”. *Nat. Commun.* 6 (2015), p. 7653.

[36] V. A. Voelz, G. R. Bowman, K. Beauchamp, and V. S. Pande. “Molecular Simulation of *Ab Initio* Protein Folding for a Millisecond Folder NTL9(1-39)”. *J. Am. Chem. Soc.* 132.5 (2010), pp. 1526–1528.

[37] N. Plattner, S. Doerr, G. D. Fabritiis, and F. Noé. “Complete Protein–Protein Association Kinetics in Atomic Detail Revealed by Molecular Dynamics Simulations and Markov Modelling”. *Nat. Chem.* 9.10 (2017), p. 1005.

[38] K. Lindorff-Larsen, P. Maragakis, S. Piana, M. P. Eastwood, R. O. Dror, and D. E. Shaw. “Systematic Validation of Protein Force Fields against Experimental Data”. *PLoS ONE* 7.2 (2012). Ed. by D. J. Muller, e32131.

[39] A. Raval, S. Piana, M. P. Eastwood, R. O. Dror, and D. E. Shaw. “Refinement of Protein Structure Homology Models via Long, All-Atom Molecular Dynamics Simulations”. *Proteins* 80 (2012), pp. 2071–2079.

[40] S. Rensi, R. B Altman, T. Liu, Y.-C. Lo, G. McInnes, A. Derry, and A. Keys. *Homology Modeling of TMPRSS2 Yields Candidate Drugs That May Inhibit Entry of SARS-CoV-2 into Human Cells*. Preprint. 2020. url: https://chemrxiv.org/articles/Homology_Modeling_of_TMPRSS2_Yields_Candidate_Drugs_That_May_Inhibit_Entry_of_SARS-CoV-2_into_Human_Cells/12009582.

[41] L. Hedstrom. “Serine Protease Mechanism and Specificity”. *Chem. Rev.* 102.12 (2002), pp. 4501–4524.

[42] W. C. Swope, J. W. Pitera, and F. Suits. “Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 1. Theory”. *J. Phys. Chem. B* 108.21 (2004), pp. 6571–6581.

[43] N. Singhal, C. D. Snow, and V. S. Pande. “Using Path Sampling to Build Better Markovian State Models: Predicting the Folding Rate and Mechanism of a Tryptophan Zipper Beta Hairpin”. *J. Chem. Phys.* 121.1 (2004), pp. 415–425.

[44] F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl. “Constructing the Equilibrium Ensemble of Folding Pathways from Short Off-Equilibrium Simulations”. *Proc. Natl. Acad. Sci.* 106.45 (2009), pp. 19011–19016.

[45] 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.* 134.17 (2011), p. 174105.

[46] B. E. Husic and V. S. Pande. “Markov State Models: From an Art to a Science”. *J. Am. Chem. Soc.* 140.7 (2018), pp. 2386–2396.

[47] B. Zerner, R. P. M. Bond, and M. L. Bender. “Kinetic Evidence for the Formation of Acyl-Enzyme Intermediates in the α-Chymotrypsin-Catalyzed Hydrolyses of Specific Substrates”. *J. Am. Chem. Soc.* 86.18 (1964), pp. 3674–3679.

[48] M. K. Ramjee, I. M. Henderson, S. B. McLoughlin, and A. Padova. “The Kinetic and Structural Characterization of the Reaction of Nafamostat with Bovine Pancreatic Trypsin”. *Thromb. Res.* 98.6 (2000), pp. 559–569.

[49] E. S. Radisky, J. M. Lee, C.-J. K. Lu, and D. E. Koshland. “Insights into the Serine Protease Mechanism from Atomic Resolution Structures of Trypsin Reaction Intermediates”. *Proc. Natl. Acad. Sci.* 103.18 (2006), pp. 6835–6840.

[50] K. W. Rickert, P. Kelley, N. J. Byrne, R. E. Diehl, D. L. Hall, A. M. Montalvo, J. C. Reid, J. M. Shipman, B. W. Thomas, S. K. Munshi, P. L. Darke, and H.-P. Su. “Structure of Human Prostasin, a Target for the Regulation of Hypertension”. *J. Biol. Chem.* 283.50 (2008), pp. 34864–34872.

[51] W. F. Mangel, P. T. Singer, D. M. Cyr, T. C. Umland, D. L. Toledo, R. M. Stroud, J. W. Pflugrath, and R. M. Sweet. “Structure of an Acyl-Enzyme Intermediate during Catalysis: (Guanidinobenzoyl)Trypsin”. *Biochemistry* 29.36 (1990), pp. 8351–8357.

[52] Z. Rappoport. *Handbook of Tables for Organic Compound Identification*. Cleveland: Chemical Rubber Co, 1967.

[53] P. Pearce and R. Simkins. “Acid Strengths of Some Substituted Picric Acids”. *Can. J. Chem.* 46.2 (1968), pp. 241–248.

[54] C. Ritz, F. Baty, J. C. Streibig, and D. Gerhard. “Dose-Response Analysis Using R”. *PLoS ONE* 10.12 (2015). Ed. by Y. Xia, e0146021.

[55] P. Virtanen et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. *Nat. Methods* 17 (2020), pp. 261–272.

[56] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, and V. S. Pande. “OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics”. *PLOS Comput. Biol.* 13.7 (2017), e1005659.

[57] R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig, and A. D. MacKerell. “Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ_1 and χ_2 Dihedral Angles”. *J. Chem. Theory Comput.* 8.9 (2012), pp. 3257–3273.

[58] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, and E. E. Bolton. “PubChem 2019 Update: Improved Access to Chemical Data”. *Nucleic Acids Res.* 47.D1 (2019), pp. D1102–D1109.

[59] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. D. Mackerell. “CHARMM General Force Field: A Force Field for Drug-like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields”. *J. Comput. Chem.* 31 (2009), pp. 671–690.

[60] S. Jo, T. Kim, V. G. Iyer, and W. Im. “CHARMM-GUI: A Web-Based Graphical User Interface for CHARMM”. *J. Comput. Chem.* 29.11 (2008), pp. 1859–1865.

[61] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. “Comparison of Simple Potential Functions for Simulating Liquid Water”. *J. Chem. Phys.* 79.2 (1983), pp. 926–935.

[62] W. Humphrey, A. Dalke, and K. Schulten. “VMD: Visual Molecular Dynamics”. *J. Mol. Graph.* 14.1 (1996), pp. 33–38.

[63] V. Kumar, J. K. Dhanjal, P. Bhargava, A. Kaul, J. Wang, H. Zhang, S. C. Kaul, R. Wadhwa, and D. Sundar. “Withanone and Withaferin-A Are Predicted to Interact with Transmembrane Protease Serine 2 (TMPRSS2) and Block Entry of SARS-CoV-2 into Cells”. *J. Biomol. Struct. Dyn.* (2020), pp. 1–13.

[64] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard. “A Direct Approach to Conformational Dynamics Based on Hybrid Monte Carlo”. *J. Comput. Phys.* 151.1 (1999), pp. 146–168.

[65] F. Noé, I. Horenko, C. Schütte, and J. C. Smith. “Hierarchical Analysis of Conformational Dynamics in Biomolecules: Transition Networks of Metastable States”. *J. Chem. Phys.* 126.15 (2007), p. 155102.

[66] F. Noé. “Probability Distributions of Molecular Observables Computed from Markov Models”. *J. Chem. Phys.* 128.24 (2008), p. 244103.

[67] 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.* 139.18 (2013), p. 184114.

[68] L. R. Rabiner. “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”. *Proc. IEEE* 77.2 (1989), pp. 257–286.

[69] M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz, and F. Noé. “PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models”. *J. Chem. Theory Comput.* 11.11 (2015), pp. 5525–5542.

6

Deep learning to decompose macromolecules into independent Markovian domains

This Chapter has been published as

Andreas Mardt*, Tim Hempel*, Cecilia Clementi, and Frank Noé. "Deep Learning to Decompose Macromolecules into Independent Markovian Domains". *Nature Communications* 13.1 (2022), p. 7101. https://doi.org/10.1038/s41467-022-34603-z

* contributed equally

This Chapter is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

Contributions TH was one of two lead authors in this project. He has co-designed the research (with the other authors) and designed the method. TH and AM jointly conducted the research. In particular, TH contributed benchmarking systems and evaluated the synaptotagmin results. TH created the figures 1, 3, 4, and 5, and was one of two main authors of the manuscript (with AM). All authors contributed to writing the manuscript. (This paragraph summarizes TH's contributions alone, it is not an exhaustive list of other authors' contributions.)

Abstract

The increasing interest in modeling the dynamics of ever larger proteins has revealed a fundamental problem with models that describe the molecular system as being in a global configuration state. This notion limits our ability to gather sufficient statistics of state probabilities or state-to-state transitions because for large molecular systems the number of metastable states grows exponentially with size. In this manuscript, we approach this challenge by introducing a method that combines our recent progress on independent Markov decomposition (IMD) with VAMPnets, a deep learning approach to Markov modeling. We establish a training objective that quantifies how well a given decomposition of the molecular system into independent subdomains with Markovian dynamics approximates the overall dynamics. By constructing an end-to-end learning framework, the decomposition into such subdomains and their individual Markov state models are simultaneously learned, providing a data-efficient and easily interpretable summary of the complex system dynamics. While learning the dynamical coupling between Markovian subdomains is still an open issue, the present results are a significant step towards learning Ising models of large molecular complexes from simulation data.

6.1 Introduction

The understanding of protein function is often interlinked with understanding protein dynamics. Molecular dynamics (MD) simulations are a valuable tool to study these dynamics on an atomistic level [1–6]. However, further methods are necessary to extract the statistically relevant information and to help overcome the discrepancy between feasible simulation length and the timescales of relevant processes. A common approach to enhance sampling of a specific process of interest is to bias the simulation along a reaction coordinate aligning with the process [7–13]. In comparison, the Markov modeling approach [14–20] extracts kinetic information and tackles the sampling problem without requiring the definition of few predefined reaction coordinates by combining arbitrary numbers of short unbiased distributed simulations to model the long-timescale behavior of target systems. Consequently, multiple software packages [21, 22] have been developed over the last decade providing assistance in estimating these models. They often include a pipeline for feature selection [21–24], dimension reduction [25–31], clustering [32–35], transition matrix estimation [15, 19, 36, 37], and coarse graining [38–44]. Markov state models (MSMs) have been applied to a wide range of molecular biology problems such as protein aggregation [45–47] or ligand binding [48–50]

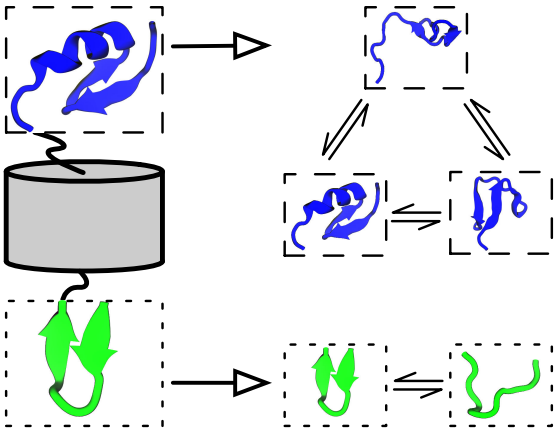


Figure 6.1: The iVAMP concept as visualized by modeling dynamics of a protein that has two independent, flexible regions separated by a rigid barrel. iVAMPnets learn an assignment of the C- (blue/top) and N-termini (green/bottom) into independent subsystems from molecular dynamics trajectories (left column). Moreover, the dynamics of both termini are modeled with statistically independent VAMPnets (right column).

and can be a valuable tool to understand experimental data on the atomistic scale [51, 52].

The necessity to assess a model's performance and thereby rank its quality encouraged the development of variational methods [53, 54], in particular the variational approach for Markov processes (VAMP) [55]. This variational formulation has allowed us to replace the aforementioned pipeline with an end-to-end deep learning framework called VAMPnet [56], which simultaneously learns a dimension reduction of the molecular system to the collective variables best describing the rare event processes and an MSM on these variables. The framework can be used to further drive MD simulations along these learned collective variables [57, 58]. We can also use this framework to estimate statistically reversible MSMs and incorporate constraints from experimental observables [59–61].

Despite these developments, there is a fundamental scaling problem in describing MD in terms of transitions between global system states: While the assignment of MD configurations to discrete global states representing the metastable groups of structures is an excellent model for small cooperative molecular systems, such as small to medium proteins, larger molecular systems (e.g. proteins with hundreds of amino acids) have an increasing number of subsystems whose dynamics are (nearly) independent [62] (Fig. 6.1). Consider, for example, a solution of N proteins which undergo transitions between their open and closed states independently when these proteins are dissociated and these transitions only (partially) couple when they are associated with other proteins. The number of global system states is 2^N, i.e. grows exponentially with the

number of subsystems N [63, 64]. This means any form of simulation or analysis which explicitly distinguishes global system states will not scale to large molecular systems.

At the same time, the (approximate) independence between subsystems is also key to the solution of the problem. A scalable solution needs to address two separate issues: (a) dividing the protein system into approximately Markovian subsystems and (b) learning the coupling between them. Olsson & Noé [63] made a first attempt at (b), by learning a dynamic graphical model between predefined subsystems. This approach leads to a graphical model, or Markov random field, resembling Ising or Potts models in physics, with the key difference that both the definition of the individual subsystems or spins as well as their transition dynamics need to be learned. In contrast, Hempel et al. [64] proposed a solution for (a) by approximating the global system dynamics as a set of independent (uncoupled) Markov models (termed Independent Markov decomposition, IMD). They furthermore propose a pairwise independence score of features, which allows to detect nearly uncoupled regions where independent Markov state models can be estimated subsequently.

In this manuscript, we present a joint IMD and VAMP approach (termed independent VAMPnet, or shorthand iVAMPnet) that significantly advances our ability to identify approximately independent Markovian subsystems (issue a) by generalizing IMD to neural network basis functions. iVAMPnets are an integrated end-to-end learning approach that decomposes the macromolecular structure into subsystems that are dynamically weakly coupled, and estimates a VAMPnet for each of these subsystems to promote a comprehensible analysis of the subsystem dynamics (Fig. 6.1). In comparison to previous implementations of IMD, our approach learns an optimal decomposition into independent subsystems and can find collective variables that are nonlinear combinations of the input features.

6.2 Results

6.2.1 Markov state models and Koopman models

Markovian dynamics can be modeled by the transition density:

$$p_\tau(\mathbf{y}|\mathbf{x}) = \mathbb{P}(\mathbf{x}_{t+\tau} = \mathbf{y}|\mathbf{x}_t = \mathbf{x}), \tag{6.1}$$

which is the probability density to observe configuration $\mathbf{y}$ at time $t + \tau$ given that the system was in configuration $\mathbf{x}$ at time t. Based on the transition density we can charac-

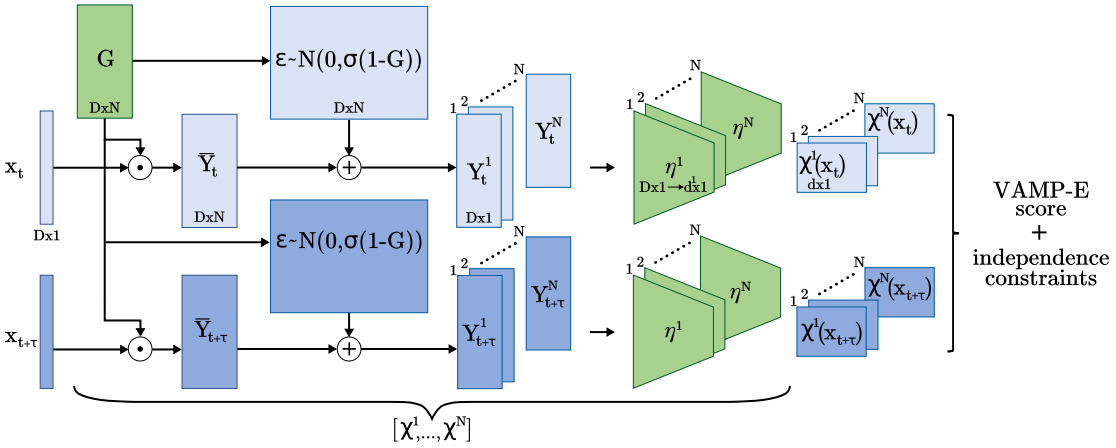


Figure 6.2: Architecture of an iVAMPnet for N subsystems, where trainable parts are shaded green. Two lobes are given for configuration pairs $\mathbf{x}_t$ and $\mathbf{x}_{t+\tau}$, where the weights are shared. Firstly, the input features are element wise weighted $\bar{\mathbf{Y}}_t = \mathbf{G} \odot \mathbf{x}_t$ with a mask $\mathbf{G} \in \mathbb{R}^{D \times N}$, where each subsystem learns its individual weighting. The mask values can be interpreted as probabilities to which subsystem the input feature belongs. In order to prevent the subsequent neural network to reverse the effects of the mask, we draw for each input feature i and subsystem j an independent, normally distributed random variable $\epsilon_{ij} \sim \mathcal{N}(0, \sigma(1-G_{ij}))$. This noise is added to the weighted features $\mathbf{Y}_t = \bar{\mathbf{Y}}_t + \boldsymbol{\epsilon}$. Thereby, the attention weight linearly interpolates between input feature and Gaussian noise, i.e., if the attention weight $G_{ij} = 1$, Y_{ij} carries exclusively the input feature x_i, if $G_{ij} = 0$, Y_{ij} is simple Gaussian noise. Afterwards, the transformed feature vector is split for each individual subsystem $\mathbf{Y}_t = [\mathbf{Y}_t^1, \ldots, \mathbf{Y}_t^N]$ and passed through the subsystem specific neural network $\boldsymbol{\eta}^i$. We call the whole transformation for a subsystem i the fuzzy state assignment $\boldsymbol{\chi}^i(\mathbf{x}_t) = \boldsymbol{\eta}^i(\mathbf{Y}_t^i)$.

terize the time evolution of a probability density χ as:

$$\chi_{t+\tau}(\mathbf{y}) = \int p_\tau(\mathbf{y}|\mathbf{x})\chi_t(\mathbf{x})\mathrm{d}\mathbf{x}. \tag{6.2}$$

By discretizing the molecular state space in a suitable way and defining a transition matrix $\mathbf{T}$ between discrete states, we can linearize this equation as:

$$\boldsymbol{\chi}_{t+\tau}(\mathbf{y}) = \mathbf{T}_\tau^\top \boldsymbol{\chi}_t(\mathbf{x}) \tag{6.3}$$

This is the equation of a Markov state model, where the element i of the vector $\boldsymbol{\chi}_{t+\tau}(\mathbf{y})$ is the probability to be in the discrete state i at time $t + \tau$. Furthermore, the transition matrix elements $(\mathbf{T}_\tau)_{ij}$ describe the transition probabilities for jumping to state j given state i within a time τ. In the case of fuzzy state assignments, e.g., as with VAMPnets, Eq. (6.3) describes the more general Koopman model [65] and $\mathbf{T}_\tau$ becomes the Koopman matrix. This means that probability densities are still propagated but the matrix elements cannot be interpreted as transition probabilities.

The lag time τ is common to all Markovian models and is usually chosen with the aid of an implied timescales test [66]. If a too small τ is chosen, the resulting model is not a valid Markov model (resulting in errors of the predicted variables) – a too large lag time produces a model that unnecessarily discards kinetic information. We therefore usually choose the smallest lag time above which the implied timescales are approximately constant.

We now seek to find a state assignment $\boldsymbol{\chi}$ and model matrix $\mathbf{T}$ that satisfy Eq. (6.3) and also succeed in predicting the long-time behavior, i.e., for multiples of the lag time τ. Formally, $\boldsymbol{\chi}$ are (initially unknown) basis functions, i.e., we assume that the relevant dynamic features can be expressed by a linear combination of them. VAMP [55] tells us that an optimal solution is reached when $\boldsymbol{\chi}$ can span the left $(\psi_1, ..., \psi_k)^\top$ and right singular functions $(\phi_i, ..., \phi_k)^\top$ of the transition operator. They can be found by maximizing the singular values of a matrix that can be estimated from simulation data (see Eqs. (6.9)-(6.13) in Methods). In the case of a VAMPnet [56], deep neural networks are trained by maximizing the VAMP score, so as to represent optimal fuzzy state assignments.

In equilibrium, the singular functions correspond to the eigenfunctions of the Markov state model and the singular values to its eigenvalues. As the Koopman model still propagates densities, it is instructive to inspect the eigenfunctions and implied timescales of $\mathbf{T}$ since they describe the slow dynamics of a given system.

6.2.2 iVAMPnets and iVAMP-score

To implement iVAMPnets, we need to bridge the gap between the deep neural networks of VAMPnets and the spatial decomposition of independent Markov models. The general idea is to set up multiple parallel VAMPnets, each modeling the Markovian dynamics of a separate, independent subsystem of the molecule, together with an attention mechanism that identifies these subsystems. Thus, each independent VAMPnet should only receive the time dependent molecular geometry features representing its specific subsystem. For example, such an attention mechanism could separate different protein domains and channel the data of individual domains to separate VAMPnets. We therefore develop an architecture that combines a meaningful attention mechanism and parallel VAMPnets and trains them with a loss function that simultaneously promotes dynamic independence between the subsystems and slow kinetics within each subsystem (Fig. 6.2). iVAMPnets are designed to optimize both these objectives simultaneously.

In practice, we extract all time-lagged data pairs $\mathbf{x}_t, \mathbf{x}_{t+\tau}$ that contain all molecular geometry features (e.g., distances, contacts, torsions) of our simulation data and pass

them through the architecture presented in Fig. 6.2. The data is fed through an attention mechanism (represented by the matrix $\mathbf{G}$) that yields subsystem specific vectors $\mathbf{Y}_t^i$, each of which attends to features relevant for subsystem i. These vectors then serve as inputs to N parallel feature transformations $\boldsymbol{\eta}^i$ (parallel VAMPnets) which transform those into output features $\boldsymbol{\chi}^1, \ldots \boldsymbol{\chi}^N$ (with $\boldsymbol{\chi}^i(\mathbf{x}_t) = \boldsymbol{\eta}^i(\mathbf{Y}_t^i(\mathbf{x}_t))$) that represent slow collective coordinates or directly fuzzy assignments to metastable Markov states of each molecular subsystem. Equipped with the state assignments, we can compute correlation matrices (Eq. (6.9)) and derive a Koopman model matrix from those (Eq. (6.10)). As in VAMPnets, the feature transformations $\boldsymbol{\eta}^1, \ldots \boldsymbol{\eta}^N$ are represented by deep neural networks. In the present study we use multilayer perceptrons with a SoftMax output layer representing fuzzy state assignments. However, other architectures could be chosen, e.g. graph convolution networks when parameter sharing is desired [67, 68], and a linear output layer could be chosen if the aim is to represent slow collective variable rather than discrete states [57, 58]. The parameters of the feature transformations $\boldsymbol{\eta}$ and the attention matrix are learned end-to-end via backpropagation.

In more detail, given N individual subsystem models, the global system state can be given by the Kronecker product of all subsystem states:

$$\boldsymbol{\chi}^G(\mathbf{x}_t) = \bigotimes_i \boldsymbol{\chi}^i(\mathbf{x}_t) \tag{6.4}$$

and by computing the global correlation matrices $(\mathbf{C}_{00}^G, \mathbf{C}_{0\tau}^G, \mathbf{C}_{\tau\tau}^G)$ from Eqs. (6.9) using $\boldsymbol{\chi}^G$. We note that this step does not require that we have independent Markovian models, but it is simply a formalism to express global states in terms of a combination of local states.

Furthermore, we construct a candidate for the global Koopman model from the subsystem models by combining the individual singular values and vectors with a Kronecker product [64]:

$$\hat{\mathbf{K}}^G = \bigotimes_i \mathbf{K}^i \qquad \hat{\mathbf{U}}^G = \bigotimes_i \mathbf{U}^i \qquad \hat{\mathbf{V}}^G = \bigotimes_i \mathbf{V}^i. \tag{6.5}$$

The matrices $\hat{\mathbf{U}}^G$ and $\hat{\mathbf{V}}^G$ map the global state assignments onto the constructed singular functions and are computed from the local matrices as defined in Eqs. (6.11)-(6.12). The diagonal matrix $\hat{\mathbf{K}}^G$ encodes the singular values and is computed from the subsystem singular value matrices via Eq. (6.10).

In order to evaluate the performance of the constructed model to predict the dynamics in the global state space, the VAMP-E validation [55] score can be exploited,

$$\mathcal{R}_E^G = \mathrm{tr}[2\hat{\mathbf{K}}^G(\hat{\mathbf{U}}^G)^\top \mathbf{C}_{0\tau}^G \hat{\mathbf{V}}^G - \hat{\mathbf{K}}^G(\hat{\mathbf{U}}^G)^\top \mathbf{C}_{00}^G \hat{\mathbf{U}}^G \hat{\mathbf{K}}^G(\hat{\mathbf{V}}^G)^\top \mathbf{C}_{\tau\tau}^G \hat{\mathbf{V}}^G]. \tag{6.6}$$

The VAMP-E score measures the difference between the estimated Koopman model and the true dynamics. Here, it is evaluated for the global state assignments $\bigotimes_i \chi^i$ (as encoded in $\mathbf{C}_{00}^G, \mathbf{C}_{0\tau}^G, \mathbf{C}_{\tau\tau}^G$) mapped on the constructed singular functions (as encoded in $\hat{\mathbf{U}}^G, \hat{\mathbf{V}}^G$). If the subsystems are independent the constructed singular functions are optimal and the singular values of the global system are indeed the product of singular values of the subsystems (as formalized in Conditions for independent systems, also see Supplementary Note 1). In this case, the global VAMP-E score Eq. (6.6) has a product form

$$\mathcal{R}_E^G = \prod_i \mathcal{R}_E^i \tag{6.7}$$

that poses a necessary condition for subsystem independence.

To finally train the model, we develop a loss function that (i) maximizes the global VAMP-E score, assuming that they describe independent dynamics (Eqs. (6.4)-(6.6)) , and (ii) minimizes a term that penalizes statistical dependence between these subsystems (Eqs. (6.7)) scaled by a weighting factor ξ.

We evaluate the scores only pairwise, to escape the growth of the global state space, and sum over all possible pairs i,j:

$$L = -\sum_{i<j} \mathcal{R}_E^{ij} + \xi \sum_{i<j} \frac{||\mathcal{R}_E^{ij} - \mathcal{R}_E^i \mathcal{R}_E^j||}{\mathcal{R}_E^{ij}}. \tag{6.8}$$

Here, $\mathcal{R}_E^{ij}$ measures the quality of the constructed Koopman model of subsystems i and j and is computed using Eq. (6.6). The weighting factor ξ is a hyperparameter that should be chosen large enough to find decoupled systems and small enough to not interfer with the subsystem dynamics. Even though the choice of an appropriate ξ depends on the nature of the dynamics and the coupling, it is directly related to the training procedure as it, briefly, balances focus of the optimizer between kinetics and decoupling. Further conditions (Eq. (6.18)), which evaluate the independence of the singular functions and values, can be used as post training validation metrics for adjusting ξ and for testing to which degree dynamically independent subsystems were found.

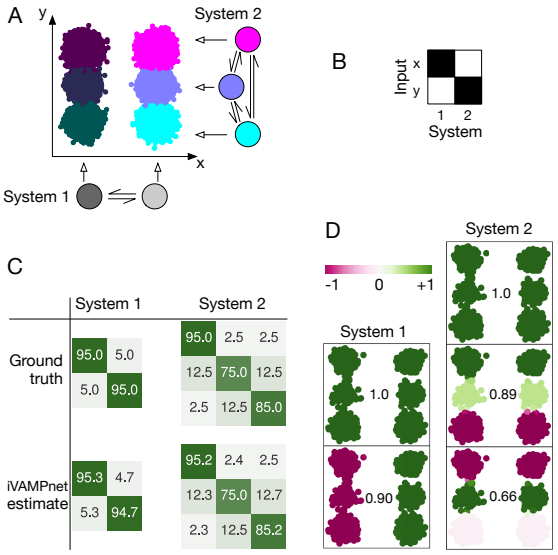


Figure 6.3: Hidden Markov state model as a benchmark example for independent subsystems: **(a)** 2 subsystems with 2 and 3 states emit independently to an x and y axis, respectively. The corresponding 2D space embeds all 6 global states. **(b)** The learned mask shows that each subsystem focuses on one input dimension. **(c)** The estimated subsystem transition matrices are compared with the ground truth (in percent). **(d)** Subsystem eigenfunctions and corresponding eigenvalues as found by iVAMPnet. Independent processes are recovered from the 2D data.

6.2.3 Benchmark model with two independent subsystems

The iVAMPnet architecture, which is implemented using PyTorch [69], is depicted in Fig. 6.2. Generally, various neural network architectures are possible; we here choose fully connected feed forward neural networks with up to 5 hidden layers with 100 nodes each. The scripts to reproduce the results including the details for the training routine, choice of hyper-parameters, and network architecture can be found in our GitHub repository. We note that an implementation of VAMPnets is available in the current version of DeepTime [70].

We first demonstrate that iVAMPnets are capable of decomposing a dynamical system into its independent Markovian subsystems based on observed trajectory data using an exactly decomposable benchmark model (Fig. 6.3).

Akin to the protein illustrated in Fig. 6.1, we define a system that consists of two independent subsystems with two and three states, respectively. It is modeled by two transition matrices with the corresponding number of states. We sample a discrete trajectory with each matrix (100k steps) [70]. The global state is defined as a combination of these discrete states. The discrete subsystem states are now interpreted as the hidden states of hidden Markov models [71] that emit to separate, subsystem-specific dimensions of a 2D space. The output of each subsystem is modeled with Gaussian noise

$N(\mu_i, \tilde{\sigma}) \in \mathbb{R}$ that is specific to the state that the system is in, specified by the mean μ_i, and a constant $\tilde{\sigma}$. The two state subsystem therefore describes a jump process between Gaussian basins along the x-axis and the three state subsystem along the y-axis, respectively (Fig. 6.3a). These variables compare to collective variables of the green (x) and blue (y) system depicted in Fig. 6.1. Please note that while in this benchmark system the relevant slow collective variables are known, iVAMPnets are generally capable of finding them (cf. 10D hypercube benchmark model and Synaptotagmin-C2A).

Since the generative benchmark model consists of perfectly independent subsystems and the pair already describes the global system, our method can simply be optimized for the global VAMP-E score (Eq. (6.6)) without the need for any further constraints. We train a model with a two and three state subsystem at a lag time of $\tau = 1$ step.

Once trained, the iVAMPnet yields a model of the dynamics in each of the identified subsystems. As expected, we find that the estimated transition matrices for both subsystems closely agree with the ground truth (Fig. 6.3c). To additionally assess the slow subsystem dynamics in more detail, we borrow concepts from MSM analysis and conduct an eigenvalue decomposition of the iVAMPnet models (cf. VAMPnets). The analysis of the eigenfunctions demonstrates that, by construction, the system exhibits one independent process along the x-axis ($\lambda_1 = 0.90$) and two along the y-axis ($\lambda_2 = 0.89$ and $\lambda_4 = 0.66$) (Fig. 6.3d). In contrast, we note that in the picture of global states, two additional processes would appear as a result of mixing the independent processes (cf. Supplementary Note 2), which makes the combined dynamical model more challenging to analyze, whereas the iVAMPnet analysis remains straightforward and simple.

Besides the dynamical models, our iVAMPnet yields assignments between input features and subsystems. We find that the method correctly identifies the two state system as the x-axis and the three states as the y-axis feature, respectively (Fig. 6.3b).

6.2.4 10D hypercube benchmark model

In a next step we test the iVAMPnet approach with ten 2-state subsystems, which corresponds to 1024 global states (Fig. 6.4a,b). As before, the dynamics is generated by ten independent hidden Markov state models with unique timescales. The system is split into five pairs of subsystems, and the two coordinates governing the transition dynamics of each pair are rotated in order to make them more difficult to separate (Fig. 6.4a). Additionally, we make the learning problem harder by adding ten noise dimensions such that the global system lives on a 10-dimensional hypercube embedded in a 20 dimensional space.

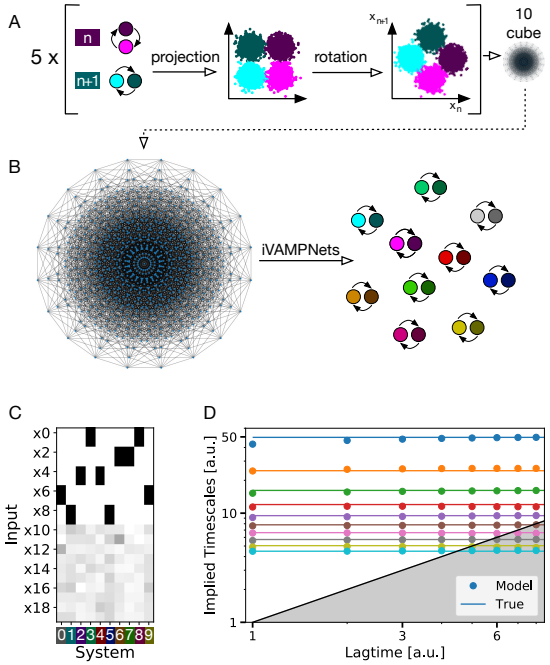


Figure 6.4: Hidden Markov state model with 1024 global states forming a 10D hypercube embedded in a 20D space. **(a)** The hypercube is composed of ten independent 2-state subsystems. A pair of two subsystems always lives in a common rotated 2D-manifold. Therefore, two subsystems need the same input features to be well approximated. **(b)** 2D depiction of the hypercube in an orthographic projection [72, 73], where the global system can jump freely between all 1024 vertices, and the ten 2-state models retrieved from it by the iVAMPnet. **(c)** Learned mask shows that for each subsystem, the network assigns two highly important input features which are shared with exactly one other subsystem, mirroring the rotated input space. Noise dimensions (x10-x19) are assigned low importance values. **(d)** Implied timescales of all ten subsystems learned by our method (dots) approximate the underlying true timescales (lines).

Although the subsystems are perfectly independent, we will estimate an iVAMPnet with the VAMP-E score in a pairwise fashion, thereby avoiding to estimate expensively large correlation matrices in $\mathbb{R}^{1024\times1024}$. As this is only justified if all systems are independent, we additionally enforce Eq. (6.7) during training by minimizing Eq. (6.8) and thereby rule out that any two subsystems approximate the same process.

The iVAMPnet estimation yields subsystem models which, as common in MSM analysis, can be validated by testing whether their implied relaxation timescales are converged in the model lag time τ. We find that the implied timescales learned by the iVAMPnet are indeed converged and accurately reproduce the ground truth (Fig. 6.4d). We note that in addition to the timescales of the individual subsystems that are identified by the iVAMPnet, a global model would also contain all timescales that result from products of eigenvalues, resulting in a total of 1024 timescales. Thus, the iVAMP-

net analysis provides a much simpler and more concise model than a global MSM or VAMPnet would.

Furthermore, the subsystem assignment mask indicates that the method correctly assigns high importance weight to two input features for each model (Fig. 6.4c). Therefore, the method proves its capability of decomposing a noisy, high dimensional global system into its independent sub-processes in a data efficient way.

We have generalized the 10-cube system to a variable number of subsystems (N-cube) to conduct a performance benchmark, finding that iVAMPnets outperform VAMPnets for this particular system. We however note that this result may not be generalizable to arbitrary systems as the N-cube features truely independent 2-state subsystems (compare Supplementary Note 6 for details).

6.2.5 Synaptotagmin-C2A

Finally, we test iVAMPnets on an all-atom protein system. In comparison to our benchmark examples, we expect the underlying global dynamics to be only approximately decomposable into independent subsystems. Our test data consists of 184 µs aggregate MD data of each 2 µs length ($92 \times 2\,\mu s$) of the C2A domain of synaptotagmin (Supplementary Note 7) that was described previously [74]; synaptotagmin plays a crucial role in the regulation of neurotransmitter release [75]. It was shown to consist of approximately uncoupled subsystems containing the calcium binding region (CBR) and the C78 loop, respectively [64].

First, we attempted to model the protein with a global model, i.e., with a single (regular) VAMPnet. Indeed, this approach failed because there were not enough simulation statistics to estimate a reversibly connected transition model between all global metastable states, resulting in diverging implied timescales (Supplementary Note 3 and Supplementary Fig. D.2). This is exactly the scenario where iVAMPnets should provide an advantage, by only relying on locally rather than globally converged transition statistics.

Next, we train an iVAMPnet to seek two subsystems of twelve and six states, respectively, each at a lag time of $\tau = 10\,\text{ns}$ where we enforce constraint Eq. (6.7) to find uncoupled subsystems.

The trained iVAMPnet identifies one subsystem comprising all three CBR loops (CBR-1, CBR-2, CBR-3; Fig. 6.5a). The second subsystem consists not only of the aforementioned C78 loop but also of the loop connecting beta sheets 3 and 4 [76] (termed C34 henceforth). When mapping the residue positions on the protein structure it becomes obvious that the two subsystems are physically well separated (Fig. 6.5a), supporting the conclusion that both regions are only weakly coupled [64].

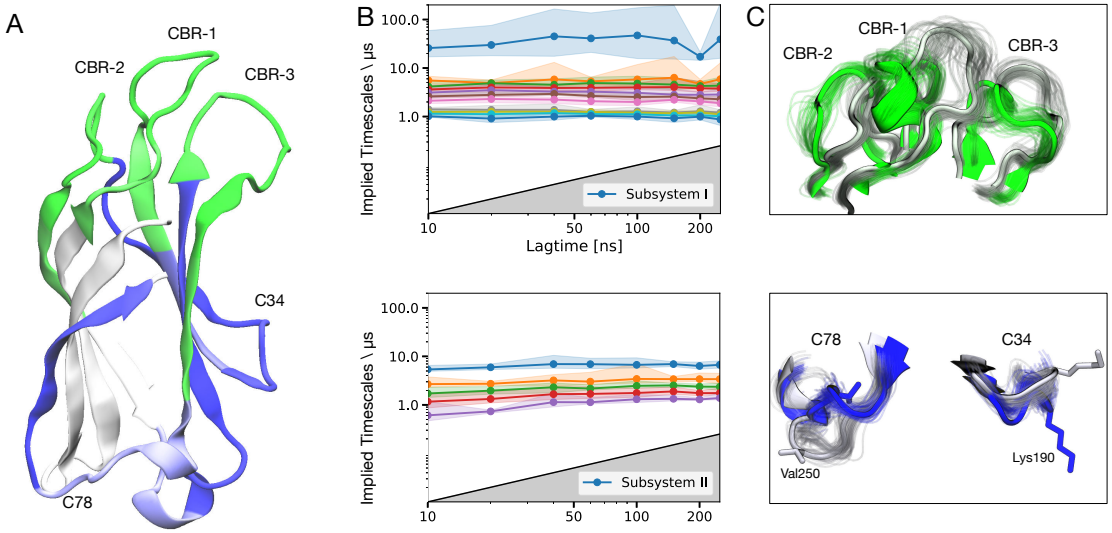


Figure 6.5: iVAMPnet of synaptotagmin-C2A with two subsystems and twelve and six states, respectively. **(a)** Importance values of the trainable mask depicted as color-coded protein secondary structure, indicating assignment to subsystem I (II) in green (blue). **(b)** Implied timescales of the two subsystems with a 90% percentile over 20 runs. **(c)** Superposed representative structures of both extrema of the slowest resolved eigenfunctions of each subsystem (residues not assigned a high importance value or not showing significant movement are omitted for clarity). The slowest process of subsystem 1 changes between green and gray structures showing an orchestrated movement of the full Calcium Binding Region (CBR1, CBR2, and CBR3). The slowest process of the second subsystem occurs between the blue and gray structures and describes a combined movement of C78 and C34.

The implied timescales of both systems are approximately constant in the model lag time τ. Most timescales are in the range of $1 - 10\,\mu s$, with the exception of one much slower process with a $100\,\mu s$ relaxation time found in the first subsystem (Fig. 6.5b), which has not been found previously. Analysis of the structural changes governing this process reveals that it involves an orchestrated transition of all CBR loops (Fig. 6.5c). Such a process could however not be resolved by the previous study [74] where the CBR was modeled as individual loops. The process of the second system involves a simultaneous movement of the C78 and C34 loops (Fig. 6.5c).

iVAMPnets find metastable structures in the local features that are comparable to the ones described in our previous work [74]. Specifically, α-helices in two distinct positions and a state burying a methionine residue (Met173) can be found in the CBR1. In the adjacent CBR2 site, both tightly bound and loose configurations are identified, and the C78 site features all three previously described valine residue conformations (Val250, Val255). In addition to the features modeled in our preceding study [74], iVAMPnets identify dynamics in a lysine rich cluster (Lys189-192) that was previously reported as important for membrane interaction [77]. Please compare Supplementary Note 4 for a detailed view on the metastable states and exchange kinetics. In contrast

to our previous work, the kinetic models in the local subsystems are more complex and incorporate a larger number of dynamic processes, providing a more comprehensive picture without the need to define a partitioning manually. In fact, conducting domain-decomposition and local kinetic modeling simultaneously has enabled the identification of very subtle dynamical features as long as they contribute significantly to the local VAMP-scores.

Although estimating a global VAMPnet model for synaptotagmin was not feasible given the sparse data sample, iVAMPnets use the same data efficiently and estimate a statistically valid dynamical model. This result is especially striking because the iVAMP-net approach also simplifies the subsequent task of interpreting models by separating dynamically independent protein domains.

6.2.6 Counterexample: folding of the villin miniprotein

Finally, we conducted an experiment on a villin protein folding trajectory of 125 µs length [78] as a negative example (Supplementary Note 7). Small proteins such as villin are typically cooperative, i.e., the slowest processes related to folding involve all residues (Supplementary Note 5). Thus, these processes cannot be resolved when decomposing the system into several subsystems. Indeed, we find that a splitting into two subsystems with two states each results in timescales that are not converged, and whose relaxation processes approximate a partial folding on disjoint areas (cf. Supplementary Fig. D.6).

6.2.7 Testing statistical independence of the learned dynamical subsystems

As constraint Eq. (6.7) was used as a penalty during training (as independence score Eq. (6.19)), we assess the validity of an estimated subsystem assignment by evaluating the constraints that were not enforced during training (Eq. (6.17)) as post-training independence scores M_U, M_V, and M_{UV} (defined in Eq. (6.18)). Low values for M_U and M_V imply that the constructed left and right singular functions are indeed valid candidates for singular functions in the global state space. A small value for M_{UV} indicates that the kinetics in the global state space is well predicted by the Kronecker product of subsystem models. We find that the three metrics are well suited to indicate independence of the learned subsystems (Tab. 6.1). Out of the tested systems only villin cannot be split into independent parts (all scores > 0.1). In comparison, the benchmark models and synaptotagmin can be decomposed into statistically uncoupled subsystems (all scores

	M_U	M_V	M_{UV}	M_R
Benchmark 2	0.0058	0.0059	0.0055	0.0002
10-Cube	0.0039	0.0039	0.0046	0.0005
Synaptotagmin	0.0042	0.0042	0.0044	0.0018
Villin	0.1353	0.1364	0.1493	0.0021

Table 6.1: Post-training independence validation. The scores in columns 1-3 (M_U, M_V, M_{UV}, cf. Eq. (6.18)) are computed from independence constraints that were not enforced during the training. The score in the last column (M_R) is used during the training and shown for reference. The three post-training validation scores M_U, M_V, and M_{UV} indicate that the final subsystems of both benchmark examples and synaptotagmin are indeed independent, whereas the scores for villin strictly oppose this conjunction. The standard deviations (SD) over 10 different runs are on the order of 10^{-5} for all systems except villin, which has an SD $\sim 10^{-4}$.

< 0.01). The slightly increased M_R-value for synaptotagmin suggest that its subsystems might be weakly coupled.

6.3 Discussion

We have proposed an unsupervised deep learning framework that, using only molecular dynamics simulation data, learns to decompose a complex molecular system into subsystems which behave as approximately independent Markov models. Thereby, iVAMPnet is an end-to-end learning framework that points a way out of the exponentially growing demand for simulation data that is required to sample increasingly large biomolecular complexes.

Specifically, we have developed and demonstrated iVAMPnets for molecular dynamics, but the approach is, in principle, also applicable to different application areas, such as fluid dynamics. The specific implementation, such as the representation of the input vectors $\mathbf{x}_t$ and the neural network architecture of the χ-functions, depend on the application and can be adapted as needed.

We now have a hierarchy of increasingly powerful models ranging from MSMs over VAMPnets to iVAMPnets. MSMs always consist of (1) a state space decomposition and (2) a Markovian transition matrix governing the dynamics between these states. VAMPnets provide a deep learning framework for MSMs, and thereby (3) learn the collective coordinates in which the state space discretization (1) is best made. iVAMPnets additionally learn (4) a physical separation of the molecular system into subsystems, each of which has its own slow coordinates, Markov states, and transition matrix.

We have demonstrated that iVAMPNets are a powerful multiscale learning method that succeeds in finding and modeling molecular subsystems when these subsystems in-

deed evolve statistically independently. Additionally, iVAMPnets are capable of learning from high dimensional MD data. To prove that point, we have demonstrated that the synaptotagmin C2A domain is decomposable into two almost independent Markov state models. Importantly, we have shown that this dynamical decomposition of synaptotagmin C2A succeeds while an attempt to model the system with a global Markov state model fails due to poor sampling. This is a direct demonstration that iVAMPnets are statistically more efficient than VAMPnets, MSMs or other global-state models and may indeed scale to much larger systems.

We note, however, that iVAMPnets do not learn how the subsystems are coupled, and are therefore, in their current form, only applicable to molecular systems that consist of uncoupled or weakly coupled subsystems. Although most biomolecular complexes are known to be cooperative, there are examples that have been modeled very successfully using independent subsystems, such as the Hudgkin-Huxley model of voltage gated channel proteins [79, 80]. For other systems, the degree of coupling is a matter of debate, for example the C2-tandem (C2A and C2B domains) in synaptotagmins [81, 82]. Since isolated domains are known to conduct function by themselves in many cases, we believe that discarding couplings is a first-order modeling assumption that is suitable to identify these domains and their relevant metastable states.

Following up on Ref. [63] and introducing coupling parameters that describe how the learned MSMs are coupled, is subject to ongoing research. Furthermore, the weak-coupling assumption is made for the time-scale of the investigated molecular processes and may not be generalizable to arbitrary times. E.g., the degree of coupling between domains found in an MD simulations of a folded protein state may be very different in its unfolded state, which will be eventually encountered for a long enough simulation time.

Besides the usual hyperparameter choices in deep learning approaches, iVAMPnets require the specification of the number of sought subsystems. This choice can be guided by training an iVAMPnet for different numbers of subsystems and then interrogating the independence scores (Eq. (6.19) and Eqs. (6.18)) to choose a decomposition where statistical independence is optimal. We suggest to start with decomposing the system into two subsystems as a starting point, and to increase this number subsequently. Non-optimal choices may, e.g., reflect in non-converged implied timescales (possibly an incarnation of the sampling problem that may be mitigated by increasing the number of subsystems) or high independence scores (not possible to split the system because too many or non-optimal number of subsystems were chosen). Furthermore, the choice of the number of subsystems can be guided by the number of structural domains in a

protein (complex) or by using the network-based approach presented in Ref. [64]. Furthermore, the number of states in each subsystems needs to balance a) the quality of the singular function approximation (higher for few states) and b) model resolution (higher for more states). Ultimately, different choices may yield converged validation measures, and the number of states may be chosen to yield the desired model resolution in this case.

iVAMPnets can be improved and further developed in multiple ways, e.g. by employing more advanced network architectures, e.g. graph neural networks, where parameters could be shared across subsystems. This might result in higher quality models and a greater robustness against the hyperparameter choice. Very recently, graph neural networks were indeed successfully combined with VAMPnets, showing that the resulting method (GraphVAMPnets) is applicable to MD data and that the estimated models are high quality [83].

In summary, iVAMPnets pave a possible path for modeling the kinetics of large biological systems in a data-efficient and interpretable manner.

6.4 Methods

6.4.1 VAMPnets

Since an iVAMPnet implements multiple parallel VAMPnets representing the kinetics of separate independent subsystems, we will introduce VAMPnets first [56]. VAMPnets are multilayer perceptrons that represent feature functions χ (we omit the upper subsystem index i for the sake of clearness here). Their last layer is often chosen to be a SoftMax function, i.e., summing over all non-negative outputs yields a 1. Therefore, the output of a VAMPnet can be interpreted as a fuzzy assignment to a metastable state. Taking the linear combination of states with equal weights results in the constant singular function with the singular value 1, which will be reflected by the singular values of the Koopman matrix (Eq. (6.10) with the normalized correlation matrix). Given the feature functions χ, we can compute the following correlation matrices:

$$\begin{aligned}
\mathbf{C}_{00} &= \frac{1}{L}\sum_t \chi(\mathbf{x}_t)\chi(\mathbf{x}_t)^\top \\
\mathbf{C}_{0\tau} &= \frac{1}{L}\sum_t \chi(\mathbf{x}_t)\chi(\mathbf{x}_{t+\tau})^\top \\
\mathbf{C}_{\tau\tau} &= \frac{1}{L}\sum_t \chi(\mathbf{x}_{t+\tau})\chi(\mathbf{x}_{t+\tau})^\top,
\end{aligned} \tag{6.9}$$

where L is the number of collected data pairs in the simulations.

Training VAMPnets or iVAMPnets involves the computation of covariance matrices over minibatches. We therefore need to choose the batchsize to balance large estimator variance obtained for small batches and high memory requirements for large batches. Instead of using the trivial covariance estimator (Eqs. (6.9)) which is asymptotically unbiased [55] but has a high-variance, one can employ a shrinkage estimator [84, 85] which reduces the overall estimator error by trading larger bias for lower variance. For the current study, we assume that our benchmark and MD data has been sufficiently sampled to yield adequate approximations with the estimator given in Eqs. (6.9).

The approximation of the singular functions and values can be estimated via the singular value decomposition (SVD) of the following matrix $\bar{\mathbf{K}}$:

$$\bar{\mathbf{K}} = \mathbf{C}_{00}^{-1/2}\mathbf{C}_{0\tau}\mathbf{C}_{\tau\tau}^{-1/2} = \mathbf{A}\mathbf{K}\mathbf{B}^{\top} \tag{6.10}$$

$\mathbf{K}$ is the diagonal matrix of approximated singular values corresponding to the left and right singular functions:

$$\mathbf{f}^{\top}(\mathbf{x}_t) = \chi(\mathbf{x}_t)^{\top}\mathbf{U} = \chi(\mathbf{x}_t)^{\top}\mathbf{C}_{00}^{-1/2}\mathbf{A} \tag{6.11}$$

$$\mathbf{g}^{\top}(\mathbf{x}_{t+\tau}) = \chi(\mathbf{x}_{t+\tau})^{\top}\mathbf{V} = \chi(\mathbf{x}_{t+\tau})^{\top}\mathbf{C}_{\tau\tau}^{-1/2}\mathbf{B}. \tag{6.12}$$

The matrices $\mathbf{U}$ and $\mathbf{V}$ construct the left and right singular functions from the individual state assignments. The optimal state assignments can be found by maximizing the VAMP-E score:

$$\mathcal{R}_E = \mathrm{tr}[2\mathbf{K}\mathbf{U}^{\top}\mathbf{C}_{0\tau}\mathbf{V} - \mathbf{K}\mathbf{U}^{\top}\mathbf{C}_{00}\mathbf{U}\mathbf{K}\mathbf{V}^{\top}\mathbf{C}_{\tau\tau}\mathbf{V}]. \tag{6.13}$$

Given trained state assignments $\chi(\mathbf{x}_t)$ and correlation matrices Eq. (6.9), the Koopman matrix $\mathbf{T}$ can then be evaluated as:

$$\mathbf{T} = \mathbf{C}_{00}^{-1}\mathbf{C}_{0\tau}. \tag{6.14}$$

Furthermore, we can estimate the eigenfunction φ and timescales t_i by its eigendecomposition $\mathbf{T} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}$:

$$\varphi(\mathbf{x}) = \mathbf{Q}^{\top}\chi(\mathbf{x}), \tag{6.15}$$

$$t_i = \frac{-\tau}{\log(|\Lambda_{ii}|)}. \tag{6.16}$$

Please note that this operation is only possible if the eigendecomposition is (approximately) real-valued, a condition that is met for the presented application cases.

6.4.2 Conditions for independent systems

For Markov independent systems, the singular values and functions that are constructed by the Kronecker product match the true global ones,

$$\begin{aligned}(\hat{\mathbf{U}}^G)^\top \mathbf{C}_{00}^G \hat{\mathbf{U}}^G &= 1 \\ (\hat{\mathbf{V}}^G)^\top \mathbf{C}_{\tau\tau}^G \hat{\mathbf{V}}^G &= 1 \\ (\hat{\mathbf{U}}^G)^\top \mathbf{C}_{0\tau}^G \hat{\mathbf{V}}^G &= \hat{\mathbf{K}}^G,\end{aligned} \tag{6.17}$$

where the first two equations guarantee the orthonormality of the constructed singular functions. The latter verifies that the left and right singular functions correlate as predicted by the Kronecker product of the singular values.
These conditions can be translated to the following scores:

$$\begin{aligned}M_U &= |(\hat{\mathbf{U}}^G)^\top \mathbf{C}_{00}^G \hat{\mathbf{U}}^G - 1| \\ M_V &= |(\hat{\mathbf{V}}^G)^\top \mathbf{C}_{\tau\tau}^G \hat{\mathbf{V}}^G - 1| \\ M_{UV} &= |(\hat{\mathbf{U}}^G)^\top \mathbf{C}_{0\tau}^G \hat{\mathbf{V}}^G - \hat{\mathbf{K}}^G|\end{aligned} \tag{6.18}$$

Furthermore, using the identities Eq. (6.17) and the definition of the VAMP-E score Eq. (6.13) yields

$$M_R = \frac{|\mathcal{R}_E^G - \prod_i \mathcal{R}_E^i|}{\mathcal{R}_E^G}. \tag{6.19}$$

The norms denote simple means. The last score, M_R, is enforced during training in a pairwise fashion (cf. Eq. (6.8)).

6.4.3 Network architecture

Given a global system, which we want to decompose into N subsystems, and a time series of input features $\{\mathbf{x}_t\}_{t=1,\dots T}$, $\mathbf{x}_t \in \mathbb{R}^{D\times 1}$, we pass the features through a mask $\mathbf{G} \in \mathbb{R}^{D\times N}$, which weights each input differently for each subsystem, before the result are transformed individually by the N independent state assignment functions $\boldsymbol{\eta}^i$. It should be mentioned that the mask is merely introduced for interpretability reasons and is not essential to find independent subsystems. If the mask was omitted, the extraction

of the relevant features would simply be transferred to the downstream neural networks, remaining hidden to the practitioner.

The weighted input is assessed by an element wise multiplication $\bar{\mathbf{Y}}_t = \mathbf{G} \odot \mathbf{x}_t$. In order to prevent the neural networks to reverse the weighting of the mask in its consecutive layers, we draw for each input feature i and subsystem j an independent, normally distributed random variable $\epsilon_{ij} \sim \mathcal{N}(0, \sigma(1 - G_{ij}))$. This noise is added to the weighted features:

$$\mathbf{Y}_t = \bar{\mathbf{Y}}_t + \boldsymbol{\epsilon}. \tag{6.20}$$

Thereby, the attention weight linearly interpolates between input feature and Gaussian noise, i.e., if the attention weight $G_{ij} = 1$, Y_{ij} carries exclusively the input feature x_i, if $G_{ij} = 0$, Y_{ij} is simple Gaussian noise. By tuning the noise scaling σ, a harder assignment by $\mathbf{G}$ can be enforced. This hyperparameter should be optimized by adjusting it so that the resulting mask yields clear subsystem assignments without being binary. Subsequently, the transformed feature vector is split for each individual subsystem $\mathbf{Y}_t = [\mathbf{Y}_t^1, ..., \mathbf{Y}_t^N]$ and passed through the subsystem specific neural network $\boldsymbol{\eta}^i$ resulting in feature transformations $\boldsymbol{\chi}^i(\mathbf{x}_t) = \boldsymbol{\eta}^i(\mathbf{Y}_t^i)$. These features are then used to estimate the Koopman models.

The training framework and neural network architecture were implemented in the Python 3 programming language using numpy [86] and pyTorch [69]; benchmark system data was generated using DeepTime [70]; data visualization was performed using matplotlib [87] and VMD [24].

6.4.4 Constructing the mask

To train an interpretable mask, we use the following three premises:

1. A single subsystem should not focus on all input features.
2. Different subsystems compete for high weights for the same feature.
3. All weights should be in the range $[0, 1]$ and the matrix should be sparse.

Therefore, the mask is constructed by trainable weights $\mathbf{g} \in \mathbb{R}^{D \times N}$ which are first processed by a softmax function which normalizes along the input feature axis $\mathbf{g}_1 = \text{softmax}(\mathbf{g}, \text{dim=0})$. Thereby, if a subsystem focuses on one part of the features, a lower weight for the other parts is expected following the first premise.

In a next step, weights which are lower than a threshold θ are clipped to zero $\mathbf{g}_2 = \text{relu}(\mathbf{g}_1 - \theta)$ to guarantee sparsity. The threshold θ is a hyperparameter that can be optimized by starting with comparably small values (i.e., very little cutoff) and subsequently

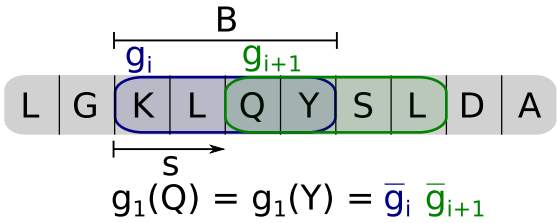


Figure 6.6: Attention scheme for amino acid chain. Windows of size B are placed along the chain with a step size of s resulting into W many windows. A trainable weight $\mathbf{g} \in \mathbb{R}^{W\times N}$ is assigned for a window in each subsystem which are made positive and normalized along the window axis through a softmax $\bar{\mathbf{g}} = \mathrm{softmax}(\mathbf{g}, \mathrm{dim}{=}0)$. Here a window size of $B = 4$ and a step size of $s = 2$ is chosen. As a consequence the weight of the amino acid glutamine (Q) is given as the product of the two windows it is part of $\mathbf{g}_1(Q) = \bar{\mathbf{g}}_i\bar{\mathbf{g}}_{i+1}$. The choice of the step size determines how many neighboring amino acids have the exact same weight within a subsystem, which applies here for the tyrosine (Y). Together with the window size it is regulated how many residues share parts of their weights. Hence, the serine (S) shares the weight $\bar{\mathbf{g}}_{i+1}$ with the previous two amino acids $\mathbf{g}_1(S) = \bar{\mathbf{g}}_{i+1}\bar{\mathbf{g}}_{i+2}$, which has a smoothing effect on the attention mechanism along the chain.

increasing it without further training – a reasonable cutoff does not alter the results in this case, as the downstream neural networks still obtain all relevant information.

Since input features could be negligible for all subsystems, a dummy system is added which has a constant value $\mathbf{c} \in \mathbb{R}^{D\times 1}$ for all features $\mathbf{g}_3 = [\mathbf{g}_2, \mathbf{c}]$. Consequently, the weights of all subsystems and the dummy system are normed for each feature $\mathbf{g}_4 = \mathbf{g}_3/\mathrm{sum}(\mathbf{g}_3, \mathrm{dim} = 1)$, which together with the clipping fulfills the premises two and three.

Finally, the mask is given by truncating the dummy system $\mathbf{g}_4 = [\mathbf{G}, \bar{\mathbf{c}}]$. Beware that only $\mathbf{g}_4$ is normalized along the system axis.

6.4.5 Application to protein dynamics

Since for proteins the final model is often expected to be invariant with respect to rotations and translations, internal coordinates are employed as input features. For Markov state modeling, the minimal heavy atom distance d_{ij} between residues i,j has been proven to be a good descriptor [56, 88]. However, for interpretability, mask weights for each residue are preferable. Therefore, the mask is of size $\mathbf{G} \in \mathbb{R}^{R\times N}$ with the number of residues R. The input features are then scaled as $x_{ij} = G_iG_j \exp(-d_{ij})$.

Furthermore, a smoothing routine is implemented such that neighboring residues along the chain have similar importance weights. W windows of size B are placed along the chain with step size s. Each window has a trainable weight $\mathbf{g} \in \mathbb{R}^{W\times N}$. Consequently, the softmax function is taken along the window axis $\bar{\mathbf{g}} = \mathrm{softmax}(\mathbf{g}, \mathrm{dim}{=}0)$. However, before applying the clipping as before the weight for each residue $\mathbf{g}_1 \in \mathbb{R}^{R\times N}$ is calculated as the product of all window weights the residue is part of (Fig. 6.6).

Data availability

The benchmark data can be generated from the Jupyter notebooks that have been deposited on GitHub under `https://github.com/markovmodel/ivampnets` [89]. The molecular dynamics data set of synaptotagmin C2A have been deposited in Zenodo under `https://zenodo.org/record/6908073` [90]. The crystal structure of synaptotagmin C2A is available under PDB ID 2R83. The villin headpiece folding data are available under restricted access and were used under license for this study as courtesy of D.E. SHAW research [78], access can be obtained from the authors upon request.

Code availability

The code that implements the presented models and reproduces the presented results has been deposited on GitHub under `https://github.com/markovmodel/ivampnets` [89].

Ackowledgements

We acknowledge financial support from Deutsche Forschungsgemeinschaft DFG (SFB/TRR 186, project A12 to TH, FN, CC; SFB 958, project A04 to AM, FN; SFB 1114, projects C03 to FN, A04 to FN, CC, B03 to CC; SFB 1078, project C7 to CC; and RTG 2433 to FN, CC), the European Commission (ERC CoG 772230 “ScaleCell” to FN), the Berlin Mathematics center MATH+ (AA1-6 and AA1-10 to FN, CC), the BMBF (Research center BIFOLD to FN), the National Science Foundation (CHE-1900374, and PHY-2019745 to CC), the Welch Foundation (C-1570 to CC), and the Einstein Foundation Berlin (project 0420815101 to CC). We further thank Manuel Dibak and Moritz Hoffmann (FU Berlin) for fruitful discussions.

Author contributions

A.M., T.H. performed research (A.M. derived loss functions and implemented deep learning framework; T.H. designed method and developed test systems); A.M., T.H. analyzed data; A.M., T.H., C.C., F.N. designed research; A.M., T.H., C.C., F.N. wrote the paper.

Competing interests

The authors declare no competing interests.

Bibliography

[1] J. C. Phillips, D. J. Hardy, J. D. C. Maia, J. E. Stone, J. V. Ribeiro, R. C. Bernardi, R. Buch, G. Fiorin, J. Hénin, W. Jiang, R. McGreevy, M. C. R. Melo, B. K. Radak, R. D. Skeel, A. Singharoy, Y. Wang, B. Roux, A. Aksimentiev, Z. Luthey-Schulten, L. V. Kalé, K. Schulten, C. Chipot, and E. Tajkhorshid. "Scalable Molecular Dynamics on CPU and GPU Architectures with NAMD". *J. Chem. Phys.* 153.4 (2020), p. 044130.

[2] J. W. Vant, D. Sarkar, C. Gupta, M. S. Shekhar, S. Mittal, and A. Singharoy. "Molecular Dynamics Flexible Fitting: All You Want to Know about Resolution Exchange". *Protein Structure Prediction*. Springer, 2020, pp. 301–315.

[3] I. Buch, M. J. Harvey, T. Giorgino, D. P. Anderson, and G. De Fabritiis. "High-Throughput All-Atom Molecular Dynamics Simulations Using Distributed Computing". *J. Chem. Inf. Model.* 50.3 (2010), pp. 397–403.

[4] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, and V. S. Pande. "OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics". *PLOS Comput. Biol.* 13.7 (2017), e1005659.

[5] R. Salomon-Ferrer, A. W. Gotz, D. Poole, S. Le Grand, and R. C. Walker. "Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald". *J. Chem. Theory Comput.* 9.9 (2013), pp. 3878–3888.

[6] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl. "GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers". *SoftwareX* 1–2 (2015), pp. 19–25.

[7] G. Bussi, A. Laio, and P. Tiwary. "Metadynamics: A Unified Framework for Accelerating Rare Events and Sampling Thermodynamics and Kinetics". *Handbook of Materials Modeling*. Ed. by W. Andreoni and S. Yip. Cham: Springer International Publishing, 2020, pp. 565–595.

[8] S.-T. Tsai, Z. Smith, and P. Tiwary. "SGOOP-d: Estimating Kinetic Distances and Reaction Coordinate Dimensionality for Rare Event Systems from Biased/Unbiased Simulations". *J. Chem. Theory Comput.* 17.11 (2021), pp. 6757–6765.

[9] C. Liu, E. Brini, A. Perez, and K. A. Dill. "Computing Ligands Bound to Proteins Using MELD-Accelerated MD". *J. Chem. Theory Comput.* 16.10 (2020), pp. 6377–6382.

[10] J. L. MacCallum, A. Perez, and K. A. Dill. "Determining Protein Structures by Combining Semireliable Data with Atomistic Physical Models by Bayesian Inference". *Proc. Natl. Acad. Sci. U.S.A.* 112.22 (2015), pp. 6985–6990.

[11] A. Perez, J. L. MacCallum, and K. A. Dill. "Accelerating Molecular Simulations of Proteins Using Bayesian Inference on Weak Information". *Proc. Natl. Acad. Sci. U.S.A.* 112.38 (2015), pp. 11846–11851.

[12] Y. Ge and V. A. Voelz. "Estimation of Binding Rates and Affinities from Multiensemble Markov Models and Ligand Decoupling". *J. Chem. Phys.* 156.13 (2022), p. 134115.

[13] J. M. L. Ribeiro, P. Bravo, Y. Wang, and P. Tiwary. "Reweighted Autoencoded Variational Bayes for Enhanced Sampling (RAVE)". *The Journal of Chemical Physics* 149.7 (2018), p. 072301.

[14] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard. “A Direct Approach to Conformational Dynamics Based on Hybrid Monte Carlo”. *J. Comput. Phys.* 151.1 (1999), pp. 146–168.

[15] 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.* 134.17 (2011), p. 174105.

[16] W. C. Swope, J. W. Pitera, and F. Suits. “Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 1. Theory”. *J. Phys. Chem. B* 108.21 (2004), pp. 6571–6581.

[17] F. Noé, I. Horenko, C. Schütte, and J. C. Smith. “Hierarchical Analysis of Conformational Dynamics in Biomolecules: Transition Networks of Metastable States”. *J. Chem. Phys.* 126.15 (2007), p. 155102.

[18] 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”. *The Journal of Chemical Physics* 126.15 (2007), p. 155101.

[19] N.-V. Buchete and G. Hummer. “Coarse Master Equations for Peptide Folding Dynamics”. *J. Phys. Chem. B* 112.19 (2008), pp. 6057–6069.

[20] H. Wan and V. A. Voelz. “Adaptive Markov State Model Estimation Using Short Reseeding Trajectories”. *J. Chem. Phys.* 152.2 (2020), p. 024103.

[21] M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz, and F. Noé. “PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models”. *J. Chem. Theory Comput.* 11.11 (2015), pp. 5525–5542.

[22] M. P. Harrigan, M. M. Sultan, C. X. Hernández, B. E. Husic, P. Eastman, C. R. Schwantes, K. A. Beauchamp, R. T. McGibbon, and V. S. Pande. “MSMBuilder: Statistical Models for Biomolecular Dynamics”. *Biophys J.* 112.1 (2017), pp. 10–15.

[23] R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane, and V. S. Pande. “MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories”. *Biophys. J.* 109.8 (2015), pp. 1528–1532.

[24] W. Humphrey, A. Dalke, and K. Schulten. “VMD: Visual Molecular Dynamics”. *J. Mol. Graph.* 14.1 (1996), pp. 33–38.

[25] G. Pérez-Hernández, F. Paul, T. Giorgino, G. D. Fabritiis, and F. Noé. “Identification of Slow Molecular Order Parameters for Markov Model Construction”. *J. Chem. Phys.* 139.1 (2013), p. 015102.

[26] A. Ziehe and K.-R. Müller. “TDSEP — an Efficient Algorithm for Blind Separation Using Time Structure”. *ICANN 98*. Springer Science and Business Media, 1998, pp. 675–680.

[27] I. Mezić. “Spectral Properties of Dynamical Systems, Model Reduction and Decompositions”. *Nonlinear Dyn* 41.1-3 (2005), pp. 309–325.

[28] P. J. Schmid. “Dynamic Mode Decomposition of Numerical and Experimental Data”. *J. Fluid Mech.* 656 (2010), pp. 5–28.

[29] J. H. Tu, C. W. Rowley, D. M. Luchtenburg, S. L. Brunton, and J. N. Kutz. “On Dynamic Mode Decomposition: Theory and Applications”. *J. Comput. Dyn.* 1.2 (2014), pp. 391–421.

[30] F. Noé and C. Clementi. “Collective Variables for the Study of Long-Time Kinetics from Molecular Trajectories: Theory and Methods”. *Curr. Opin. Struct. Biol.* 43 (2017), pp. 141–147.

[31] S. Klus, F. Nüske, P. Koltai, H. Wu, I. Kevrekidis, C. Schütte, and F. Noé. “Data-Driven Model Reduction and Transfer Operator Approximation”. *J Nonlinear Sci* 28.3 (2018), pp. 985–1010.

[32] G. R. Bowman, V. S. Pande, and F. Noé, eds. *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*. Vol. 797. Advances in Experimental Medicine and Biology. Dordrecht: Springer Netherlands, 2014.

[33] B. E. Husic and V. S. Pande. “Ward Clustering Improves Cross-Validated Markov State Models of Protein Folding”. *J. Chem. Theo. Comp.* 13.3 (2017), pp. 963–967.

[34] F. K. Sheong, D.-A. Silva, L. Meng, Y. Zhao, and X. Huang. “Automatic State Partitioning for Multibody Systems (APM): An Efficient Algorithm for Constructing Markov State Models to Elucidate Conformational Dynamics of Multibody Systems”. *J. Chem. Theory Comput.* 11.1 (2015), pp. 17–27.

[35] M. Weber, K. Fackeldey, and C. Schütte. “Set-Free Markov State Model Building”. *J. Chem. Phys.* 146.12 (2017), p. 124133.

[36] G. R. Bowman, K. A. Beauchamp, G. Boxer, and V. S. Pande. “Progress and Challenges in the Automated Construction of Markov State Models for Full Protein Systems.” *J. Chem. Phys.* 131 (2009), p. 124101.

[37] B. Trendelkamp-Schroer, H. Wu, F. Paul, and F. Noé. “Estimation and Uncertainty of Reversible Markov Models”. *J. Chem. Phys.* 143.17 (2015), p. 174101.

[38] S. Kube and M. Weber. “A Coarse Graining Method for the Identification of Transition Rates between Molecular Conformations”. *J. Chem. Phys.* 126.2 (2007), p. 024103.

[39] Y. Yao, R. Z. Cui, G. R. Bowman, D.-A. Silva, J. Sun, and X. Huang. “Hierarchical Nyström Methods for Constructing Markov State Models for Conformational Dynamics”. *J. Chem. Phys.* 138.17 (2013), p. 174106.

[40] K. Fackeldey and M. Weber. “GenPCCA – Markov State Models for Non-Equilibrium Steady States”. *WIAS Rep.* 29 (2017), pp. 70–80.

[41] S. Gerber and I. Horenko. “Toward a Direct and Scalable Identification of Reduced Models for Categorical Processes”. *Proc Natl Acad Sci USA* 114.19 (2017), pp. 4863–4868.

[42] G. Hummer and A. Szabo. “Optimal Dimensionality Reduction of Multistate Kinetic and Markov-State Models”. *J. Phys. Chem. B* 119.29 (2015), pp. 9029–9037.

[43] S. Orioli and P. Faccioli. “Dimensional Reduction of Markov State Models from Renormalization Group Theory”. *J. Chem. Phys.* 145.12 (2016), p. 124120.

[44] 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.* 139.18 (2013), p. 184114.

[45] U. Sengupta, M. Carballo-Pacheco, and B. Strodel. “Automated Markov State Models for Molecular Dynamics Simulations of Aggregation and Self-Assembly”. *J. Chem. Phys.* 150.11 (2019), p. 115101.

[46] M. Carballo-Pacheco and B. Strodel. “Advances in the Simulation of Protein Aggregation at the Atomistic Scale”. *J. Phys. Chem. B* 120.12 (2016), pp. 2991–2999.

[47] Q. Qiao, G. R. Bowman, and X. Huang. “Dynamics of an Intrinsically Disordered Protein Reveal Metastable Conformations That Potentially Seed Aggregation”. *J. Am. Chem. Soc.* 135.43 (2013), pp. 16092–16101.

[48] D.-A. Silva, G. R. Bowman, A. Sosa-Peinado, and X. Huang. “A Role for Both Conformational Selection and Induced Fit in Ligand Binding by the LAO Protein”. *PLoS Comput Biol* 7.5 (2011). Ed. by R. Nussinov, e1002054.

[49] U. Sengupta and B. Strodel. “Markov Models for the Elucidation of Allosteric Regulation”. *Phil. Trans. R. Soc. B* 373.1749 (2018), p. 20170178.

[50] N. Plattner and F. Noé. “Protein Conformational Plasticity and Complex Ligand-Binding Kinetics Explored by Atomistic Simulations and Markov Models”. *Nat. Commun.* 6 (2015), p. 7653.

[51] C. R. Baiz, Y.-S. Lin, C. S. Peng, K. A. Beauchamp, V. A. Voelz, V. S. Pande, and A. Tokmakoff. “A Molecular Interpretation of 2D IR Protein Folding Experiments with Markov State Models”. *Biophys. J.* 106.6 (2014), pp. 1359–1370.

[52] S. Olsson, H. Wu, F. Paul, C. Clementi, and F. Noé. “Combining Experimental and Simulation Data of Molecular Processes via Augmented Markov Models”. *Proc. Natl. Acad. Sci.* 114.31 (2017), pp. 8265–8270.

[53] F. Noé and F. Nüske. “A Variational Approach to Modeling Slow Processes in Stochastic Dynamical Systems”. *Multiscale Model. Simul.* 11.2 (2013), pp. 635–655.

[54] R. T. McGibbon and V. S. Pande. “Variational Cross-Validation of Slow Dynamical Modes in Molecular Kinetics”. *The Journal of Chemical Physics* 142.12 (2015), p. 124105.

[55] H. Wu and F. Noé. “Variational Approach for Learning Markov Processes from Time Series Data”. *J Nonlinear Sci* (2019).

[56] A. Mardt, L. Pasquali, H. Wu, and F. Noé. “VAMPnets for Deep Learning of Molecular Kinetics”. *Nat. Commun.* 9.1 (2018), pp. 1–11.

[57] W. Chen, H. Sidky, and A. L. Ferguson. “Nonlinear Discovery of Slow Molecular Modes Using State-Free Reversible VAMPnets”. *J. Chem. Phys.* 150.21 (2019), p. 214114.

[58] L. Bonati, G. Piccini, and M. Parrinello. “Deep Learning the Slow Modes for Rare Events Sampling”. *Proc Natl Acad Sci USA* 118.44 (2021), e2113533118.

[59] A. Mardt, L. Pasquali, F. Noé, and H. Wu. “Deep Learning Markov and Koopman Models with Physical Constraints”. *Proc. First Math. Sci. Mach. Learn. Conf.* Ed. by J. Lu and R. Ward. Vol. 107. Proceedings of Machine Learning Research. Princeton University, Princeton, NJ, USA: PMLR, 2020, pp. 451–475.

[60] H. Wu, A. Mardt, L. Pasquali, and F. Noe. “Deep Generative Markov State Models”. *ArXiv180507601 Phys. Stat* (2018). arXiv: 1805.07601 [physics, stat].

[61] A. Mardt and F. Noé. “Progress in Deep Markov State Modeling: Coarse Graining and Experimental Data Restraints”. *J. Chem. Phys.* 155.21 (2021), p. 214106.

[62] K. A. Konovalov, I. C. Unarta, S. Cao, E. C. Goonetilleke, and X. Huang. “Markov State Models to Study the Functional Dynamics of Proteins in the Wake of Machine Learning”. *JACS Au* 1.9 (2021), pp. 1330–1341.

[63] S. Olsson and F. Noé. “Dynamic Graphical Models of Molecular Kinetics”. *Proc. Natl. Acad. Sci.* 116.30 (2019), pp. 15001–15006.

[64] T. Hempel, M. J. del Razo, C. T. Lee, B. C. Taylor, R. E. Amaro, and F. Noé. “Independent Markov Decomposition: Toward Modeling Kinetics of Biomolecular Complexes”. *Proc Natl Acad Sci USA* 118.31 (2021), e2105230118.

[65] B. O. Koopman. “Hamiltonian Systems and Transformation in Hilbert Space”. *Proc. Natl. Acad. Sci. U.S.A.* 17.5 (1931), pp. 315–318.

[66] C. Wehmeyer, M. K. Scherer, T. Hempel, B. E. Husic, S. Olsson, and F. Noé. “Introduction to Markov State Modeling with the PyEMMA Software [Article v1.0]”. *LiveCoMS* 1.1 (2019), p. 5965.

[67] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller. “SchNet – A Deep Learning Architecture for Molecules and Materials”. *The Journal of Chemical Physics* 148.24 (2018), p. 241722.

[68] K. Schütt, O. Unke, and M. Gastegger. “Equivariant Message Passing for the Prediction of Tensorial Properties and Molecular Spectra”. *Proc. 38th Int. Conf. Mach. Learn.* Ed. by M. Meila and T. Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 9377–9388.

[69] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. “Pytorch: An Imperative Style, High-Performance Deep Learning Library”. *Adv. Neural Inf. Process. Syst.* 2019, pp. 8026–8037.

[70] M. Hoffmann, M. Scherer, T. Hempel, A. Mardt, B. de Silva, B. E. Husic, S. Klus, H. Wu, N. Kutz, S. L. Brunton, and F. Noé. “Deeptime: A Python Library for Machine Learning Dynamical Models from Time Series Data”. *Mach. Learn.: Sci. Technol.* 3.1 (2022), p. 015009.

[71] L. R. Rabiner. “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”. *Proc. IEEE* 77.2 (1989), pp. 257–286.

[72] W. R. Inc. *Mathematica, Version 11.2.0*. 2017. `https://www.wolfram.com/mathematica`.

[73] A. A. Hagberg, D. A. Schult, and P. J. Swart. “Exploring Network Structure, Dynamics, and Function Using NetworkX”. *Proc. 7th Python Sci. Conf.* Ed. by G. Varoquaux, T. Vaught, and J. Millman. Pasadena, CA USA, 2008, pp. 11–15.

[74] T. Hempel, N. Plattner, and F. Noé. “Coupling of Conformational Switches in Calcium Sensor Unraveled with Local Markov Models and Transfer Entropy”. *J. Chem. Theory Comput.* 16.4 (2020), pp. 2584–2593.

[75] T. C. Südhof. “Neurotransmitter Release: The Last Millisecond in the Life of a Synaptic Vesicle”. *Neuron* 80.3 (2013), pp. 675–690.

[76] J. L. Jiménez, G. R. Smith, B. Contreras-Moreira, J. G. Sgouros, F. A. Meunier, P. A. Bates, and G. Schiavo. “Functional Recycling of C2 Domains Throughout Evolution: A Comparative Study of Synaptotagmin, Protein Kinase C and Phospholipase C by Sequence, Structural and Modelling Approaches”. *J. Mol. Biol.* 333.3 (2003), pp. 621–639.

[77] J. Guillén, C. Ferrer-Orta, M. Buxaderas, D. Pérez-Sánchez, M. Guerrero-Valero, G. Luengo-Gil, J. Pous, P. Guerra, J. C. Gómez-Fernández, N. Verdaguer, and S. Corbalán-García. “Structural Insights into the Ca2+ and PI(4,5)P2 Binding Modes of the C2 Domains of Rabphilin 3A and Synaptotagmin 1”. *Proc. Natl. Acad. Sci.* 110.51 (2013), pp. 20503–20508.

[78] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw. “How Fast-Folding Proteins Fold”. *Science* 334.6055 (2011), pp. 517–520.

[79] A. L. Hodgkin and A. F. Huxley. “A Quantitative Description of Membrane Current and Its Application to Conduction and Excitation in Nerve”. *The Journal of Physiology* 117.4 (1952), pp. 500–544.

[80] Y. Rudy and J. R. Silva. “Computational Biology in the Study of Cardiac Ion Channels and Cell Electrophysiology”. *Q. Rev. Biophys.* 39.1 (2006), pp. 57–116.

[81] M. Bykhovskaia. “Calcium Binding Promotes Conformational Flexibility of the Neuronal Ca2+ Sensor Synaptotagmin”. *Biophys. J.* 108.10 (2015), pp. 2507–2520.

[82] H. T. Tran, L. H. Anderson, and J. D. Knight. “Membrane-Binding Cooperativity and Coinsertion by C2AB Tandem Domains of Synaptotagmins 1 and 7”. *Biophysical Journal* 116.6 (2019), pp. 1025–1036.

[83] M. Ghorbani, S. Prasad, J. B. Klauda, and B. R. Brooks. “GraphVAMPNet, Using Graph Neural Networks and Variational Approach to Markov Processes for Dynamical Modeling of Biomolecules”. *J. Chem. Phys.* 156.18 (2022), p. 184103.

[84] O. Ledoit and M. Wolf. “A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices”. *Journal of Multivariate Analysis* 88.2 (2004), pp. 365–411.

[85] Y. Chen, A. Wiesel, Y. C. Eldar, and A. O. Hero. “Shrinkage Algorithms for MMSE Covariance Estimation”. *IEEE Trans. Signal Process.* 58.10 (2010), pp. 5016–5029.

[86] C. R. Harris et al. “Array Programming with NumPy”. *Nature* 585.7825 (2020), pp. 357–362.

[87] J. D. Hunter. “Matplotlib: A 2D Graphics Environment”. *Comput. Sci. Eng.* 9.3 (2007), pp. 90–95.

[88] M. K. Scherer, B. E. Husic, M. Hoffmann, F. Paul, H. Wu, and F. Noé. “Variational Selection of Features for Molecular Kinetics”. *J. Chem. Phys.* 150.19 (2019), p. 194108.

[89] A. Mardt, T. Hempel, C. Clementi, and F. Noé. *Deep Learning to Decompose Macromolecules into Independent Markovian Domains*. Zenodo. 2022. `https://zenodo.org/record/7215890`.

[90] T. Hempel, N. Plattner, and F. Noe. *Molecular Dynamics Dataset of Synaptotagmin-1*. 2022. `https://zenodo.org/record/6908073`.

7

Markov field models: scaling molecular kinetics approaches to large molecular machines

This Chapter has been published as

Tim Hempel, Simon Olsson, and Frank Noé. "Markov Field Models: Scaling Molecular Kinetics Approaches to Large Molecular Machines". *Current Opinion in Structural Biology* 77 (2022), p. 102458. https://doi.org/10.1016/j.sbi.2022.102458

Contributions TH was lead author in this project. He reviewed the underlying literature (with SO), derived the presented generalization, and conducted comparative analyses. He has created the figures, and was main author of the manuscript. All authors contributed to writing the manuscript. (This paragraph summarizes TH's contributions alone, it is not an exhaustive list of other authors' contributions.)

Abstract

With recent advances in structural biology, including experimental techniques and deep learning-enabled high-precision structure predictions, molecular dynamics methods that scale up to large biomolecular systems are required. Current state-of-the-art approaches in molecular dynamics modeling focus on encoding global configurations of molecular systems as distinct states. This paradigm commands us to map out all possible structures and sample transitions between them, a task that becomes impossible for large-scale systems such as biomolecular complexes. To arrive at scalable molecular models, we suggest moving away from global state descriptions to a set of coupled models that each describe the dynamics of local domains or sites of the molecular system. We describe limitations in the current state-of-the-art global-state Markovian modeling approaches and then introduce Markov field models as an umbrella term that includes models from various scientific communities, including Independent Markov decomposition, Ising and Potts models, and (dynamic) graphical models, and evaluate their use for computational molecular biology. Finally, we give a few examples of early adoptions of these ideas for modeling molecular kinetics and thermodynamics.

7.1 Introduction

Computer simulations such as molecular dynamics (MD) are established tools for understanding the function of molecular machines on an atomistic scale. In contrast to experiments, *in silico* methods are not limited by their spatial or temporal resolution; their Achilles' heel is that enough data must be gathered to describe a biological system in thermodynamic equilibrium. Many recent advances have contributed to the solution of this so-called sampling problem, such as hardware developments like fast graphical processing units (GPUs), efficient software packages [1, 2], and enhanced sampling methods that, for example, use bias potentials along a reaction coordinate [3–8] or diffusion maps [9]. Additionally, Markov state models (MSMs) have leveraged fast parallel processing power by combining large numbers of short off-equilibrium trajectories without defining reaction coordinates or introducing bias potentials to the system [10–15]. MSMs have been profiting substantially from the development of deep learning methods in recent years [16–18]; see the study by Noé [19] for an overview of both shallow and deep machine learning (ML) methods in this area.

These combined efforts have been very successful, shedding light on complex molecular processes such as protein folding [20–23], ligand-protein binding [24–27], or even

protein-protein association kinetics [28]. These small to medium-sized protein systems are often cooperative, giving rise to a small number of rare-event processes between a few long-lived, metastable states, and thus MSMs or other kinetic models can be used to characterize their rare-event dynamics globally [29].

However, this approach does not scale with increasing size of the molecular system, as its cooperativity decreases and thus the number of globally distinct metastable states increases combinatorially. As an example, consider a solution of N dissociated proteins [30]. If each protein can be in one of two states, the number of all global states is 2^N, that is, it scales exponentially with the number of constituents. Therefore, sampling each global system configuration and the transitions between them becomes infeasible even for a small number of proteins. A biomolecular complex, of course, has more cooperativity, and the coupling between domains may reduce the number of globally accessible states. However, the fundamental problem remains, as an increasing number of loosely coupled domains will lead to an exponentially increasing number of global system states. Therefore, even though all-atom MD simulations can now be conducted with impressive system sizes such as a virus in an aerosol particle [31] or a membrane model of the endoplasmic reticulum [32], they will not lead to the ability to directly parameterize a global state model (e.g., an MSM) in the near future. This task would require us to increase the aggregate simulation time exponentially with system size, whereas it typically decreases with system size in reality.

The main idea proposed in this manuscript is to avoid the exponential scaling by adopting ideas from Ising models (Fig. 7.1a,b), a multiscale approach that explicitly recognizes that there are loosely coupled subsets of the complex structure ("domains", loosely associated with "spins" in an Ising model). Instead of modeling a biomolecular complex as a global entity, we propose to describe its dynamics by a graph consisting of domains (e.g., protein domains) that interact via edges (i.e., coupling) [33], thus forming the full sequence protein complex (or assembly of domains) as shown in Fig. 7.1c,d. In this setting, each domain has a limited number of states that can be sampled, and the coupling depends only on local states (e.g. of the pair of coupled states) without the need to explicitly encode the global state. Like in an Ising model, the global dynamics arises from a combination of simple parts.

Although Ising models are established in statistical physics, marrying them with kinetic models of biomolecular complexes is challenging, as we need to identify the domains that are useful for such a model. Furthermore, it is yet unclear which of the various dynamical models and coupling approaches will be most suitable to describe macromolecular dynamics. In contrast to simple physics models such as Ising models,

there is no *a priori* definition of discrete sites, spins or subdomains in a protein system, nor is the number or discreteness of states within each subsystem well-defined. Their definition will instead largely depend on the observables in which one is interested in computing. Here, we therefore introduce Markov field models (MFMs) as an umbrella term and discuss recent progress and ideas in this direction. A key idea is Independent Markov decomposition (IMD), which is an approach that spatially decomposes a system into independent domains that are subsequently described by domain-specific (uncoupled) Markov state models. In this review, we trace the path from IMD to higher models in which the domains are coupled, such as Dynamical graphical models. We discuss a) how domain decompositions can be estimated from data and b) ideas to represent the global thermodynamics and kinetics in terms of a coupled local dynamics between such domains.

7.2 Markovian dynamics

Markovian models describe the dynamics generated by an operator $\mathcal{P}$, which propagates the probability density ρ of a system over a finite time τ given an initial probability density ρ_0 [15],

$$\rho_\tau = \mathcal{P}(\tau) \circ \rho_0.$$

These dynamics are usually approximated by lumping protein conformations into M discrete global states. In this case, the operator $\mathcal{P}(\tau)$ becomes the transition matrix $\mathbf{P}(\tau)$ with its element (i,j) describing the conditional probability to jump from state i to state j within a lag time τ. Furthermore, the probability densities ρ become vectors $\mathbf{p}$ that encode the probability distribution over the discrete states. The dynamics of the resulting Markov State Model (MSM) is then

$$\mathbf{p}_\tau^\top = \mathbf{p}_0^\top \mathbf{P}(\tau), \tag{7.1}$$

with $(\cdot)^\top$ denoting the vector transpose. This model describes the kinetics of the underlying system and can, e.g., be depicted as a network of exchanging states as in Fig. 7.1a. In this review, we limit our discussion of Markovian models to discrete state MSMs, while a whole zoo of such models exist, including transfer operator and Koopman operator models, Master-equation models and fuzzy MSMs. Please compare the study by Noé [19] for references and the DeepTime library [34] for software implementations.

We usually assume that the eigenvalue spectrum of the MSM transition matrix has a small number of eigenvalues close to 1 at a given lagtime τ, corresponding to only a

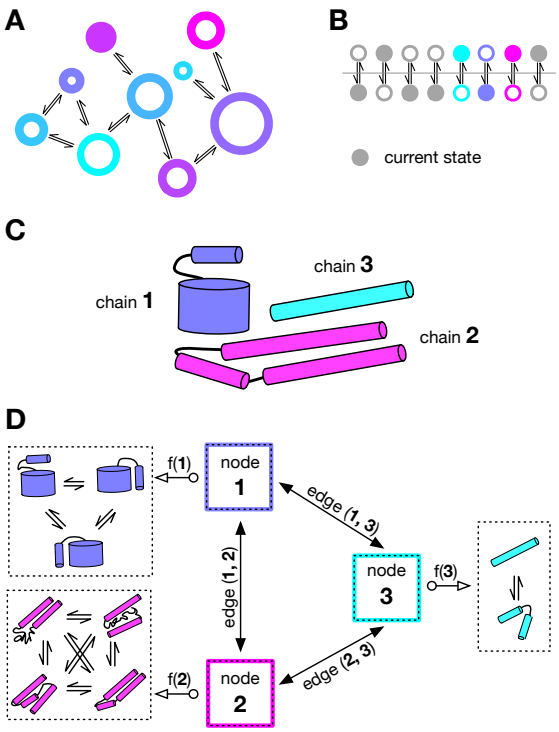


Figure 7.1: Markov field models: From a biomolecular complex to a graph of stochastically coupled subdomains. **(a)** MSM with M distinct, global states that have different equilibrium probabilities (denoted by circle size). Arrows indicate where direct transitions between states are possible. The instantaneous state of the system is defined by being in a single state, here indicated by the filled circle. **(b)** Ising model with N local spins. Each Ising spin can be in one of two states. The instantaneous state of the system is defined by the combination of all spin settings, here indicated by the filled circles. Three spins, corresponding to three MFM domains, are highlighted by color-coding as in subsequent panels. **(c)** Example of a molecular complex, where each of three protein chains is modeled as a dynamical subsystem. **d:** Graph view of the molecular complex, each protein chain is now seen as a node in a graph. The edges (◂▸) denote interaction terms, that is, where the state transitions of each domain can be coupled. Next to each node (○▷), the states and transitions between states of that node or protein subsystem are shown.

few number of transition processes whose autocorrelation times significantly exceed τ, and therefore are slow relative to τ. This is a mild assumption which has been found to be practically useful for many small- to mid-sized proteins which have folded structures and are therefore sufficiently cooperative. For example, a popular atomistic model of capped alanine has three slow processes on the nanoseconds timescale which may be identified by finding the number of implied timescales above the implied timescales gap (Fig. 7.2a, [15]). In our atomistic model of the C2A domain of Synaptotagmin, we find 24 metastable states connected by processes on the 100s of nanoseconds timescale [35], whereas we expect this number to be in the 100s for the C2AB dimer (Fig. 7.2b). Parametrizing a global MSM for this system would require extensive MD simulations, likely on the order of multiple milliseconds in aggregate trajectory data.

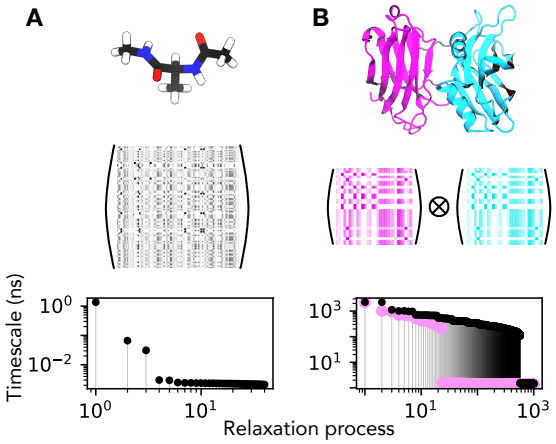


Figure 7.2: Transition matrices and implied timescale spectra of a simple test system and a multi-domain protein. Descriptions from top to bottom. **(a)** Capped alanine, depicted in licorice representation, as represented by 100x100 transition matrix. The spectrum shows three implied timescales that describe slow processes. **(b)** Syt-C2AB consists of a C2A and a C2B sub-domain [38, 39], color-coded magenta and cyan. Assuming equal dynamics in both domains and no coupling, the combined spectrum has >550 slow implied timescales (black dots). We note that this represents an upper bound as the real number may be reduced by domain-domain couplings. For comparison, the spectrum of only the C2A domain has already 23 slow processes above the implied timescales gap (magenta dots), as identified by our previous work [35].

7.3 Markov field models

The abstract idea of MFMs is to model the global transition matrix $\mathbf{P}$ by a tensor product of smaller transition matrices $\mathbf{P}_i$ that govern the dynamics of N local domains, and a coupling term $\mathbf{Y}$ which models the statistical dependence between individual Markov domains [36, 37]:

$$\mathbf{P} \approx \bigotimes_{i=1}^{N} \mathbf{P}_i + \mathbf{Y}. \tag{7.2}$$

In Fig. 7.5, we compare two methods that either explicitly model coupling terms $\mathbf{Y}$ or discard them altogether. Intuitively, MFMs are suitable for larger systems as the number of model parameters stored on the right side of Eq. (7.2) may be smaller compared to the left side, saving memory and increasing statistical efficiency as the number of independent parameters that have to be estimated is smaller.

As indicated in Eq. (7.2), MFMs can in principle generate the global transition matrix $\mathbf{P}$ as a function of local transition matrices $\mathbf{P}_i$ and their coupling, and they therefore allow us to compute thermodynamic and kinetic quantities and compare to experimental data. However, in contrast to MSMs, it is not necessary and often not feasible to actually compute $\mathbf{P}$ explicitly. As an example consider an Ising model with N spins which

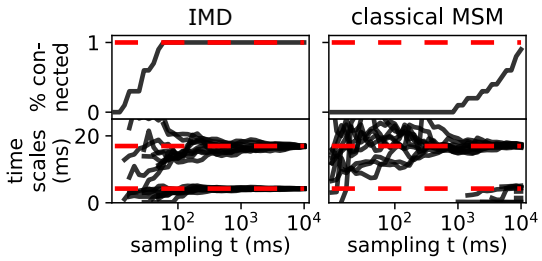


Figure 7.3: Statistical efficiency of IMD versus MSM. How much sampling is necessary for reconstructing the Hodgkin–Huxley model from simple discrete data? IMD (left column) is compared with classical MSMs (right column) for ten realizations of the underlying random process. The percentage of connected (valid) transition matrices in this ensemble is assessed as well as representative implied timescales (ground truth given by red dashed lines). Figure reproduced from the study by Hempel et al [35].

possesses 2^N global states with a $2^N \times 2^N$ transition matrix. Instead of computing and analyzing $\mathbf{P}$ directly, we will thus usually sample the dynamical models by running a simulation algorithm that employs $\mathbf{P}_i$ and $\mathbf{Y}$, for example, some form of Markov-Chain Monte Carlo.

Compared to global-state models, MFMs come with the challenge of finding a meaningful decomposition of the molecular system into domains (nodes) and learning the interaction graph (edges), which is a problem that has been extensively studied [29, 40]. In the following, we will review methods that seek to approximate the right-hand side of Eq. 7.2.

7.4 Independent Markov decomposition

When the local domain Markov processes are uncoupled, that is, statistically independent of each other, Eq. (7.2) simplifies to the Kronecker product of local transition matrices, a model termed Independent Markov decomposition (IMD) [35],

$$\mathbf{P} = \bigotimes_i \mathbf{P}_i. \tag{7.3}$$

An IMD model approximates the system dynamics by neglecting interaction terms between domains, that is, does not model couplings (the edges in Fig. 7.1d are assumed to be negligible to first order). Even though this assumption is not expected to completely model the global kinetics, its strength is that it can reduce the simulation data required to get sufficient sampling by orders of magnitude compared to an MSM (Fig. 7.3) and it extracts approximately independent Markov domains from data, often leading to a much more straightforwardly interpretable model compared to a global MSM (Tab. 7.1).

Higher-order coupling terms $\mathbf{Y}$ can then be added *a posteriori* in order to better approximate the global dynamics.

Based on the variational approach to Markov processes (VAMP) [41], an independence score could be defined that allows partitioning a system into domains, maximizing the metastability in the state definition of each domain. IMD has been shown to be applicable to high-dimensional MD data [42]. An application example is shown in Fig. 7.5. In a mathematical context, IMD can be considered as an estimator for the *nearest Kronecker product problem* [43, 44] which describes the decomposition of a matrix $\mathbf{A}$ into a product $\mathbf{B} \otimes \mathbf{C}$.

7.5 iVAMPNets

The idea of IMD was later combined with VAMPNets [16] to iVAMPNets [30], an end-to-end unsupervised deep learning system which performs IMD given MD simulation data. To this end, iVAMPNets learn a probabilistic partitioning of the protein structure to approximately independent Markov domains by using an attention mechanism, and then use VAMPnets in order to learn the nonlinear coordinate transform into collective coordinates resolving the rare events in these individual domains as well as a partitioning into their local Markov states. The local MSMs $\mathbf{P}_i$ can then be easily extracted. iVAMPNets are trained by minimizing a loss function that combines VAMP [41] and IMD [35]. The deep neural network architecture is shown and compared to the original VAMPNet architecture in Fig. 7.4.

7.6 Ising and Potts models, graphical models, and Markov random fields

Arguably the most famous models that rely on local properties instead of global state space descriptions are the Ising model [45] (Fig. 7.1b) and the Potts model [46]. They model equilibrium distributions of coupled local domains, which are called spins due to their original use-case of modeling magnetic materials. Ising and Potts models were successfully applied to biological systems, for example, in protein folding [47, 48] or direct-coupling analysis (DCA) [49]. The Ising model was extended to a continuous-time dynamical model by Glauber [50, 51]. The topology of the model is defined by means of a coupling graph (Tab. 7.1). For the simplest case of a linear periodic chain of spins with nearest-neighbor couplings and no external field, Glauber dynamics defines

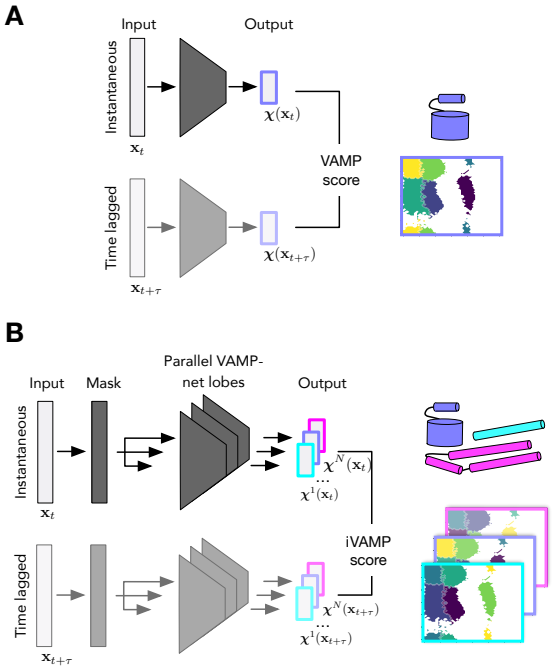


Figure 7.4: Performing an independent Markov decomposition with iVAMPNets. Simplified neural network architecture of VAMPNets **(a)** and iVAMPNets **(b)**. Dark shaded shapes contain trainable weights. Domain assignments are color-coded and associated with sample projections of VAMPNet metastable assignments.

the rate of a spin i to flip its state $\sigma_i \in \{-1, 1\}$ as a function of its neighbors,

$$q_i = q_i(\sigma_{i-1}, \sigma_{i+1}) = \frac{\alpha}{2} - \gamma \frac{\alpha}{2} \sigma_i \frac{\sigma_{i-1} + \sigma_{i+1}}{2} \tag{7.4}$$

with the flipping rate of an independent spin $\alpha/2$ and the coupling parameter γ. Sampling spin-flips from these rates (e.g. using a Markov chain sampler) will sample from the Boltzmann distribution of the Ising model Hamiltonian and define dynamics in the sense of an MFM. To compare with the general idea of MFMs (Eq. (7.2)), we can interpret the $\alpha/2$ term as the transition rate ignoring the coupling with neighbor spins and the second term as the coupling term.

Ising model approaches can be generalized to higher dimensions, other coupling graphs, other spin Hamiltonians with external fields, etc. When coupling graphs are used that do not correspond to a regular lattice topology, one usually speaks of the more general Markov random fields (MRF) or graphical models. Both models have been extensively used in statistics and machine learning in order to express the conditional dependence of random variables. In MRFs, which are a generalization of Ising and Potts models, this dependence is expressed by an undirected graph – for example

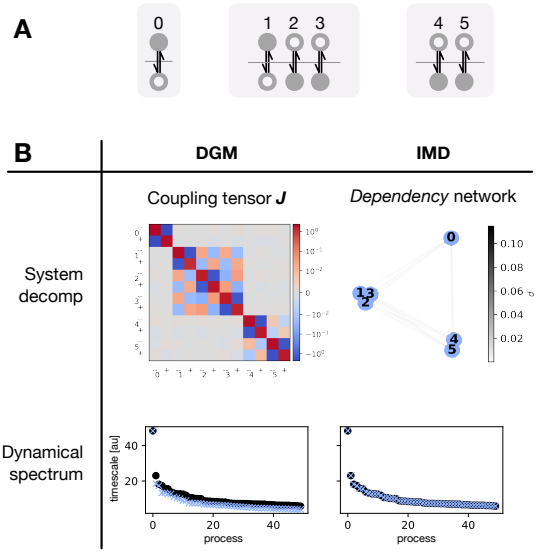


Figure 7.5: Comparison of a method explicitly modeling coupling terms (DGM) and one discarding them (IMD). **(a)** The test system is a set of spins with coupling which is constructed such that spin groups $[0]$, $[1, 2, 3]$, and $[4, 5]$ are mutually independent. **(b)** Modeling results of DGMs (left) compared to IMD (right). DGMs explicitly model the coupling in a tensor, shown as a matrix here, whereas IMD qualitatively groups coupled spins together in a graph plot. The implied timescales spectrum is well approximated by both methods, IMD matching the ground truth to numerical precision as the test systems are truly independent.

the probability distribution of spin σ_{i+1} in the 1D Ising model conditionally depends on σ_i and vice versa. The classical periodic Ising chain in one dimension would be depicted as a circle with every node interacting with its left and right neighbors. MRFs can be efficiently evaluated using energy functions defined over so-called *cliques* (i.e., fully connected subgraphs) [52, 53]. Conceptually, MRFs directly relate to the graph shown in Fig. 7.1d.

Graphical models are another generalization of MRFs and use directed graphs, that is, in a graphical model it is possible that the distribution of random variable σ_{i+1} conditionally depends on σ_i but not vice versa [54].

7.7 Dynamic graphical models

MRFs and graphical models specify the conditional dependence of random variables and define an equilibrium distribution when sampled. However, they both do not define the dynamics, for example, they do not model how often a given spin flip or state transition is attempted in a given physical time window. In order to model biomolecular kinetics, we therefore need to additionally specify a dynamical model.

In their 2017 article [55], Gerber and Horenko derive a method to model a system's dynamics from domain-specific feature vectors, e.g., encoding discrete backbone dihedral states. Although it is not directly based on MRFs or graphical models, it can be considered to be in this tradition. Olsson & Noé [29] derived a method explicitly developed as an extension of MRFs to dynamical systems under the name Dynamic graphical models (DGMs). Given a set of known domains, DGMs address the problem of learning the coupling graph (Tab. 7.1) and, subsequently, the global transition matrix from the data by solving an independent logistic regression problem per domain. They assume conditional independence of the future states of a domain, given all domain states at a previous time. More specifically, DGMs model the transition probabilities for a domain configuration $\mathbf{s}_t = (\sigma_{0,t}, \sigma_{1,t}, \ldots, \sigma_{N,t})$ given an initial configuration $\mathbf{s}_{t-\tau}$. The coupling is explicitly modeled with a coupling tensor $\mathbf{J} = \mathbf{J}(\tau)$. To compare to the general idea (Eq. (7.2)), we re-write the DGM transition probabilities in the binary case and in absence of an external field,

$$p(\mathbf{s}_t|\mathbf{s}_{t-\tau}) \sim \exp\left[\sum_{i=1}^{N} \sigma_{i,t}^{\top}\mathbf{J}_{ii}\sigma_{i,t-\tau} + \sum_{i=1}^{N} \sigma_{i,t}^{\top}\sum_{\substack{j=1\\ j\neq i}}^{N} \mathbf{J}_{ij}\sigma_{j,t-\tau}\right], \tag{7.5}$$

highlighting how self- and pair-coupling terms enter the model. Here, $\sigma_{i,t}$ encode the i'th domain state at time t and $\mathbf{J}_{ij}$ is a sub-matrix of $\mathbf{J}$ encoding the coupling between domains i and j. We note that Eq. (7.5) yields a probability for a given global spin configuration $\mathbf{s}$; the construction of a global transition matrix is equivalent to the Glauber model. An example of a DGM application and the coupling tensor is shown in Fig. 7.5.

7.8 Stochastic automata networks

Stochastic automata networks (SANs) [56, 57] are MFMs that aim to reduce the dynamics of a system with a large state space into a network of weakly coupled random processes (or stochastic automata) using Kronecker products. They operate on a known set of domains and model couplings with fixed functional forms (Tab. 7.1), which are occasional synchronization events and environment-dependent changes of transition probabilities. SANs were originally developed in the computer science community and have been applied to biophysical applications, for example, for modeling the state transition dynamics of coupled Ca^{2+} channels [58].

Within the SAN formulation, the transition matrix of the global system $\mathbf{P}$ can be constructed from the local domain transition matrices $\mathbf{P}_i^{(l)}$ and matrices $\mathbf{P}_i^{(s)}$ and

Model	Domains	Couplings
Glauber dynamics	pre-specified	Pre-specified
DGM	pre-specified	Learned
IMD	Learned	n/a
iVAMPNets	Learned	n/a
SAN	Pre-specified	Pre-specified

Table 7.1: Classification of different Markov field models.

$\mathbf{P}_i^{(s,n)} = \mathrm{diag}(\sum_{\mathrm{rows}} \mathbf{P}_i^{(s)})$ that encode synchronization events* $s \in \epsilon$. The lower index i denotes the local system, ϵ is the set of synchronization events, and l is the label that determines the type of transition. The global transition matrix can be written as [59]

$$\mathbf{P} = \bigotimes_{i=1}^{N} \mathbf{P}_i^{(l)} + \left[\sum_{s \in \epsilon} \bigotimes_{i=1}^{N} \mathbf{P}_i^{(s)} - \bigotimes_{i=1}^{N} \mathbf{P}_i^{(s,n)} \right], \tag{7.6}$$

which again can be interpreted in light of the general idea of MFMs (Eq. (7.2)). When stochastic automata – or, in other words, Markov processes – are mutually independent, the SAN equation (Eq. (7.6)) simplifies to the Kronecker product of the local transition matrices [57] which is equivalent to IMD (Eq. (7.3)).

7.9 Related models

A slightly different approach to modeling dynamical systems which may be seen as related to MFMs is causality modeling, which is often based on Granger's notion of causality [60] or Schreiber's definition of transfer entropy [61]. Causality models often estimate directed graphs between local domains but without attempting to model the global dynamics of the system. Causality modeling may, for example, proceed by inferring pairwise causality between measurement channels of recorded time-series data [62].

Gerber and Horenko present a mathematically rigorous approach to estimate such models for MD simulations [63]. Estimating the causality graph from time-series data of short peptide dihedral torsion angles, they shed light on the spatial and temporal structure of residue-residue interactions. Furthermore, causality modeling has been applied to quantify allosteric couplings in proteins [42, 64], that is, to model directional influence between spatially distant residues or loops.

*Simplified notation, please consult B. Plateau and K. Atif. "Stochastic Automata Network of Modeling Parallel Systems". *IIEEE Trans. Software Eng.* 17.10 (1991), pp. 1093–1108. for details.

7.10 Discussion and outlook

MFMs take the view that the dynamics of a complex system, for example, a protein, can be modeled by multiple weakly coupled, or independent dynamic domains. Even though MFMs are proper generalizations of global-state methods such as MSMs, they only unfold their full potential if subsystems exist such that their coupling is weak or negligible, or when the subsystems and their coupling share similarities. In these cases MFMs will require fewer parameters than MSMs and other global dynamic models in which each transition probability is assumed to be an independent parameter, and MFMs are then more statistically data efficient , avoiding the exponential scaling problem of MSMs for large systems. Even in cases where MFMs are not statistically more efficient, learning MFMs from data can be very insighful as they can partition large dynamical systems into smaller, weakly coupled domains and present a simpler and better interpretable picture of the individual domain dynamics and their coupling than a global state model in whose parameters the local dynamics and their couplings are compounded.

The MFM concept has already been successfully applied as a simulation model to generate data on large molecular complexes. Notable examples are Ultra-Coarse-Graining (UCG), which treats local domains as multi-state coarse-grained beads [65, 66] (compare Fig. 7.1d), or MSM/RD, which describes whole proteins as multi-state particles in a reaction-diffusion (RD) setting [67, 68]. Therefore, we believe that MFMs have a broad application basis for recent and future problems of kinetic modeling.

Research on MFMs is still in its infancy and faces significant challenges. First, learning the dependency graph between subdomains is a highly nontrivial task, and a well-known hard problem in computer science [69, 70]. Algorithms must be able to cope with real world application cases that range somewhere in between the extreme cases of strongly coupled systems and systems with well-defined, completely independent domains. Finding the right trade-off between these extremes is related to defining the domain decomposition optimally. Compared to the hard problem of identifying the true graphical model structure underlying a dataset, MFMs for simulation data are more forgiving in that we are usually mainly interested in learning an MFM that can correctly model statistical observables of the overall systems. This task can be completed, within statistical uncertainty, even when it is not possible to uniquely determine a single MFM structure from the simulation data.

Second, systematic model errors are introduced by neglecting or sparsifying couplings, possibly both at the local domain level and globally. Therefore, it is important

to quantitatively validate MFMs, to model couplings explicitly with methods such as DGMs [29] and / or to merge strongly coupled subsystems in such analyses. Quantifying and controlling such model errors is a task for future work.

Third, validating computational models with experiments usually requires us to estimate the ensemble averages of the global system. In MFMs, it is often unfeasible to explicitly construct and analyze the global transition matrix **P**, which means we have to compute its properties by sampling the coupled subsystem dynamics, which results in statistical errors and possibly sampling problems in order to compute such ensemble averages.

In summary, we believe that MFMs have great potential for modeling and understanding the dynamics of large-scale biomolecular complexes and may significantly reduce the sampling problem, thus advancing the applicability of computationally expensive, high-resolution simulation methods such as atomistic MD. However, many challenges need to be solved until these models have reached a similar maturity and practicality as MSMs, providing fertile soil for future investigations.

Acknowledgments

TH thanks Patrick Gelß for fruitful discussions. We acknowledge the financial support of Deutsche Forschungsgemeinschaft DFG (SFB/TRR 186, Project A12), the European Commission (ERC CoG 772230 “ScaleCell”), the Berlin mathematics research center MATH+ (AA1-6), and Berlin Institute for the Foundations of Learning and Data (BIFOLD). This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation (to S.O.).

Bibliography

[1] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, and V. S. Pande. "OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics". *PLOS Comput. Biol.* 13.7 (2017), e1005659.

[2] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl. "GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers". *SoftwareX* 1–2 (2015), pp. 19–25.

[3] A. Laio and M. Parrinello. "Escaping Free-Energy Minima". *Proceedings of the National Academy of Sciences* 99.20 (2002), pp. 12562–12566.

[4] O. Valsson, P. Tiwary, and M. Parrinello. "Enhancing Important Fluctuations: Rare Events and Metadynamics from a Conceptual Viewpoint". *Annu. Rev. Phys. Chem.* 67.1 (2016), pp. 159–184.

[5] P. Tiwary and B. J. Berne. "Spectral Gap Optimization of Order Parameters for Sampling Complex Molecular Systems". *Proc Natl Acad Sci USA* 113.11 (2016), pp. 2839–2844.

[6] P. Tiwary and M. Parrinello. "From Metadynamics to Dynamics". *Phys. Rev. Lett.* 111.23 (2013), p. 230602.

[7] J. M. L. Ribeiro, P. Bravo, Y. Wang, and P. Tiwary. "Reweighted Autoencoded Variational Bayes for Enhanced Sampling (RAVE)". *The Journal of Chemical Physics* 149.7 (2018), p. 072301.

[8] L. Bonati, G. Piccini, and M. Parrinello. "Deep Learning the Slow Modes for Rare Events Sampling". *Proc Natl Acad Sci USA* 118.44 (2021), e2113533118.

[9] J. Preto and C. Clementi. "Fast Recovery of Free Energy Landscapes via Diffusion-Map-Directed Molecular Dynamics". *Phys. Chem. Chem. Phys.* 16.36 (2014), pp. 19181–19191.

[10] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard. "A Direct Approach to Conformational Dynamics Based on Hybrid Monte Carlo". *J. Comput. Phys.* 151.1 (1999), pp. 146–168.

[11] W. C. Swope, J. W. Pitera, and F. Suits. "Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 1. Theory". *J. Phys. Chem. B* 108.21 (2004), pp. 6571–6581.

[12] N. Singhal, C. D. Snow, and V. S. Pande. "Using Path Sampling to Build Better Markovian State Models: Predicting the Folding Rate and Mechanism of a Tryptophan Zipper Beta Hairpin". *J. Chem. Phys.* 121.1 (2004), pp. 415–425.

[13] 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". *The Journal of Chemical Physics* 126.15 (2007), p. 155101.

[14] F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl. "Constructing the Equilibrium Ensemble of Folding Pathways from Short Off-Equilibrium Simulations". *Proc. Natl. Acad. Sci.* 106.45 (2009), pp. 19011–19016.

[15] 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.* 134.17 (2011), p. 174105.

[16] A. Mardt, L. Pasquali, H. Wu, and F. Noé. “VAMPnets for Deep Learning of Molecular Kinetics”. *Nat. Commun.* 9.1 (2018), pp. 1–11.

[17] W. Chen, H. Sidky, and A. L. Ferguson. “Nonlinear Discovery of Slow Molecular Modes Using State-Free Reversible VAMPnets”. *J. Chem. Phys.* 150.21 (2019), p. 214114.

[18] H. Sidky, W. Chen, and A. L. Ferguson. “High-Resolution Markov State Models for the Dynamics of Trp-Cage Miniprotein Constructed Over Slow Folding Modes Identified by State-Free Reversible VAMPnets”. *J. Phys. Chem. B* 123.38 (2019), pp. 7999–8009.

[19] F. Noé. “Machine Learning for Molecular Dynamics on Long Timescales”. *Machine Learning Meets Quantum Physics*. Ed. by K. T. Schütt, S. Chmiela, O. A. von Lilienfeld, A. Tkatchenko, K. Tsuda, and K.-R. Müller. Vol. 968. Cham: Springer International Publishing, 2020, pp. 331–372.

[20] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw. “How Fast-Folding Proteins Fold”. *Science* 334.6055 (2011), pp. 517–520.

[21] F. Noé, I. Horenko, C. Schütte, and J. C. Smith. “Hierarchical Analysis of Conformational Dynamics in Biomolecules: Transition Networks of Metastable States”. *J. Chem. Phys.* 126.15 (2007), p. 155102.

[22] V. A. Voelz, G. R. Bowman, K. Beauchamp, and V. S. Pande. “Molecular Simulation of *Ab Initio* Protein Folding for a Millisecond Folder NTL9(1-39)”. *J. Am. Chem. Soc.* 132.5 (2010), pp. 1526–1528.

[23] V. A. Voelz, M. Jäger, S. Yao, Y. Chen, L. Zhu, S. A. Waldauer, G. R. Bowman, M. Friedrichs, O. Bakajin, L. J. Lapidus, S. Weiss, and V. S. Pande. “Slow Unfolded-State Structuring in Acyl-CoA Binding Protein Folding Revealed by Simulation and Experiment”. *J. Am. Chem. Soc.* 134.30 (2012), pp. 12565–12577.

[24] D.-A. Silva, G. R. Bowman, A. Sosa-Peinado, and X. Huang. “A Role for Both Conformational Selection and Induced Fit in Ligand Binding by the LAO Protein”. *PLoS Comput Biol* 7.5 (2011). Ed. by R. Nussinov, e1002054.

[25] U. Sengupta and B. Strodel. “Markov Models for the Elucidation of Allosteric Regulation”. *Phil. Trans. R. Soc. B* 373.1749 (2018), p. 20170178.

[26] N. Plattner and F. Noé. “Protein Conformational Plasticity and Complex Ligand-Binding Kinetics Explored by Atomistic Simulations and Markov Models”. *Nat. Commun.* 6 (2015), p. 7653.

[27] I. Buch, T. Giorgino, and G. De Fabritiis. “Complete Reconstruction of an Enzyme-Inhibitor Binding Process by Molecular Dynamics Simulations”. *Proceedings of the National Academy of Sciences* 108.25 (2011), pp. 10184–10189.

[28] N. Plattner, S. Doerr, G. D. Fabritiis, and F. Noé. “Complete Protein–Protein Association Kinetics in Atomic Detail Revealed by Molecular Dynamics Simulations and Markov Modelling”. *Nat. Chem.* 9.10 (2017), p. 1005.

[29] S. Olsson and F. Noé. “Dynamic Graphical Models of Molecular Kinetics”. *Proc. Natl. Acad. Sci.* 116.30 (2019), pp. 15001–15006.

[30] A. Mardt, T. Hempel, C. Clementi, and F. Noe. *Deep Learning to Decompose Macromolecules into Independent Markovian Domains*. Preprint. Biophysics, 2022. url: `http://biorxiv.org/lookup/doi/10.1101/2022.03.30.486366`.

[31] A. Dommer et al. *#COVIDisAirborne: AI-Enabled Multiscale Computational Microscopy of Delta SARS-CoV-2 in a Respiratory Aerosol*. Preprint. Biophysics, 2021. url: http://biorxiv.org/lookup/doi/10.1101/2021.11.12.468428.

[32] N. Trebesch and E. Tajkhorshid. "Embracing Biological Complexity in Atomistic Simulations of Cellular Membranes". *Biophysical Journal* 118.3 (2020), 88a.

[33] F. Harary and G. Gupta. "Dynamic Graph Models". *Mathematical and Computer Modelling* 25.7 (1997), pp. 79–87.

[34] M. Hoffmann, M. Scherer, T. Hempel, A. Mardt, B. de Silva, B. E. Husic, S. Klus, H. Wu, N. Kutz, S. L. Brunton, and F. Noé. "Deeptime: A Python Library for Machine Learning Dynamical Models from Time Series Data". *Mach. Learn.: Sci. Technol.* 3.1 (2022), p. 015009.

[35] T. Hempel, M. J. del Razo, C. T. Lee, B. C. Taylor, R. E. Amaro, and F. Noé. "Independent Markov Decomposition: Toward Modeling Kinetics of Biomolecular Complexes". *Proc Natl Acad Sci USA* 118.31 (2021), e2105230118.

[36] P. Gelß. "The Tensor-Train Format and Its Applications. Modeling and Analysis of Chemical Reaction Networks, Catalytic Processes, Fluid Flows, and Brownian Dynamics". PhD thesis. Freie Universität Berlin, 2017, xvii, 160 Seiten.

[37] P. Gelß, S. Klus, S. Matera, and C. Schütte. "Nearest-Neighbor Interaction Systems in the Tensor-Train Format". *Journal of Computational Physics* 341 (2017), pp. 140–162.

[38] T. C. Südhof. "Calcium Control of Neurotransmitter Release". *Cold Spring Harb Perspect Biol* 4.1 (2012), a011353.

[39] K. L. Fuson, M. Montes, J. J. Robert, and R. B. Sutton. "Structure of Human Synaptotagmin 1 C2AB in the Absence of Ca2+ Reveals a Novel Domain Association," *Biochemistry* 46.45 (2007), pp. 13041–13048.

[40] S. Parise and M. Welling. "Structure Learning in Markov Random Fields". *Adv. Neural Inf. Process. Syst.* 29 (2006), p. 54.

[41] H. Wu and F. Noé. "Variational Approach for Learning Markov Processes from Time Series Data". *J Nonlinear Sci* (2019).

[42] T. Hempel, N. Plattner, and F. Noé. "Coupling of Conformational Switches in Calcium Sensor Unraveled with Local Markov Models and Transfer Entropy". *J. Chem. Theory Comput.* 16.4 (2020), pp. 2584–2593.

[43] C. F. Van Loan. "The Ubiquitous Kronecker Product". *Journal of Computational and Applied Mathematics* 123.1-2 (2000), pp. 85–100.

[44] C. F. Loan and N. Pitsianis. "Approximation with Kronecker Products". *Linear Algebra for Large Scale and Real-Time Applications*. Ed. by M. S. Moonen, G. H. Golub, and B. L. R. Moor. Dordrecht: Springer Netherlands, 1993, pp. 293–314.

[45] E. Ising. "Beitrag zur Theorie des Ferromagnetismus". *Z. Physik* 31.1 (1925), pp. 253–258.

[46] F. Y. Wu. "The Potts Model". *Rev. Mod. Phys.* 54.1 (1982), pp. 235–268.

[47] V. Muñoz. "What Can We Learn about Protein Folding from Ising-like Models?" *Current Opinion in Structural Biology* 11.2 (2001), pp. 212–216.

[48] E. R. Henry, R. B. Best, and W. A. Eaton. “Comparing a Simple Theoretical Model for Protein Folding with All-Atom Molecular Dynamics Simulations”. *Proc. Natl. Acad. Sci. U.S.A.* 110.44 (2013), pp. 17880–17885.

[49] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa. “Identification of Direct Residue Contacts in Protein–Protein Interaction by Message Passing”. *Proc. Natl. Acad. Sci. U.S.A.* 106.1 (2009), pp. 67–72.

[50] R. J. Glauber. “Time-Dependent Statistics of the Ising Model”. *Journal of Mathematical Physics* 4.2 (1963), pp. 294–307.

[51] N. Ito. “Glauber Dynamics of the Ising Model”. *Nonequilibrium Statistical Mechanics in One Dimension*. Ed. by V. Privman. Cambridge: Cambridge University Press, 1997, pp. 93–110.

[52] C. J. Preston. “Gibbs States and Markov Random Fields”. *Gibbs States on Countable Sets*. Cambridge Tracts in Mathematics. Cambridge: Cambridge University Press, 1974, pp. 1–9.

[53] R. Kindermann and J. L. Snell. *Markov Random Fields and Their Applications*. Contemporary Mathematics ; v. 1. Providence, R.I: American Mathematical Society, 1980.

[54] K. P. Murphy. “Dynamic Bayesian Networks: Representation, Inference and Learning”. PhD thesis. University of California, Berkeley, 2002.

[55] S. Gerber and I. Horenko. “Toward a Direct and Scalable Identification of Reduced Models for Categorical Processes”. *Proc Natl Acad Sci USA* 114.19 (2017), pp. 4863–4868.

[56] B. Plateau. “On the Stochastic Structure of Parallelism and Synchronization Models for Distributed Algorithms”. *Proc. 1985 ACM SIGMETRICS Conf. Meas. Model. Comput. Syst. - SIGMETRICS 85*. Austin, Texas, United States: ACM Press, 1985, pp. 147–154.

[57] B. Plateau and W. J. Stewart. “Stochastic Automata Networks”. *Computational Probability*. Ed. by F. S. Hillier and W. K. Grassmann. Vol. 24. Boston, MA: Springer US, 2000, pp. 113–151.

[58] V. Nguyen, R. Mathias, and G. Smith. “A Stochastic Automata Network Descriptor for Markov Chain Models of Instantaneously Coupled Intracellular Ca Channels”. *Bulletin of Mathematical Biology* 67.3 (2005), pp. 393–432.

[59] B. Plateau and K. Atif. “Stochastic Automata Network of Modeling Parallel Systems”. *IIEEE Trans. Software Eng*. 17.10 (1991), pp. 1093–1108.

[60] C. W. J. Granger. “Investigating Causal Relations by Econometric Models and Cross-spectral Methods”. *Econometrica* 37.3 (1969), pp. 424–438.

[61] T. Schreiber. “Measuring Information Transfer”. *Phys. Rev. Lett.* 85.2 (2000), pp. 461–464.

[62] C. J. Quinn, T. P. Coleman, N. Kiyavash, and N. G. Hatsopoulos. “Estimating the Directed Information to Infer Causal Relationships in Ensemble Neural Spike Train Recordings”. *J Comput Neurosci* 30.1 (2011), pp. 17–44.

[63] S. Gerber and I. Horenko. “On Inference of Causality for Discrete State Models in a Multiscale Context”. *Proceedings of the National Academy of Sciences* 111.41 (2014), pp. 14651–14656.

[64] A. Hacisuleyman and B. Erman. “Entropy Transfer between Residue Pairs and Allostery in Proteins: Quantifying Allosteric Communication in Ubiquitin”. *PLOS Comput. Biol.* 13.1 (2017), e1005319.

[65] J. F. Dama, A. V. Sinitskiy, M. McCullagh, J. Weare, B. Roux, A. R. Dinner, and G. A. Voth. “The Theory of Ultra-Coarse-Graining. 1. General Principles”. *J. Chem. Theory Comput.* 9.5 (2013), pp. 2466–2480.

[66] J. M. A. Grime, J. F. Dama, B. K. Ganser-Pornillos, C. L. Woodward, G. J. Jensen, M. Yeager, and G. A. Voth. “Coarse-Grained Simulation Reveals Key Features of HIV-1 Capsid Self-Assembly”. *Nat Commun* 7.1 (2016), p. 11568.

[67] M. Dibak, M. J. del Razo, D. De Sancho, C. Schütte, and F. Noé. “MSM/RD: Coupling Markov State Models of Molecular Kinetics with Reaction-Diffusion Simulations”. *The Journal of Chemical Physics* 148.21 (2018), p. 214107.

[68] M. J. del Razo, M. Dibak, C. Schütte, and F. Noé. “Multiscale Molecular Kinetics by Coupling Markov State Models and Reaction-Diffusion Dynamics”. *J. Chem. Phys.* 155.12 (2021), p. 124109.

[69] M. J. Wainwright, J. Lafferty, and P. Ravikumar. “High-Dimensional Graphical Model Selection Using L1-Regularized Logistic Regression”. *Adv. Neural Inf. Process. Syst.* Vol. 19. 2006.

[70] X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing. “Dags with No Tears: Continuous Optimization for Structure Learning”. *Adv. Neural Inf. Process. Syst.* Vol. 31. 2018.

A

Supplemental information: Independent Markov decomposition: Toward modeling kinetics of biomolecular complexes

This appendix has been published as supplementary material to

Tim Hempel, Mauricio J. del Razo, Christopher T. Lee, Bryn C. Taylor, Rommie E. Amaro, and Frank Noé. "Independent Markov Decomposition: Toward Modeling Kinetics of Biomolecular Complexes". *Proceedings of the National Academy of Sciences* 118.31 (2021), e2105230118.

A.1 Markov operators

A.1.1 Infinitesimal generator

Given a stochastic dynamical system, such as an MD simulation, the operator P can be understood as a propagator of the probability density $f(x, t)$, where x is on the phase space of the dynamical system. We can define P using the infinitesimal generator L [1]

$$\partial_t f = Lf. \tag{A.1}$$

This equation is a generalization to non-deterministic systems of the Liouville equation of statistical mechanics, which describes the time evolution of a density of an ensemble

of systems. Its solution, given an initial density $f(t=0)=f_0$, is

$$f(\cdot,t)=\exp(tL)f_0=P(t)f_0, \tag{A.2}$$

where $P(t)$ is the Perron-Frobenius operator. It can be understood as the propagator of the probability density f_0. In its decomposed form for two systems A and B, the operator is written as

$$f(x,t)=(P_A(t)\otimes P_B(t))f_0(x)=P(t)f_0(x). \tag{A.3}$$

A.1.2 MSM transition matrix decomposition

The Perron-Frobenius operator P can be approximated by a Markov model. The Markov model formulation propagates probability densities between discrete states; therefore, the problem requires performing a Galerkin discretization of P using a discrete basis set. This can be done by partitioning the phase space completely or into the metastable regions, say $\{A_1,\ldots,A_k\}$. The most common basis set are indicator functions on these regions,

$$1_{A_i(x)}=\begin{cases}1 & \text{if } x\in A_i\\ 0 & \text{else.}\end{cases} \tag{A.4}$$

The Galerkin discretization will yield a low rank approximations of the operator P. When using indicator functions, the output will usually be in the form of a discrete-time MSM [2]. Assume that the dynamics of interest can be separated into two independent disjoint regions in phase space. The Galerkin discretization of Eq. (A.2) in each region yields two MSMs,

$$f_A(t+\tau)=T_Af_A(t),\qquad f_B(t+\tau)=T_Bf_B(t), \tag{A.5}$$

where τ is the lagtime; f_A and f_B are the probability vectors of the corresponding MSMs; and T_A, and T_B are the corresponding transition probability matrices. In this case, the matrices T_A and T_B approximate Perron-Frobenius operators, so following Eq. (A.3), the solution of the whole systems is given by

$$f(t+\tau)=(T_A\otimes T_B)f_0(t), \tag{A.6}$$

with $f_0=f_{0A}\otimes f_{0B}$. The individual transition probability matrices are linear maps given by $T_A:\mathbb{R}^k\to\mathbb{R}^k$ and $T_B:\mathbb{R}^{k'}\to\mathbb{R}^{k'}$, where k and k' are the number of discretized states in A and B, respectively. The joint space $A\otimes B$ will then have kk' discretized states that

are defined analogously to Eq. (A.4) by indicator functions $1_{(A\otimes B)_{(i,j)}}(x) = 1$ iff $x \in A_i \cap B_j$, so

$$T_A \otimes T_B : \mathbb{R}^{kk'} \to \mathbb{R}^{kk'}. \tag{A.7}$$

In particular, the product $T_A \otimes T_B$ is the Kronecker product [3]. An analogous expression can be derived for continuous-time MSMs (see below).

In summary, we need the operator P to model the full system, and approximate it by the Kronecker product between the transition probability matrices of the MSMs of independent sub-systems.

A.1.3 Observable operator decomposition

To score dependency of subsystems in feature space, it is most natural to directly work with the operator that propagates these features. This operator is called the Koopman operator K; it propagates observable functions f,

$$Kf(x) = \mathbb{E}[f(\Phi(x))], \tag{A.8}$$

i.e., is described by the expectation value of the observable of a particular configuration, x, after the dynamics Φ has been applied. It is a infinite-dimensional linear operator [4]. It is particularly interesting for the current application because the variational approach for Markov processes (VAMP) and the related VAMP scores are derived from the Koopman operator [5].

As the Koopman formulation is a more general framework to deal with Markov processes, we only refer to the estimator of the Koopman operator which reads [5]

$$K = {C_{00}}^{-1/2} C_{0t} {C_{tt}}^{-1/2} \tag{A.9}$$

with time-lagged covariance matrix C_{0t}, "instantaneous" covariance matrices at times 0 and t C_{00} and C_{tt}, respectively. The lag time is t.

We define the common space of observables of two processes as a stacked vector $\Psi_{AB} = [\Psi_A, \Psi_A]$. For example, if $\Psi_A = (\psi_A^1, \psi_A^2, \ldots)$ and $\Psi_B = (\psi_B^1, \psi_B^2, \ldots)$ are the one-dimensional time series of features $\psi \in \mathbb{R}$ of two systems A and B, the joint space would be spanned by $\Psi_{AB} = ((\psi_A^1, \psi_B^1), (\psi_A^2, \psi_B^2), \ldots)$. Note that this means that the separation of processes happens *a priori* by the choice of Ψ_A, Ψ_B, a situation which comes closest to applied modeling situations. Given the above definition of the full system observable Ψ_{AB}, the full system Koopman operator is the direct sum of Koopman sub-operators K_A, K_B, $K_{AB} = K_A \oplus K_B$ or, more generally,

$$K \begin{pmatrix} \Psi_1 \\ \Psi_2 \\ \vdots \\ \Psi_n \end{pmatrix} (x) = \begin{pmatrix} K_1 & 0 & \dots & 0 \\ 0 & K_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & K_n \end{pmatrix} \begin{pmatrix} \Psi_1 \\ \Psi_2 \\ \vdots \\ \Psi_n \end{pmatrix} (x) \tag{A.10}$$

as all off-diagonal blocks must vanish by definition and each subsystem operator only acts on the features of its space. Therefore, the decomposition can be written as the direct sum

$$K = \bigoplus_i K_i. \tag{A.11}$$

In particular, this means that the Koopman operator has the shape of a block diagonal matrix. It can be seen from the Koopman estimator (Eq. (A.9)) that the above structure of the joint operator implies that independent processes are also uncorrelated.

A.2 VAMP score decomposition of independent systems

The VAMP-p score R_p can be interpreted as the Schatten-p norm $\| \cdot \|_p$ of the estimated Koopman operator to the p-th power [5], i.e.

$$R_p(K) = \|K\|_p^p. \tag{A.12}$$

This general form is valid for both MSMs as well as Koopman models, but note that the estimator for K is different in these cases (see below). To simplify this expression, on the one hand, we can exploit the property of the Schatten-p norm to be invariant under unitary transformations for unitarian matrices U and V,

$$\|A\|_p = \|UAV\|_p. \tag{A.13}$$

On the other hand, we can write the Koopman operator in a singular value decomposition with its singular value diagonal matrix Λ as $K = U\Lambda V$ such that, using Eq. (A.13), we find

$$\|K\|_p = \|\Lambda\|_p = \left(\sum_i \lambda_i^p \right)^{\frac{1}{p}} \tag{A.14}$$

with the real valued singular values of the Koopman matrix λ_i.

A.2.1 Sum space decomposition

Given a joint space that is spanned by the direct sum of subspaces, such as described with molecular observable vectors, and a decomposable Koopman operator $K_{AB} = K_A \oplus K_B$ of two systems A and B, we can thus write

$$\begin{aligned} K_{AB} &= U_{AB}\Lambda_{AB}V_{AB} && \text{(A.15)} \\ &= (U_A \oplus U_B)(\Lambda_A \oplus \Lambda_B)(V_A \oplus V_B) && \text{(A.16)} \\ &= (U_A\Lambda_A V_A) \oplus (U_B\Lambda_B V_B) && \text{(A.17)} \\ &= K_A \oplus K_B && \text{(A.18)} \end{aligned}$$

and hence

$$\begin{aligned} \|K_{AB}\|_p^p &= \|\Lambda_{AB}\|_p^p && \text{(A.19)} \\ &= \|\Lambda_A \oplus \Lambda_B\|_p^p. && \text{(A.20)} \end{aligned}$$

Further, the singular values of the direct sum joint operator are the set of subsystem operator singular values. In detail, writing the p-th power of the p-Schatten norm of a real valued diagonal matrix (Eq. (A.14)) reads

$$\left\| \begin{pmatrix} \lambda_{A,1} & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \lambda_{B,1} & 0 \\ 0 & \cdots & 0 & \ddots \end{pmatrix} \right\|_p^p = \mathrm{Tr} \begin{pmatrix} \lambda_{A,1}^p & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \lambda_{B,1}^p & 0 \\ 0 & \cdots & 0 & \ddots \end{pmatrix}$$

such that it follows that we can further simplify Eq. (A.20) to

$$\begin{aligned} &= \|\Lambda_A\|_p^p + \|\Lambda_B\|_p^p && \text{(A.21)} \\ &= \|K_A\|_p^p + \|K_B\|_p^p && \text{(A.22)} \end{aligned}$$

which is the VAMP-p score of two independent systems in this particular basis. We can see that the decomposability depends on the block diagonal shape of the joint Koopman operator, which is also inherent to the covariance matrix itself. I.e., a decomposition of the covariance matrix would be possible in the same way, however its trace and Frobenius norm do not represent VAMP scores.

A.2.2 Product space decomposition

When operating in a joint space that is spanned by the tensor product, as shown above, the joint operator is formed by the Kronecker product $T_{AB} = T_A \otimes T_B$. However, the VAMP-score of a transition matrix T is not directly computed from T but from the associated Koopman operator. We first show that a decomposition of $T_{AB} = T_A \otimes T_B$ also implies a decomposition of K_{AB} in the same way. We note that that the instantaneous correlation matrices are diagonal for MSMs. In the following, we make use of the transition matrix estimator $T = C_{00}^{-1} C_{0t}$ and the mixed product rule of Kronecker products.

$$
\begin{aligned}
K_{AB} &= {}^{AB}C_{00}^{1/2} T_{AB} {}^{AB}C_{tt}^{-1/2} && \text{(A.23)}\\
&= \left({}^{A}C_{00}^{1/2} \otimes {}^{B}C_{00}^{1/2}\right) (T_A \otimes T_B) \left({}^{A}C_{tt}^{-1/2} \otimes {}^{B}C_{tt}^{-1/2}\right) && \text{(A.24)}\\
&= \left({}^{A}C_{00}^{1/2} T_A {}^{A}C_{tt}^{-1/2}\right) \otimes \left({}^{B}C_{00}^{1/2} T_B {}^{B}C_{tt}^{-1/2}\right) && \text{(A.25)}\\
&= K_A \otimes K_B && \text{(A.26)}
\end{aligned}
$$

Please note that this simple proof is only valid for indicator function basis sets such as for classical MSMs.

We can further make use of a simple rule that applies to the singular value decomposition of the Kronecker product. If the subsystem operators have n and m singular values $\lambda_{A,i} \in \mathbb{R}$ and $\lambda_{B,i} \in \mathbb{R}$, respectively, the singular values of its Kronecker product are $\{\lambda_{A,i} \cdot \lambda_{B,j} : 0 < i < n, 0 < j < m\}$. It thus follows that

$$
\begin{aligned}
\|K_{AB}\|_p^p &= \sum_i \lambda_{AB,i}^p && \text{(A.27)}\\
&= \sum_{i,j} (\lambda_{A,i} \cdot \lambda_{B,j})^p && \text{(A.28)}\\
&= \sum_i \lambda_{A,i}^p \cdot \sum_j \lambda_{B,j}^p && \text{(A.29)}\\
&= \|K_A\|_p^p \cdot \|K_B\|_p^p. && \text{(A.30)}
\end{aligned}
$$

This is the decomposition for the VAMP-p score of two independent systems in a product basis such as the one applied for MSM transition matrices.

A.3 Continuous-time MSM decomposition

Discretizations of the operator P can also yield continuous-time MSMs [6, 7]. Analogously to the analysis done with discrete time MSMs, assume two independent regions in phase space that are discretized into two continuous-time MSMs. Their solution is

$$f_A(t) = \exp(tR_A)f_{0A}, \qquad f_B(t) = \exp(tR_B)f_{0B}, \tag{A.31}$$

where R_A and R_B are the transition rate matrices; f_A and f_B the probability densities in the corresponding regions; and f_{0A} and f_{0B} the initial conditions. The operator P is approximated by the exponential functions, so the solution of the whole system is given by the tensor product of exponentials,

$$f(t) = \exp(t(R_A \oplus R_B))f_0, \tag{A.32}$$

which yields a Kronecker sum $\oplus$ for the matrices in the exponent.

In summary, in order to approximate the operator P of the full system, we need to either use the Kronecker sum on rate matrices of continuous-time MSMs, or the Kronecker product on transition probability matrices of discrete-time MSMs. In general, the full system Perron-Frobenius operator can be reassembled by using the tensor product on all the subsystems operators.

A.4 Weakly coupled systems

Practical situations – for example, an ion channel with quasi-independent subunits – might often involve weak coupling. The transition matrix $\widetilde{T}$ of a weakly coupled system can be expressed as a perturbation of the transition matrix T of the non-coupled system,

$$\widetilde{T}^{\mathrm{T}} = (1-\epsilon)T^{\mathrm{T}} + \epsilon P^{\mathrm{T}}, \quad \epsilon \in [0,1] \tag{A.33}$$

where P is another Markov transition matrix defined on the same state space as T, and $\epsilon \ll 1$ corresponds to small perturbations/weak coupling. Note this definition enforces the required MSM condition that columns sum to one.

As the eigenvalues of $\widetilde{T}$ are continuous functions of ϵ, the eigenvalues of the coupled system will be arbitrarily close to those of the uncoupled one as $\epsilon \to 0$. Further analysis on the convergence speed of the eigenvalues as $\epsilon \to 0$ is system dependent and not easy to assess in general. However, upper error bounds for the stationary distribution error exist and can be assessed in multiple ways [8, 9]. We focus on one formulation

framed in terms of mean first passage times m_{ij}, since it provides physical intuition on the sensitivity of the MSM [8, 10]. Assume T and P define finite, irreducible and homogeneous MSMs, as the MSMs of interest within the scope of this work, then

$$\|\pi - \tilde{\pi}\|_\infty \leq \frac{1}{2} \max_j \left[\frac{\max_{i \neq j} m_{ij}}{m_{jj}} \right] \|(T - \widetilde{T})^{\mathrm{T}}\|_\infty, \tag{A.34}$$

where π denotes the stationary distribution; the tilde denotes quantities of the perturbed system; the ∞-norm is the maximum absolute row sum, and m_{jj} is the mean return time of state j, i.e. the time to return to j for the first time, starting from j.

In terms of our application, if the coupling is sufficiently weak ($\epsilon \ll 1$), the eigenvalues of the uncoupled system will be close to those of the weakly coupled system, providing a good approximation of the implied timescales. Furthermore, an upper bound for the stationary distribution error can be easily calculated using software like PyEMMA [11]. The bound is very effective for MSMs consisting of a dominant central state with strong connections to and from all other states [8].

A.5 Effective counts and sampling

For comparing classical MSMs and IMD models, one can assess the total number of transition counts (going into and out of a particular state) in a global state space. It is either estimated directly based on state definitions in the global system (MSM) or computed from the Kronecker product of subsystem transition matrices (IMD model).

Let us consider two independent systems with transition matrices T_i, count matrices C_i, and total counts N_i. The latter is a diagonal matrix for classical MSMs that describes the total number of counts for each state. One can write $T_i = N_i^{-1} C_i$ (maximum likelihood estimator of the transition matrix). We can compute the total transition matrix from the Kronecker product as follows

$$T_{AB} = T_A \otimes T_B = N_A^{-1} C_A \otimes N_B^{-1} C_B = (N_A \otimes N_B)^{-1} (C_A \otimes C_B). \tag{A.35}$$

We write the global count matrix as $N_{AB} = N_A \otimes N_B$. It can be interpreted as the effective number of counts for each state in the global system when estimated from the Kronecker product, i.e., each diagonal element is the product of the sub-system total counts of a particular state. These numbers, which could be interpreted as the number of "effective transitions" in global state space, will necessarily be greater than the ones from a classical MSM in the same space.

A.6 Toy models

A.6.1 Scaling behavior: uncoupled 3 state sub-systems

A system consisting of n independent sub-systems with 3 states each was set up to exemplify scaling with number of sub-systems. The transition matrix of each sub-system is given by

$$T_i = \begin{pmatrix} 1-p & p/2 & p/2 \\ p/2 & 1-p & p/2 \\ p/2 & p/2 & 1-p \end{pmatrix} \tag{A.36}$$

with $p = 0.1$, i.e., the probability to stay in a particular state is $1 - p = 0.9$. The full system is described with a Kronecker product $T_{\text{full}} = \bigotimes_i^n T_i$. Markov chains of length N are sampled from this transition matrix using PyEMMA / msmtools [11] until the desired set of states is connected. To quantify the confidence, 30 trial runs are conducted for each number of sub-systems.

A.6.2 Approximation quality: 2 weakly coupled 2 state sub-systems

In the following, we utilize a system comprised of 2 sub-systems with 2 states each in order to exemplify the IMD framework. We further analyze its behavior with regard to limited sampling and weak couplings. The toy model consists of two sub-systems with transition matrices T_1, T_2 that each have a probability to transition to another state of $\epsilon = 0.1$.

These sub-systems are coupled in a tunable fashion. A parameter λ is introduced which results in two independent sub-systems for $\lambda = 0$ and weakly coupled subsystems for $\lambda > 0$. The full system is represented by a reversible transition matrix T for any given $\lambda \in]0, \epsilon(1-\epsilon)[$. The transition matrix of the applied toy model can explicitly be written as

$$T = \begin{pmatrix} (1-\epsilon)^2 - \lambda & \epsilon(1-\epsilon) - \lambda & \epsilon(1-\epsilon) + \lambda & \epsilon^2 + \lambda \\ \epsilon(1-\epsilon) - \lambda & (1-\epsilon)^2 - \lambda & \epsilon^2 + \lambda & \epsilon(1-\epsilon) + \lambda \\ \epsilon(1-\epsilon) + \lambda & \epsilon^2 + \lambda & (1-\epsilon)^2 - \lambda & \epsilon(1-\epsilon) - \lambda \\ \epsilon^2 + \lambda & \epsilon(1-\epsilon) + \lambda & \epsilon(1-\epsilon) - \lambda & (1-\epsilon)^2 - \lambda \end{pmatrix} \tag{A.37}$$

$$\overset{\lambda=0}{=} T_1 \otimes T_2. \tag{A.38}$$

In the un-coupled case, T reduces to the Kronecker product of the two sub-system transition matrices. The sub-system transition matrices are given by

$$T_1, T_2 = \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix}. \tag{A.39}$$

We sample discrete trajectories from T, de-compose into sub-system trajectories and estimate the models presented in Fig. A.1.

As expected, all properties of the Markov model can be easily retained in the uncoupled case (Fig. A.1). Stronger coupling yields less accurate results; especially transition probabilities are over or underestimated (Fig. A.1a) while the error on the implied timescales is comparably small, possibly yielding underestimated implied timescales (Fig. A.1b). We note that the stationary probabilities are not affected by the coupling, i.e. that $p_1 \cdot p_2 = p_{1,2}$ holds in any case (Fig. A.1c). We find that indeed, the *dependency d* in both its forms, trace and Frobenius norms, is a fast converging and significant indicator for the approximation quality (Fig. A.1d).

Due to its small size, this particular example is not suitable to demonstrate that convergence is reached faster with the decomposed model.

A.7 Dimer model

The following model system serves the purpose to demonstrate that the presented dependency scores can bisect coupled from weakly coupled systems. Our example models a dimer of protein channels. Each of those channels resembles a Hodgkin-Huxley potassium channel but possesses an additional deactivation switch. This switch alters the dynamics completely, i.e. upon activation each gate will close or stay closed with a high probability. The deactivation switch is a Markov process itself and switches state with a probability of $p_{\text{switch}} = 0.01$. Thus, each channel has strongly coupled sub-units and cannot be described by individual gate MSMs as in the previous example.

Our test system consists of two such channels. They possess some weak cooperativity which we model by a slight shift in gate opening probability if both deactivation switches are disabled at the same time.

In the following, we define a block matrix that describes the whole system dynamics. For the sake of simplicity, we present it in multiple layers. The highest layer describing the full system is given by

$$T_{\text{dimer}} = \begin{pmatrix} T\begin{pmatrix} S1:0\to 0 \\ S2:0\to 0 \end{pmatrix} & T\begin{pmatrix} S1:0\to 0 \\ S2:0\to 1 \end{pmatrix} & T\begin{pmatrix} S1:0\to 1 \\ S2:0\to 0 \end{pmatrix} & T\begin{pmatrix} S1:0\to 1 \\ S2:0\to 1 \end{pmatrix} \\ T\begin{pmatrix} S1:0\to 0 \\ S2:1\to 0 \end{pmatrix} & T\begin{pmatrix} S1:0\to 0 \\ S2:1\to 1 \end{pmatrix} & T\begin{pmatrix} S1:0\to 1 \\ S2:1\to 0 \end{pmatrix} & T\begin{pmatrix} S1:0\to 1 \\ S2:1\to 1 \end{pmatrix} \\ T\begin{pmatrix} S1:1\to 0 \\ S2:0\to 0 \end{pmatrix} & T\begin{pmatrix} S1:1\to 0 \\ S2:0\to 1 \end{pmatrix} & T\begin{pmatrix} S1:1\to 1 \\ S2:0\to 0 \end{pmatrix} & T\begin{pmatrix} S1:1\to 1 \\ S2:0\to 1 \end{pmatrix} \\ T\begin{pmatrix} S1:1\to 0 \\ S2:1\to 0 \end{pmatrix} & T\begin{pmatrix} S1:1\to 0 \\ S2:1\to 1 \end{pmatrix} & T\begin{pmatrix} S1:1\to 1 \\ S2:1\to 0 \end{pmatrix} & T\begin{pmatrix} S1:1\to 1 \\ S2:1\to 1 \end{pmatrix} \end{pmatrix}. \tag{A.40}$$

Its block elements depend on deactivation switches of the individual channels, $S1$ and $S2$. On the next layer, for each transition pair of the deactivation switches,

$$T\begin{pmatrix} S1:i\to j \\ S2:n\to m \end{pmatrix} = \begin{cases} T_c \otimes T_c & \text{if } n=m=i=j=0 \\ T(S1:i\to j)\otimes T(S2:n\to m) & \text{else.} \end{cases} \tag{A.41}$$

This implements the coupling between channels by selecting different transition probabilities if both deactivation gates are inactive at the same time. The next layer describes individual channel transition probabilities (rescaled such that the full system transition matrix has row-sum 1):

$$T(S : i \to j) = \begin{cases} p_{\text{switch}} \cdot 1_{16} & n \neq m \text{ switching switch} \\ (1 - p_{\text{switch}}) \cdot T_{\text{HH}} & n = m = 0 \text{ inactive switch} \\ (1 - p_{\text{switch}}) \cdot T_{\text{inactive}}(\lambda) & n = m = 1 \text{ active switch} \end{cases} \tag{A.42}$$

with the Hodgkin-Huxley transition matrix $T_{\text{HH}} = T_{hh} \otimes T_{hh} \otimes T_{hh} \otimes T_{hh}$ with individual gate transition matrices T_{hh} that describe gate opening and closing in the native state. Further, a transition matrix describing gate dynamics if the deactivation switch is active is given. For fully activated coupling between gates and deactivation switch, it reads $\tilde{T}_{\text{inactive}} = T_o \otimes T_o \otimes T_o \otimes T_o$ with T_o being the single gate transition matrices for that case. In order to control the intensity of the gate-deactivation switch coupling, we use a linear mixture parameter λ,

$$T_{\text{inactive}}(\lambda) = \lambda \tilde{T}_{\text{inactive}} + (1 - \lambda) T_{\text{HH}}, \tag{A.43}$$

i.e., the coupling can gradually be turned off by adjusting $\lambda \in]0, 1[$, and the case $\lambda = 0$ leaves the deactivation switch with no effect on the gate dynamics.

Finally, on the last layer, the single gate matrices are given by

$$T_{hh} = \begin{pmatrix} 0.9483 & 0.0517 \\ 0.0055 & 0.9945 \end{pmatrix} \quad \text{unperturbed} \tag{A.44}$$

$$T_o = \begin{pmatrix} 0.9483 & 0.0517 \\ 0.95 & .05 \end{pmatrix} \quad \text{active deactivation switch} \tag{A.45}$$

$$T_c = \begin{pmatrix} 0.8 & 0.2 \\ 0.0055 & 0.9945 \end{pmatrix} \quad \text{both deactivation switches inactive} \tag{A.46}$$

The Markov chain is sampled from the transition matrix T_{dimer} using PyEMMA / msmtools [11] with a time step of 20 steps for 1 million time steps. The code used to generate and analyze the example can be found in our GitHub repository.

A.8 Supplementary figures

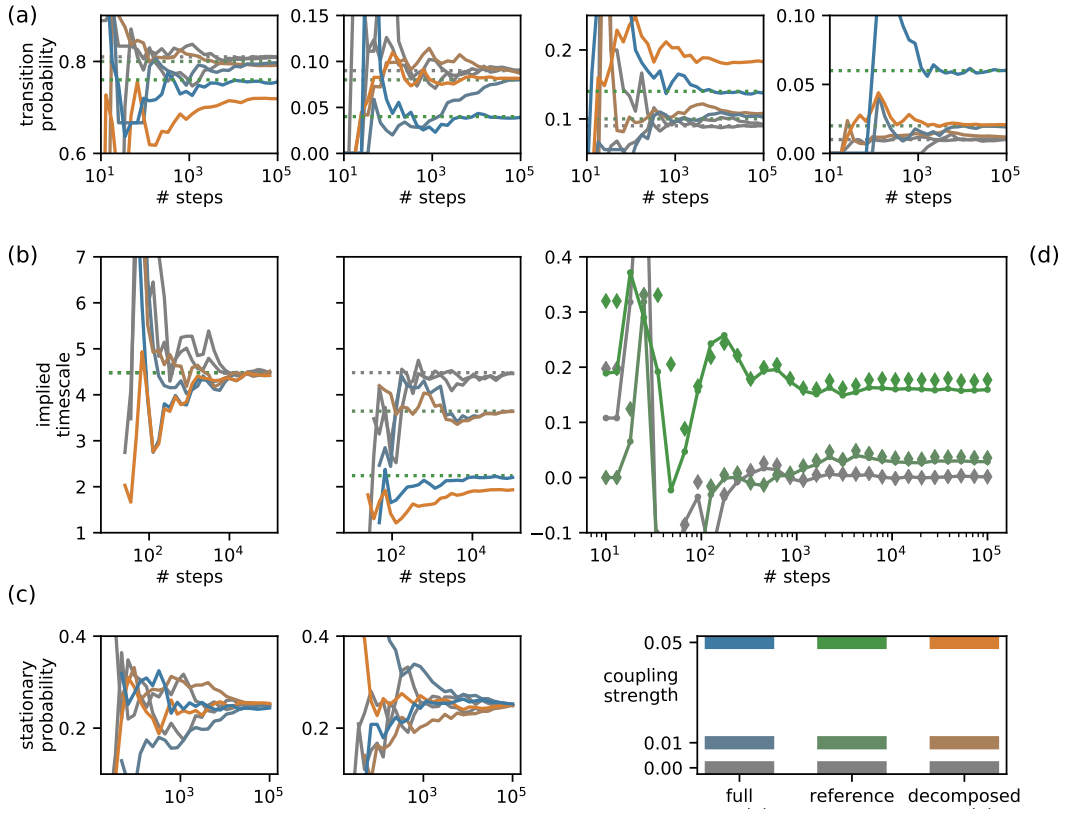


Figure A.1: Analysis of error from weak couplings and limited sampling. MSM properties of full-system and decomposed estimates are shown as functions of sampling (x-axis) and coupling (color code). **(a)** first row of transition probability matrix, **(b)** two highest implied timescales. **(c)** Stationary probabilities (shown for two example states). **(d)** *dependency* d as difference in trace norms (line) and Frobenius norms (diamonds).

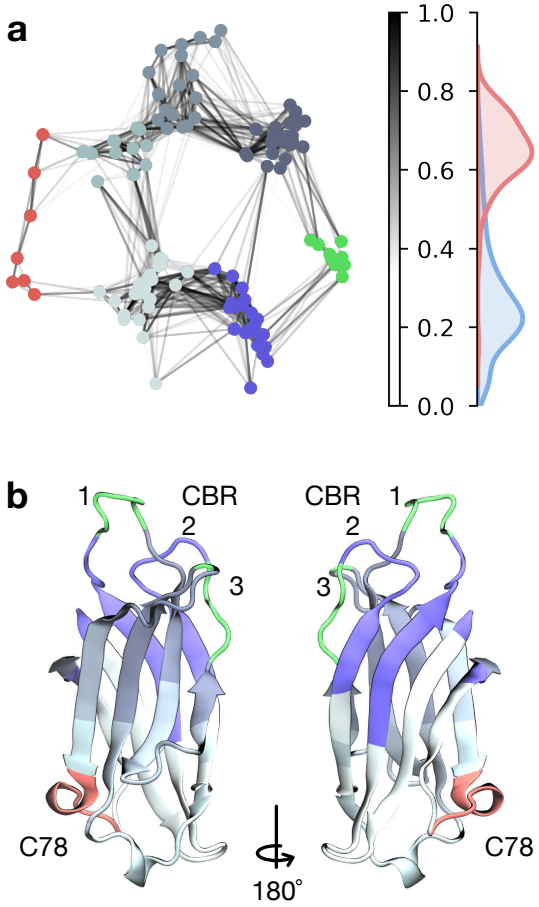


Figure A.2: Dependency-network between residues of Syt-1 C2A depicted using a standard graph layout (Fruchterman-Rheingold algorithm). **(a)** VAMP-1 normalized dependency network. Edge weights are indicated by colorbar. Nodes are colored according to an unsupervised classification by the k-means algorithm ($k = 7$). **(b)** Visualization of protein structure with color coded segments from our VAMP-1 analysis, i.e., same color code as in panel a.

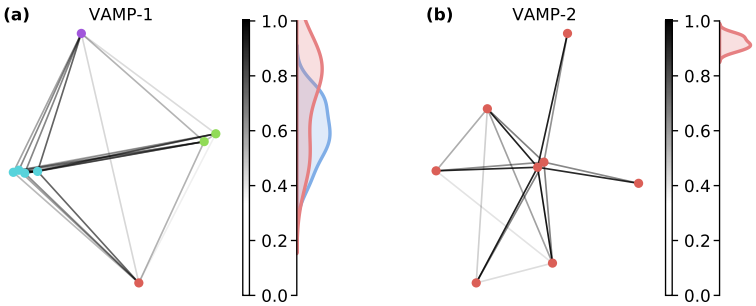


Figure A.3: Counterexample to IMD with *dependency*-network between residues of Chignolin [12]. Analysis is based on flexible torsion angles [13]. We show VAMP-1 **(a)** and VAMP-2 **(b)** normalized *dependency* networks. Edge weights are indicated by colorbar. **(a)** VAMP-1 dependency network with nodes colored according to an unsupervised classification by the k-means algorithm ($k = 4$). *Dependency* histograms depict coupling strength of residues within a subsystem cluster (red) and between different subsystem clusters (blue). Note that links between residue clusters express high normalized *dependency* scores, which is also mirrored in the two distributions having significant overlap. Therefore, the peptide cannot be split into independent subsystems. **(b)** VAMP-2 *dependency* network shows no clustering; every residue is connected to the network with scores > 0.8, further indicating that Chignolin cannot be modeled with IMD.

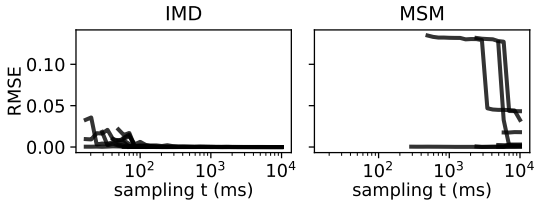


Figure A.4: Deviations of Hodgkin-Huxley ion channel models (IMD, MSM) from the ground truth, assessed with Root Mean Square Error (RMSE). RMSE is computed between estimated eigenvalue spectrum (IMD, MSM) and spectrum of the generator transition matrix (ground truth) for all cases where connected transition matrices could be estimated.

Bibliography

[1] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard. “A Direct Approach to Conformational Dynamics Based on Hybrid Monte Carlo”. *J. Comput. Phys.* 151.1 (1999), pp. 146–168.

[2] 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.* 134.17 (2011), p. 174105.

[3] I. Satake. *Linear Algebra*. Pure and Applied Mathematics. New York: Dekker, 1975.

[4] S. Klus, P. Koltai, and C. Schütte. “On the Numerical Approximation of the Perron-Frobenius and Koopman Operator”. *J. Comput. Dyn.* 3.1 (2016), pp. 1–12.

[5] H. Wu and F. Noé. “Variational Approach for Learning Markov Processes from Time Series Data”. *J Nonlinear Sci* (2019).

[6] N.-V. Buchete and G. Hummer. “Coarse Master Equations for Peptide Folding Dynamics”. *J. Phys. Chem. B* 112.19 (2008), pp. 6057–6069.

[7] D. De Sancho and A. Aguirre. “MasterMSM: A Package for Constructing Master Equation Models of Molecular Dynamics” (2019).

[8] G. E. Cho and C. D. Meyer. “Comparison of Perturbation Bounds for the Stationary Distribution of a Markov Chain”. *Linear Algebra Its Appl.* 335.1-3 (2001), pp. 137–150.

[9] C. D. Meyer. “Sensitivity of the Stationary Distribution of a Markov Chain”. *SIAM J. Matrix Anal. & Appl.* 15.3 (1994), pp. 715–728.

[10] G. E. Cho and C. D. Meyer. “Markov Chain Sensitivity Measured by Mean First Passage Times”. *Linear Algebra and its Applications*. Special Issue: Conference Celebrating the 60th Birthday of Robert J. Plemmons 316.1 (2000), pp. 21–28.

[11] M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz, and F. Noé. “PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models”. *J. Chem. Theory Comput.* 11.11 (2015), pp. 5525–5542.

[12] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw. “How Fast-Folding Proteins Fold”. *Science* 334.6055 (2011), pp. 517–520.

[13] M. K. Scherer, B. E. Husic, M. Hoffmann, F. Paul, H. Wu, and F. Noé. “Variational Selection of Features for Molecular Kinetics”. *J. Chem. Phys.* 150.19 (2019), p. 194108.

B

Supplemental information: Coupling of conformational switches in calcium sensor unraveled with local Markov models and transfer entropy

This appendix has been published as supplementary material to

Tim Hempel, Nuria Plattner, and Frank Noé. "Coupling of Conformational Switches in Calcium Sensor Unraveled with Local Markov Models and Transfer Entropy". *Journal of Chemical Theory and Computation* 16.4 (2020), pp. 2584–2593.

Reprinted with permission from T. Hempel, N. Plattner, and F. Noé. "Coupling of Conformational Switches in Calcium Sensor Unraveled with Local Markov Models and Transfer Entropy". *J. Chem. Theory Comput.* 16.4 (2020), pp. 2584–2593. Copyright 2020 American Chemical Society.

B.1 Additional Syt-1 C2A analyses

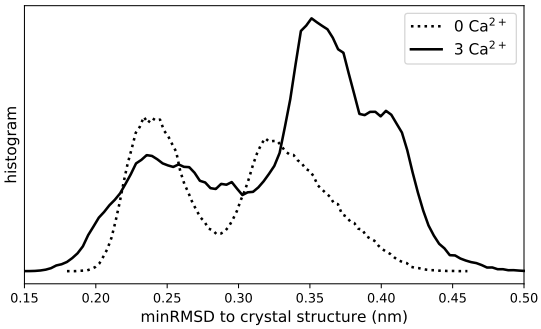


Figure B.1: Histograms of minimum root-mean-square deviation (RMSD) between calcium bound and unbound trajectories to their respective PDB structures (1BYN: bound; 2R83: unbound). The multi-peak structure shows that besides the crystal structure, reasonably populated conformations exist which become more populated in the bound case.

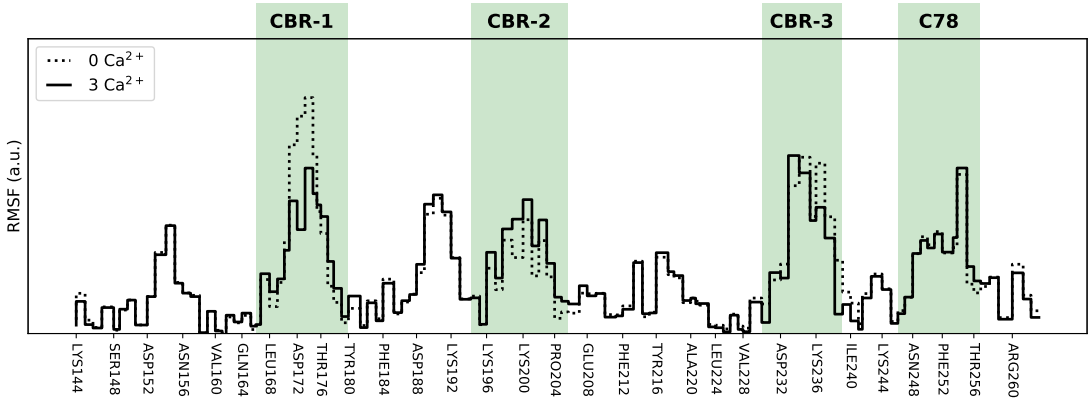


Figure B.2: Syt-C2A RMSF per residue as computed from the full set of trajectories. Several regions appear to have significant motion, most of them correspond to CBR loops.

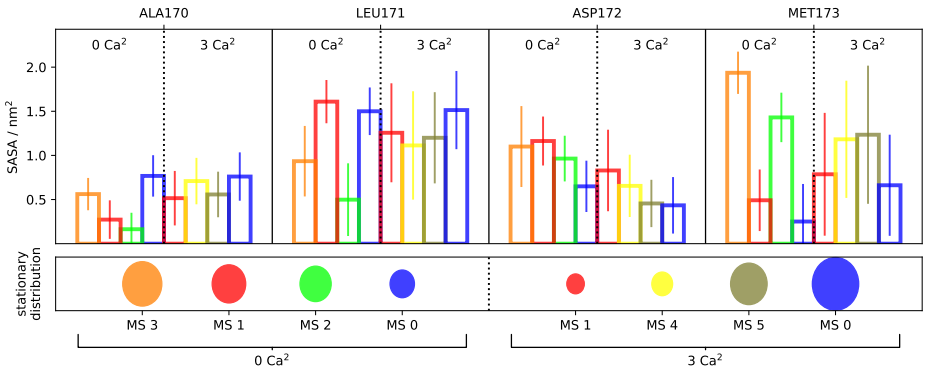


Figure B.3: SASA of charged polar or hydrophobic residues in CBR-1 (upper panel). The bars are color-coded according to the convention used throughout this paper. Stationary probabilities are added (lower panel). SASA and stationary distributions of calcium bound and unbound states are sorted left and right of the vertical dotted lines for each residue.

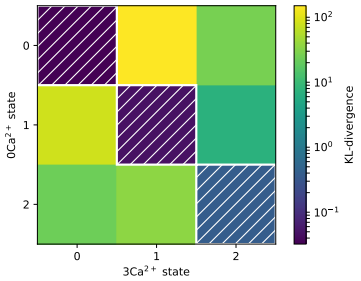


Figure B.4: HMM macrostate identification: KL-divergences between state observation probabilities per macrostate for calcium bound and unbound datasets. Both HMMs have 3 hidden states. White hatches depict states that are identified between datasets.

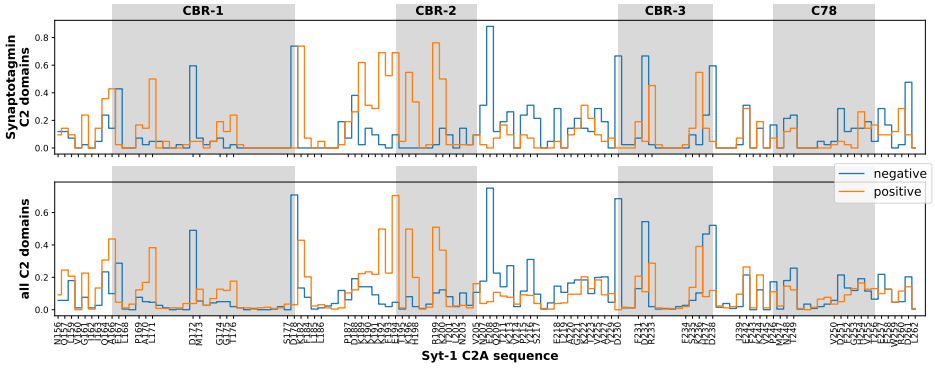


Figure B.5: Conservation of positive and negative charges. Bottom: Sequence alignment of C2 domain family members (PFAM entry PF00168, seed). Top: Sub-sample of synaptotagmin C2 domains.

B.2 Method supplement

B.2.1 openMM run script

```
import simtk.openmm.app as app
import simtk.openmm as mm
import simtk.unit as u
from parmed import gromacs

# input files
topfile = 'path/to/gromacs_top_file.top'
grofile = 'path/to/gromacs_gro_file.gro'

# user parameters
integrator_timestep_ps = 0.005  # picoseconds
simulation_time_ns = 2000  # nanoseconds
equilibration_steps = 1500 # steps

# physical values:
temperature = 300 * u.kelvin
pressure = 1 * u.bar


# load pdb, force field and create system
gromacs.GROMACS_TOPDIR = 'path/to/gromacs-5.1.1/share/top'
top = gromacs.GromacsTopologyFile(topfile)
gro = gromacs.GromacsGroFile.parse(grofile)
top.box = gro.box

# create system object
system = top.createSystem(
    nonbondedMethod=app.PME,
    nonbondedCutoff=1.0 * u.nanometer,
    constraints=app.AllBonds,
    rigidWater=True,
    hydrogenMass=4 * u.amu,
    ewaldErrorTolerance=0.0005)

# integrator
integrator = mm.LangevinIntegrator(
    temperature,
    1/u.picoseconds,
    integrator_timestep_ps * u.picoseconds)
integrator.setConstraintTolerance(1e-5)

platform = mm.Platform.getPlatformByName('CUDA')
properties = {'CudaPrecision': 'mixed'}
```

```
# pressure coupling
system.addForce(mm.MonteCarloBarostat(pressure, temperature))

# initialize simulation object
simulation = app.Simulation(
    top.topology, system, integrator,
    platform, properties)

# load positions and velocities
simulation.context.setPositions(gro.positions)
simulation.context.setVelocitiesToTemperature(temperature)

# HBond constraints
simulation.context.applyConstraints(1e-12)

# minimize energy
simulation.minimizeEnergy(maxIterations=1000)
simulation.step(equilibration_steps)

# start simulation
simulation.step(int(simulation_time_ns /
                    (integrator_timestep_ps * 1e-3)))
```

B.2.2 Calcium position convergence

We have assessed if the calcium ions have relaxed to equilibrium by computing the start and end distributions of the RMSD to the crystal structure. After superposition of the protein to the crystal structure, RMSD of calcium ion coordinates to the crystal structure was computed for the first and last 10 ns of each trajectory (individual trajectory length is 2 μs). We note no significant difference between the start and end distributions.

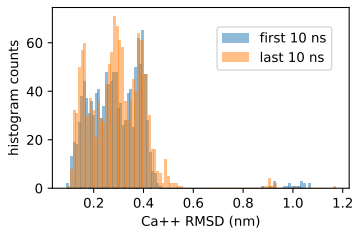


Figure B.6: Histograms of calcium ion position RMSDs from crystal structure. Compared are the first 10 ns (blue) to the last 10 ns (orange) of each trajectory.

B.2.3 Local clustering

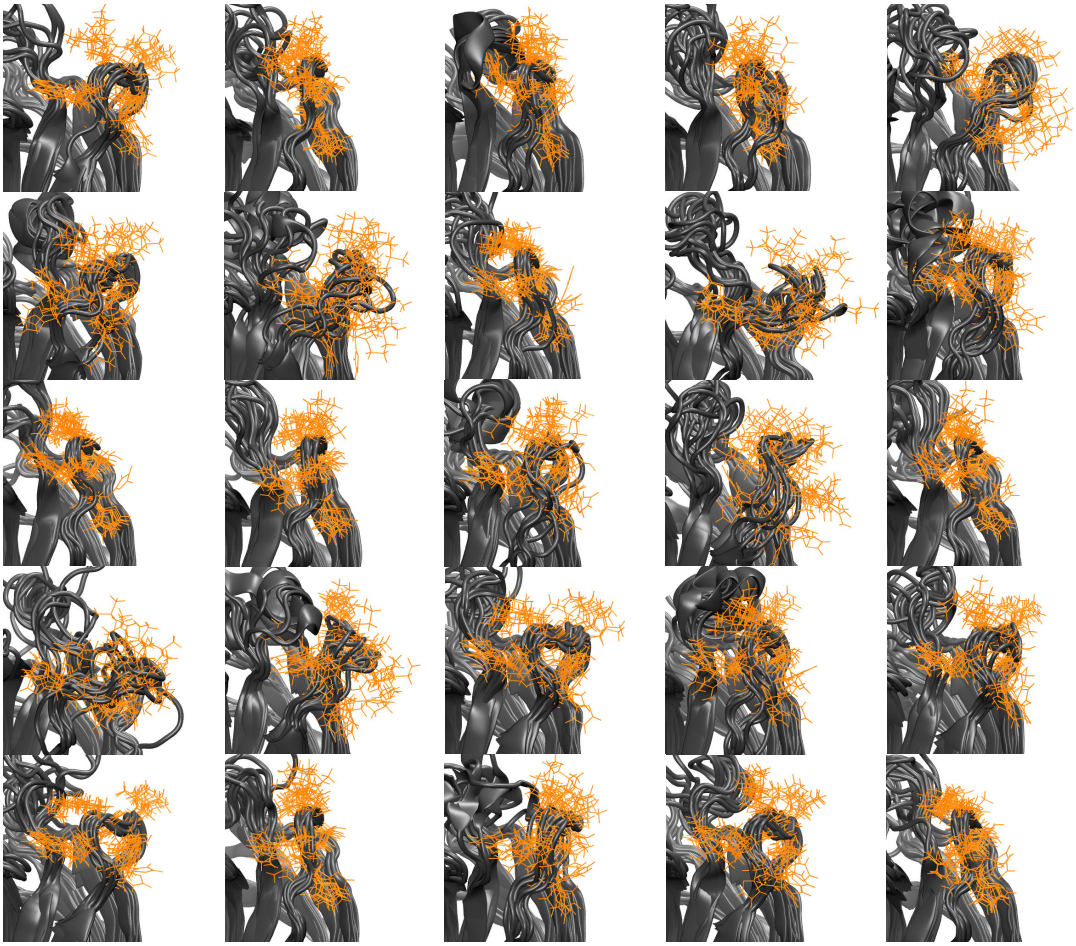


Figure B.7: Representative structures of microstates of CBR-2 obtained from k-means clustering. Residues AR199-N203 that were used for discretization are color coded in orange. Microstates mirror internal configurations of CBR-2 as well as its distance to the protein body.

B.2.4 Markov model validation

Generally, when building a Markov model, it needs to be checked if the process in discrete state space is Markovian for a certain lag time τ. This can be done by checking for the convergence of model properties with respect to the model parameter τ [1] and by testing for the Chapman-Kolmogorov equation.

Implied timescales and stationary properties convergence. When sampling rare processes in a large data set however, one faces the problem that processes can be

lost at high time scales. This is why the model parameter was tuned to a value which allows the implied timescales and the stationary distribution of all six models to be constant within error.

As depicted in Fig. B.8, especially in the case of calcium bound CBR-1, the interval between timescale convergence and loosing the process is very short. The reason is that this rare event is sampled poorly. Nevertheless it is assumed that this model is valid since the chosen lag time of $\tau = 50$ ns ensures Markovianity in all of the other models. Further, the behavior is reflected in the error estimate of the derived properties such as stationary distribution and mean first passage times (MFPT).

Chapman-Kolmogorov test Consistency of the Chapman-Kolmogorov equation $T(n \cdot \tau) = T(\tau)^n$ is tested for all the models presented here. As already mentioned, we are operating in the data sparse regime, so estimates can only be done for a finite number of multiples of the lag time τ.

Predictions as well es estimates from the presented models are depicted in Fig. B.9. Generally, we note that transition probabilities have the same trends, i.e. stay in the same order of magnitude. The presented error is estimated by Bayesian sampling of the posterior, it shows significant overlap in all cases. However, data sparsity leads to divergence at lag times of $3 \cdot \tau = 150$ ns.

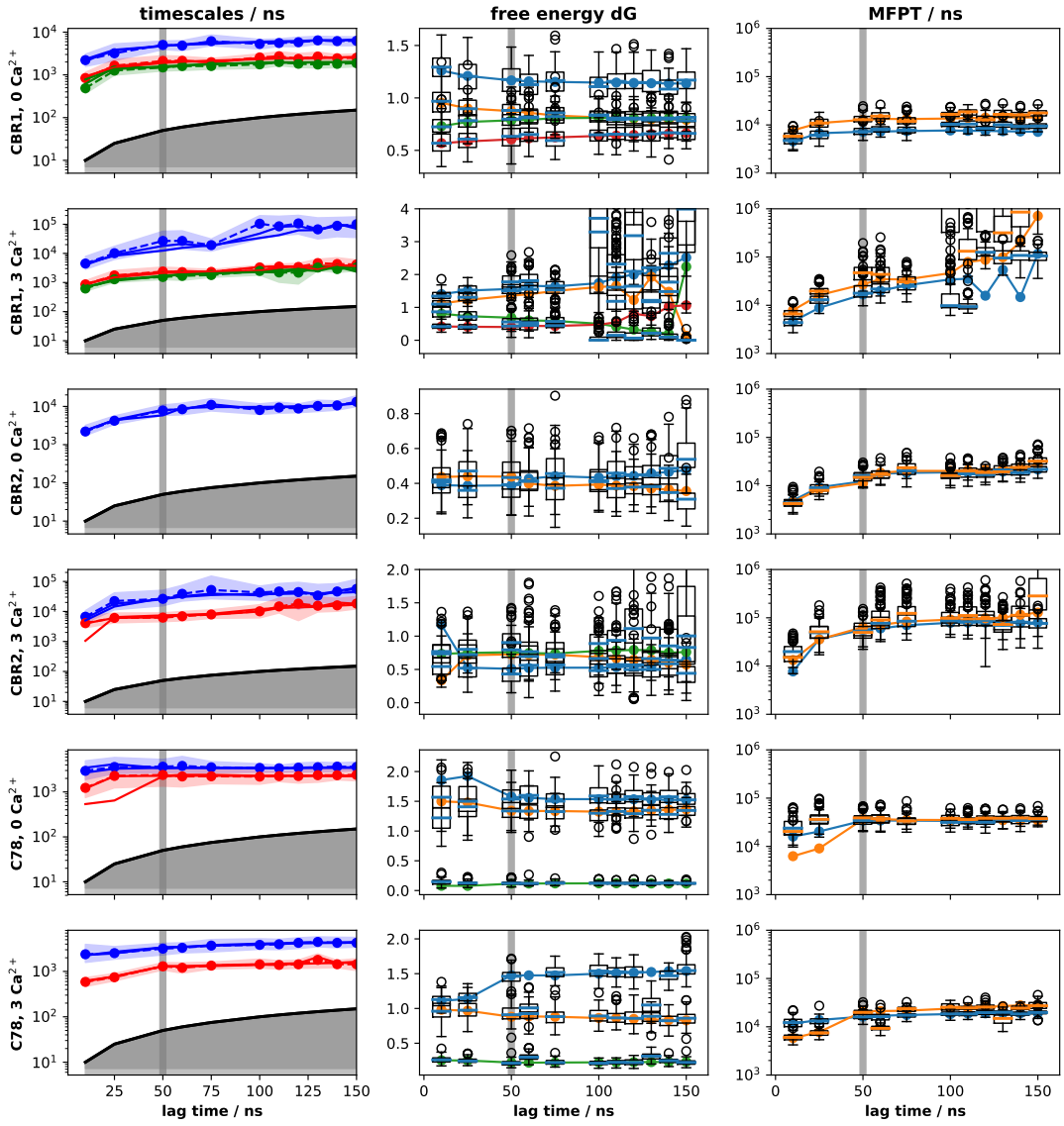


Figure B.8: Convergence of implied timescales (left column), relative free energy (middle) and mean first passage times between arbitrarily chosen states (right).

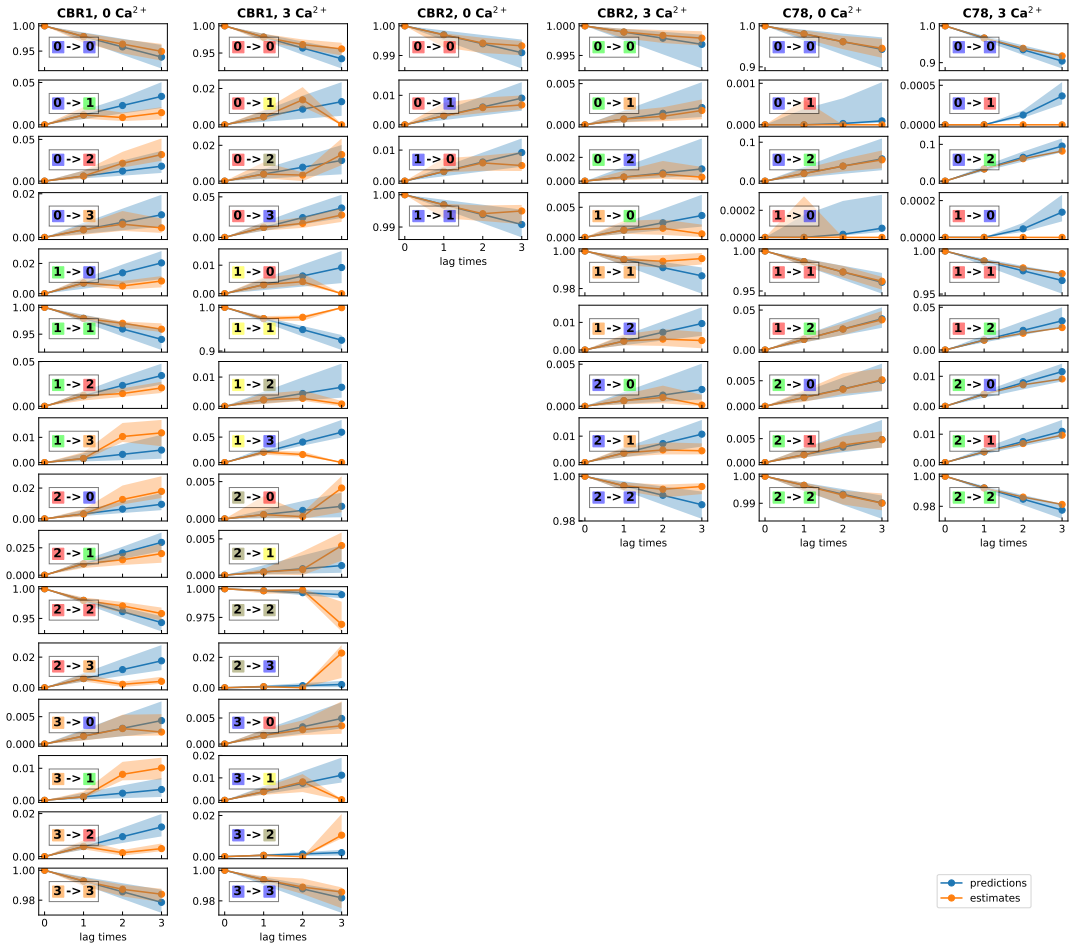


Figure B.9: Result of the Chapman-Kolmogorov test column-wise ordered by model. Predictions from the presented models are depicted in blue, new estimates in orange. Transitions are encoded in macro state numbers and colored as in the results section.

B.2.5 Validation of transfer entropy and mutual information

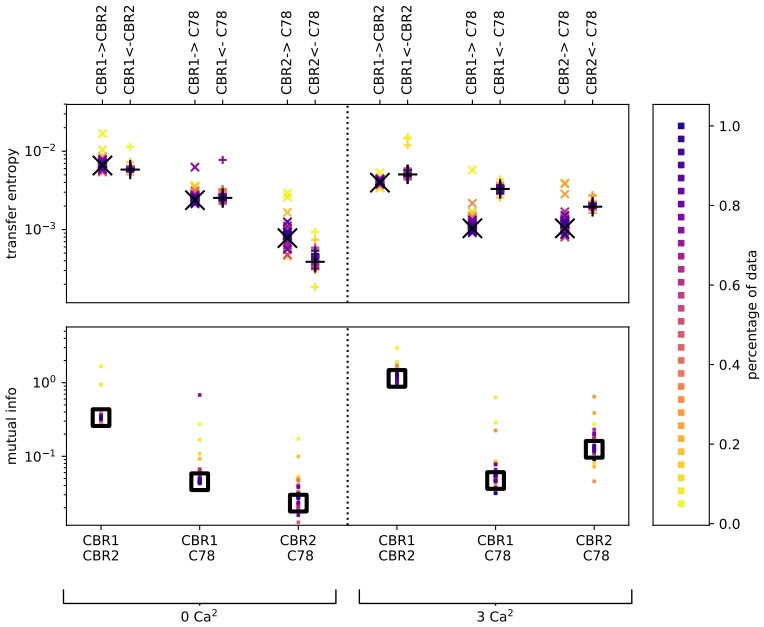


Figure B.10: Bootstrapping validation for mutual information (bottom) and transfer entropy (top). Estimates using the full set of data are depicted with large black symbols. Left panels show calcium unbound, right panels calcium bound data.

Mutual information and transfer entropy were validated using bootstrapping of trajectory data (cf. Fig. B.10). In order to assess if the results are significantly different to zero, a comparison was made to shuffled trajectories, i.e. the time information within the trajectories was kept constant while trajectories were combined that did not happen at the same time. As depicted in Fig. B.11, the results in this case are at least one order of magnitude smaller than the results in the correct time frame.

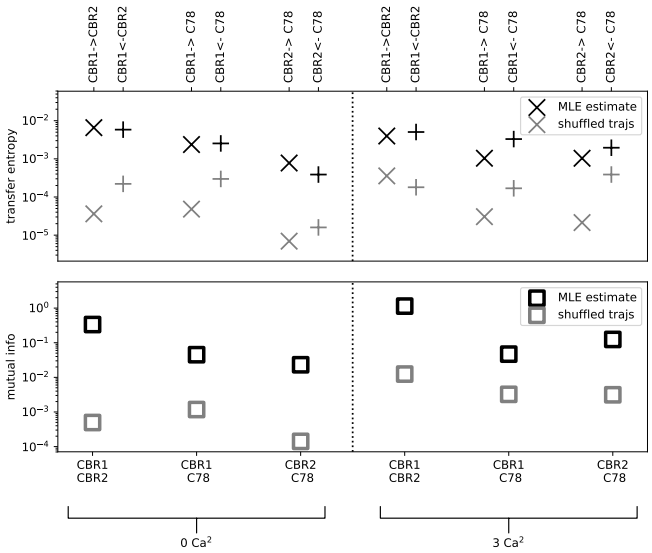


Figure B.11: Mutual information and transfer entropy validation by comparing to results obtained from shuffling trajectories among each other but keeping frames within the single sub-system trajectories.

B.2.6 Ion model validation by Alchemical free energy perturbation

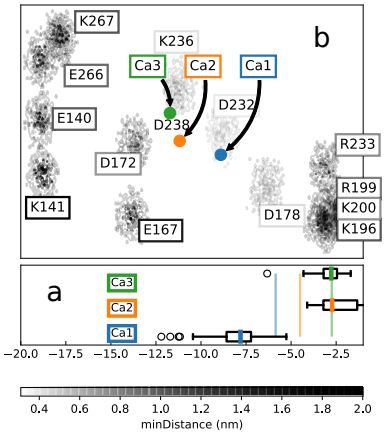


Figure B.12: Alchemical free energy computations and binding pocket projection for crystal structure. Box plots of free energy results shows that experimental values (solid vertical lines) can be matched within the error. The only exception is ion Ca2 which in terms of its binding free energy is indistinguishable from Ca1. Configuration of (color coded) calcium ions is shown within binding pocket of crystal structure.

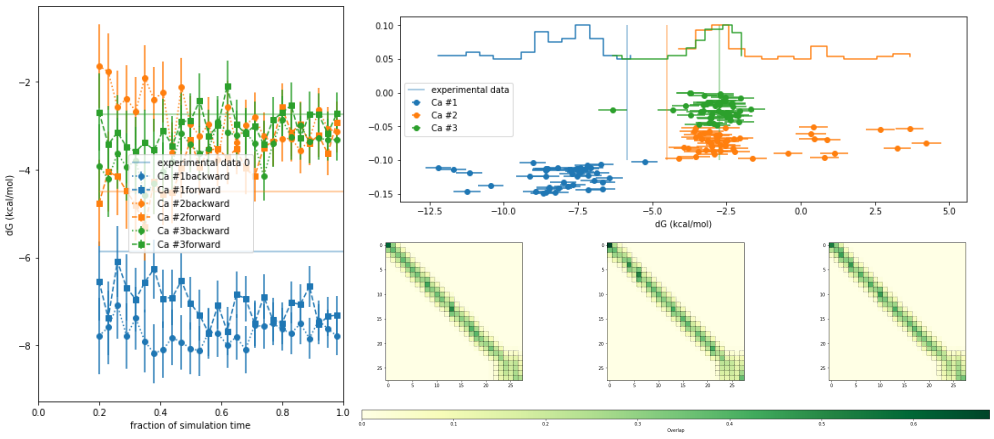


Figure B.13: Validation of alchemical free energy computations with MBAR using crystal structure. Left: Forward-backward convergence. Top right: results of about 50 indipendent computations. Bottom right: Overlap matrices with boxes denoting sufficient overlap according to Ref. [2].

Bibliography

[1] 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.* 134.17 (2011), p. 174105.

[2] P. V. Klimovich, M. R. Shirts, and D. L. Mobley. "Guidelines for the Analysis of Free Energy Calculations". *J. Comput. Aided Mol. Des.* 29.5 (2015), pp. 397–411.

C

Supplemental information: Molecular mechanism of inhibiting the SARS-CoV-2 cell entry facilitator TMPRSS2 with Camostat and Nafamostat

This appendix has been published as supplementary material to

Tim Hempel, Lluís Raich, Simon Olsson, Nurit P. Azouz, Andrea M. Klingler, Markus Hoffmann, Stefan Pöhlmann, Marc E. Rothenberg, and Frank Noé. "Molecular Mechanism of Inhibiting the SARS-CoV-2 Cell Entry Facilitator TMPRSS2 with Camostat and Nafamostat". *Chemical Science* (2021), 10.1039.D0SC05064D.

This Chapter is licensed under the Creative Commons Attribution Non-Commercial 3.0 Unported License. To view a copy of this license, visit https://creativecommons.org/licenses/by-nc/3.0/.

C.1 Supplementary figures

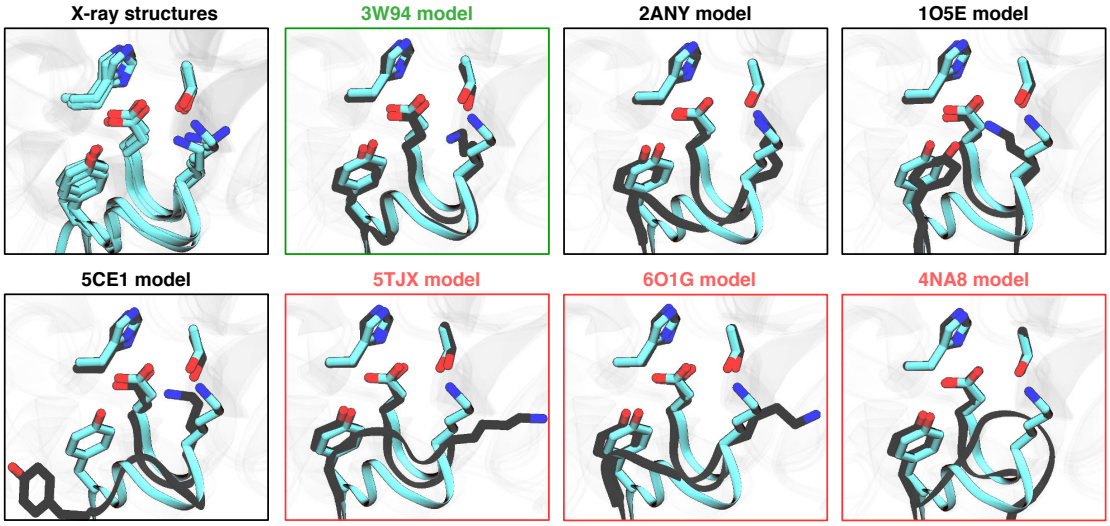


Figure C.1: Selection of homology model from Ref. 38 by specific interactions around the catalytic triad. Comparison of TMPRSS2 models (black) and four serine protease structures that contain a lysine residue next to the catalytic aspartate (cyan, PDBs 1EKB, 1FUJ, 3W94 and 4DGJ). Note that the four crystal structures show a very conserved and rigid environment around the catalytic aspartate, with just few fluctuations of the lysine head (K99 in 1EKB). The structural models of TMPRSS2, instead, show a wide variability of conformations of both the backbone and sidechains, with 3W94 being the most conservative model.

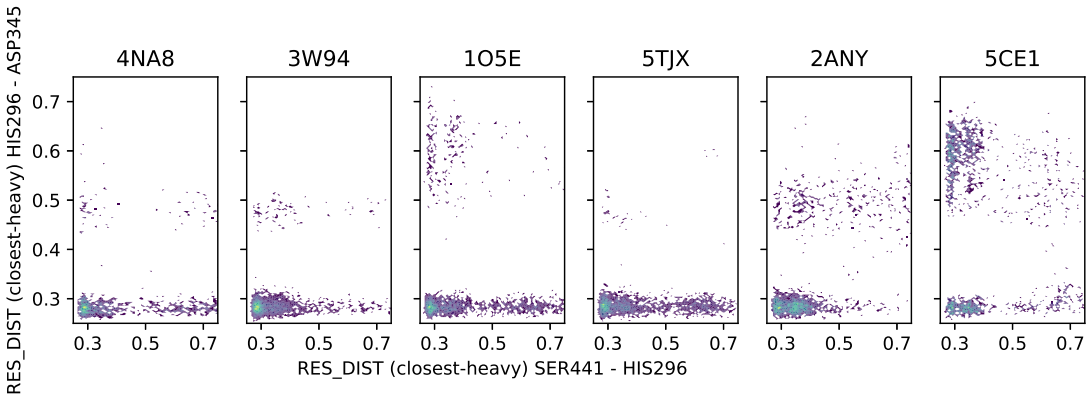


Figure C.2: Relevant distances of the catalytic triad for different homology models from Rensi et al. [1] as computed from roughly $30\mu s$ of MD data for the drug free protein.

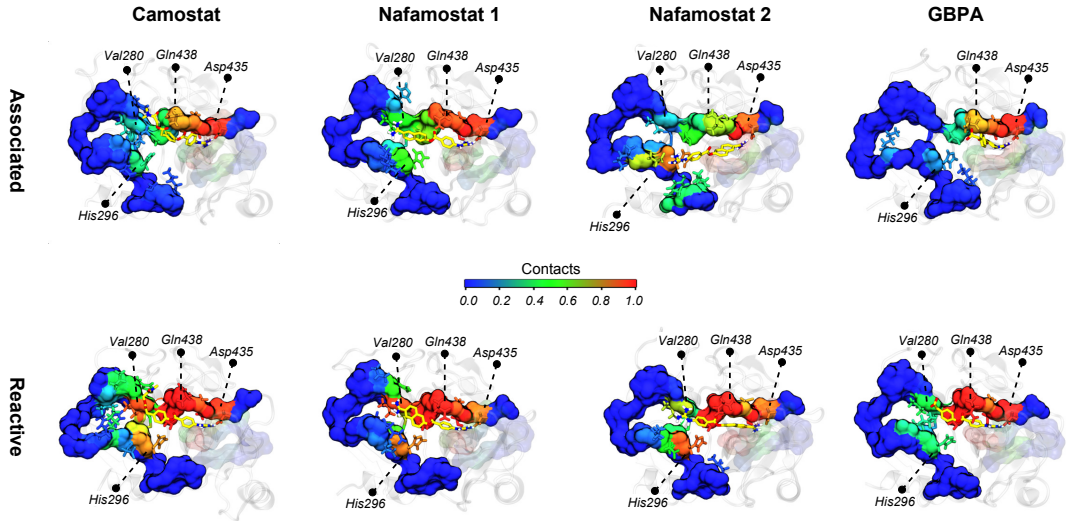


Figure C.3: Contact frequency map of TMPRSS2 with camostat, nafamostat, and the camostat product GBPA. Contacts are defined from any atom of the drug to any atom of TMPRSS2 with a distance below 0.35 nm. Residues above with binding frequencies above 0.05 are shown as colored sticks.

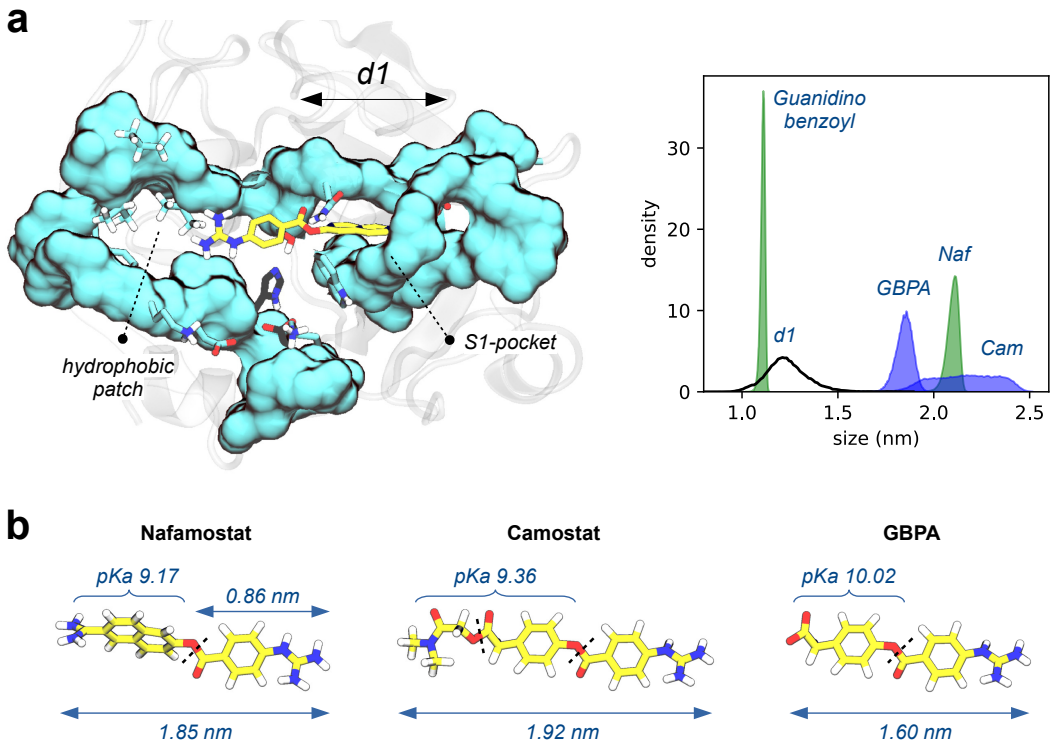


Figure C.4: Distribution of TMPRSS2 cavity lengths and end-to-end drug sizes. **(a)** histogram of drug end-to-end distances and S1 pocket length (d1, minimal distance Ser441(Og)-Asp435(Od1/Od2)). Nafamostat end-to-end distance is show in green, together with the size of the guanidinobenzoyl moiety. Camostat and its metabolized form (GBPA) are shown in blue. Note the wide distribution of camostat, indicating a high degree of flexibility going from compact to extended conformations. Distributions of the drugs are shifted by 0.25 nm to account for the distance of hydrogen bonds. **(b)** structures of nafamostat, camostat, and GBPA. Note that the three drugs share the same guanidinobenzoyl moiety (right part of each molecule). Hydrolyzable bonds are indicated by dashed lines. Average end-to-end distances by two headed lines. The pKa value of the leaving group (phenol) is shown above each moiety, as predicted by Schrödinger Epik (Schrödinger Release 2020-2).

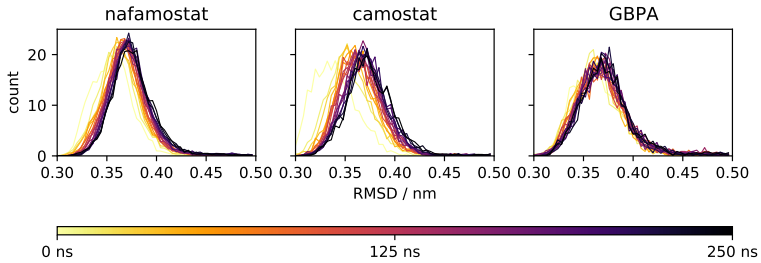


Figure C.5: Convergence of root mean squared deviation (RMSD) with trajectory time. As all three datasets include 100s of trajectories, we show the time evolution of RMSD histograms over all available trajectories of a given drug (time as displayed by color code). Please note that the distributions converge over time.

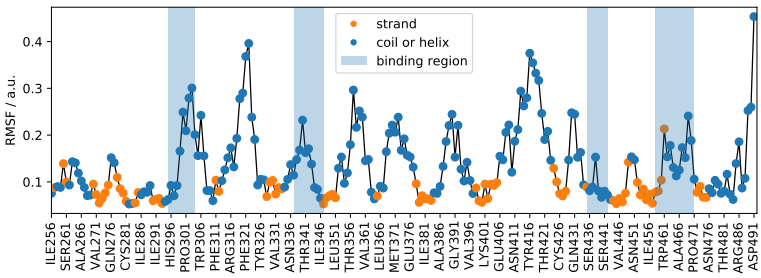


Figure C.6: Per-residue root mean squared fluctuations (RMSF) along the protein sequence. The protein core, mainly consisting of β-strands, has a low RMSF, i.e. is rigid compared to coil or helical protein segments such as a large part of the binding region.

Bibliography

[1] S. Rensi, R. B Altman, T. Liu, Y.-C. Lo, G. McInnes, A. Derry, and A. Keys. *Homology Modeling of TMPRSS2 Yields Candidate Drugs That May Inhibit Entry of SARS-CoV-2 into Human Cells*. Preprint. 2020. url: https://chemrxiv.org/articles/Homology_Modeling_of_TMPRSS2_Yields_Candidate_Drugs_That_May_Inhibit_Entry_of_SARS-CoV-2_into_Human_Cells/12009582.

D

Supplemental information: Deep learning to decompose macromolecules into independent Markovian domains

This appendix has been published as supplementary material to

Andreas Mardt*, Tim Hempel*, Cecilia Clementi, and Frank Noé. "Deep Learning to Decompose Macromolecules into Independent Markovian Domains". *Nature Communications* 13.1 (2022), p. 7101.

D.1 Supplementary Note 1: Independent Koopman operators

The true Koopman operator $\mathcal{K}_\tau$ can be written in two ways. First, if the operator is a Hilbert-Schmidt operator, the following singular value decomposition (SVD) exists:

$$\mathcal{K}_\tau g(\mathbf{x}) = \sum_{i=1}^{\infty} \sigma_i \langle g, \phi_i \rangle_{\rho_1} \psi_i(\mathbf{x}). \tag{D.1}$$

A low-rank approximation to the Koopman operator is obtained by truncating the sum after $k \ll \infty$ terms.

This Chapter is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

Second, the Koopman operator can be expressed via the transition density $p_\tau(\mathbf{y}|\mathbf{x})$ that describes the transitions from configuration $\mathbf{x}$ to $\mathbf{y}$ within a time window τ:

$$\mathcal{K}_\tau g(\mathbf{x}) = \int p_\tau(\mathbf{y}|\mathbf{x}) g(\mathbf{y}) \mathrm{d}\mathbf{y}. \tag{D.2}$$

Given two independent systems with configurations $\mathbf{x}^1, \mathbf{x}^2$ and $\mathbf{y}^1, \mathbf{y}^2$ and their transition densities $p^1_\tau(\mathbf{y}^1|\mathbf{x}^1)$, $p^2_\tau(\mathbf{y}^2|\mathbf{x}^2)$, respectively, the global transition density is then:

$$p^G_\tau(\mathbf{y}^1, \mathbf{y}^2|\mathbf{x}^1, \mathbf{x}^2) = p^1_\tau(\mathbf{y}^1|\mathbf{x}^1) \cdot p^2_\tau(\mathbf{y}^2|\mathbf{x}^2). \tag{D.3}$$

As a consequence, if we study observables which can be expressed as the product of the subsystem specific observables (hence the Kronecker product of the feature functions χ) $g^G(\mathbf{x}^1, \mathbf{x}^2) = g^1(\mathbf{x}^1) g^2(\mathbf{x}^2)$, the global Koopman operator decomposes into the product of subsystem operators:

$$\begin{aligned} \mathcal{K}^G_\tau g^G(\mathbf{x}^1, \mathbf{x}^2) &= \iint p^G_\tau(\mathbf{y}^1, \mathbf{y}^2|\mathbf{x}^1, \mathbf{x}^2) g^G(\mathbf{y}^1, \mathbf{y}^2) \mathrm{d}\mathbf{y}^1 \mathrm{d}\mathbf{y}^2 \\ &= \int p^1_\tau(\mathbf{y}^1|\mathbf{x}^1) g^1(\mathbf{y}^1) \mathrm{d}\mathbf{y}^1 \int p^2_\tau(\mathbf{y}^2|\mathbf{x}^2) g^2(\mathbf{y}^2) \mathrm{d}\mathbf{y}^2 \\ &= \mathcal{K}^1_\tau g^1(\mathbf{x}^1) \mathcal{K}^2_\tau g^2(\mathbf{x}^2). \end{aligned} \tag{D.4}$$

The low rank approximation of the decomposed Koopman operator (defined in Eq. (D.1)) for independent systems can now be written as follows, taking into account only k^1 and k^2 singular functions:

$$\begin{aligned} \hat{\mathcal{K}}^G_\tau g^G(\mathbf{x}^1, \mathbf{x}^2) &= \hat{\mathcal{K}}^1_\tau g^1(\mathbf{x}^1) \hat{\mathcal{K}}^2_\tau g^2(\mathbf{x}^2) \\ &= \sum_{i=1}^{k^1} \sigma^1_i \langle g^1, \phi^1_i \rangle_{\rho^1_1} \psi^1_i(\mathbf{x}^1) \sum_{j=1}^{k^2} \sigma^2_j \langle g^2, \phi^2_j \rangle_{\rho^2_1} \psi^2_j(\mathbf{x}^2) \\ &= \sum_{i=1}^{k^1} \sum_{j=1}^{k^2} \sigma^1_i \sigma^2_j \langle g^1 g^2, \phi^1_i \phi^2_j \rangle_{\rho^G_1} \psi^1_i(\mathbf{x}^1) \psi^2_j(\mathbf{x}^2) \\ &= \sum_{l=1}^{k^1 k^2} \sigma^G_l \langle g^G, \phi^G_l \rangle_{\rho^G_1} \psi^G_l, \end{aligned} \tag{D.5}$$

Therefore in case of independent systems, the optimal singular functions and values of the *global* system are given by the Kronecker product of the *subsystem* singular functions and values, $\sigma^G_l = \sigma^1_i \sigma^2_j$, $\psi^G_l = \psi^1_i \psi^2_j$, and $\phi^G_l = \phi^1_i \phi^2_j$. This procedure can be applied to arbitrarily many independent subsystems.

D.2 Supplementary Note 2: Global model of 3x2 benchmark system

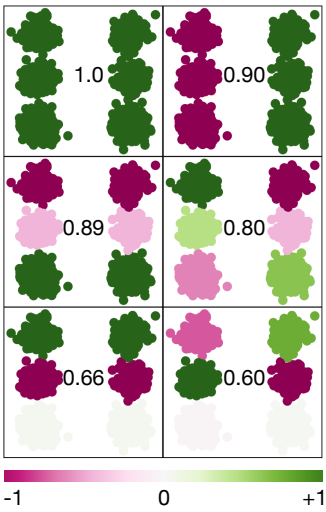


Figure D.1: Hidden Markov state model as a benchmark example for independent subsystems: The 6 global eigenfunctions supplied with their eigenvalues revealing the 4 independent processes and the 2 resulting mixed product processes. Eigenvalues of the latter are computed from the product of independent process eigenvalues: $\lambda_4 = \lambda_2 \cdot \lambda_3 = 0.80$ and $\lambda_6 = \lambda_2 \cdot \lambda_5 = 0.60$

D.3 Supplementary Note 3: Global model of Syt-C2A

To compare iVAMPnets to existing methods, we estimate a classical (global) VAMPnet model with 8 output nodes and no attention mechanism, but otherwise the same hyperparameters that were used for the iVAMPnet estimation (lag time, batch size, architecture, and training routine).

The training score converges to the theoretical maximum of 8. However, when projecting on the eigenfunctions, it becomes apparent that some of them do not describe any real transition event, i.e., stay constant during each single trajectory. Rather, these eigenfunctions model disconnected configurations that belong to different trajectories, which is an artifact of sparse sampling seeded from multiple distinct configurations. It manifests in the implied timescales (Fig. D.2) that become infinite for these processes, i.e., the eigenvalues are ≈ 1, resulting in numerically unstable implied timescales calculations.

These results imply that in order to model the global Syt-C2A system with classical VAMPnets, more simulation data has to be collected to connect these structures, i.e., the amount of data is not sufficient to build a global model. The iVAMPnet model does not

suffer from these shortcomings because products of the mentioned disconnected processes will not be observable in the data, which would result in a lower score. Therefore, the VAMP-E score will favor the processes which are truly observed.

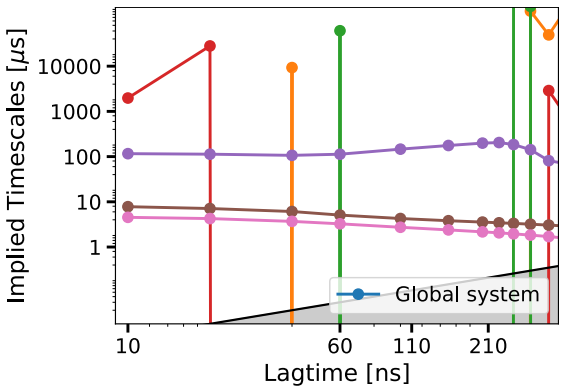


Figure D.2: Failed global implied timescales test with a classical VAMPnet of synaptotagmin with 8 output nodes. Since the model resolves processes which are not connected, the eigenvalues are ≈ 1, making the implied timescales estimation numerical unstable. It indicates that the amount of data is insufficient to build a global model. The model has no attention mechanism, but utilizes otherwise the same hyper-parameter than the ones used in iVAMPNets.

D.4 Supplementary Note 4: iVAMPnet model of Syt-C2A

The iVAMPnet model of synaptotagmin presented in this paper identifies two distinct subsystems that roughly correspond to the calcium binding region (CBR) and the opposite site of the protein, in particular, two loops that we call C34 and C78. In the following, we discuss the features of our iVAMPnet models that describe these subsystems individually. To that end, we make use of the system description provided in our earlier publication [1]. Furthermore, we approximate transition probabilities and stationary probabilities by computing the transition operator $C_{00}^{-1}C_{0\tau}$ using the iVAMPnet basis functions in each of the subsystems. We note that this approximate transition model is given only for orientation and comparison with the previous model, since the approximated transition operator is generally not a row-stochastic matrix (also compare [2]).

In general, we find that the structural features resolved by our previous study are also resolved by iVAMPnets. In the CBR (cf. Table D.1), we find CBR1 α-helices at two locations and a state burying Met173, as well as a structural rearrangement in the CBR2 that may be described as attachment / detachment of that loop to the protein body. Furthermore, iVAMPnets identify metastable dynamics in the CBR3 loop that was not previously described. In comparison to our previous model, the iVAMPnet model for

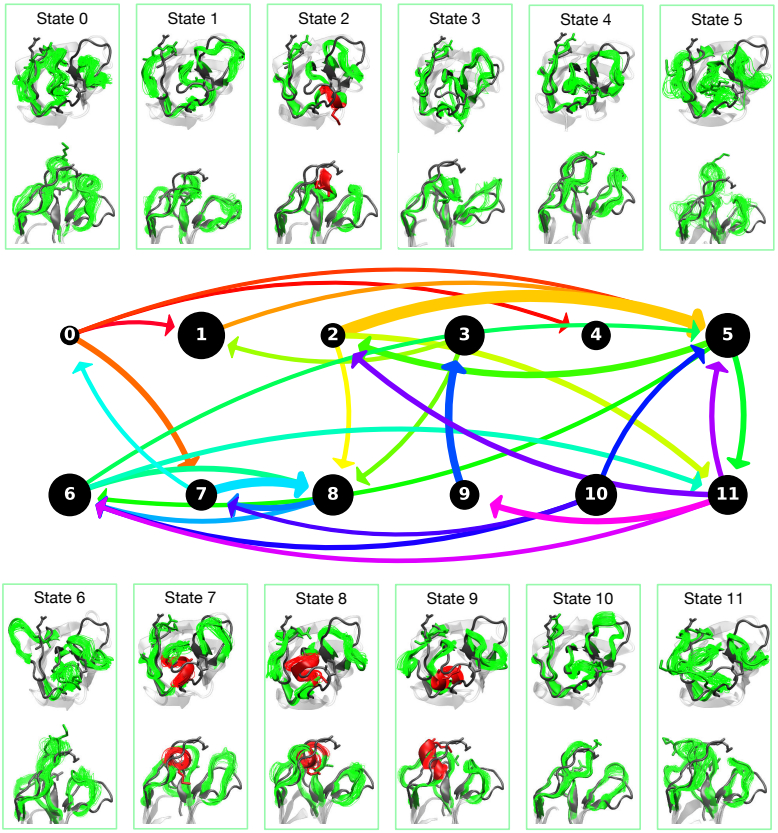


Figure D.3: Metastable states and transition model of the first subsystem of synaptotagmin that is located in the CBR. For each state, structural renders from two perspectives are given. Helical conformations are highlighted in red. The transition model is an approximation; we show only the most important pathways ($T_{ij} \geq 0.0008$) and depict transition probabilities by arrow thickness. Additionally, approximated stationary probabilities are shown as circle diameters for each state. Arrows are colorized to make them distinguishable.

the CBR describes kinetics involving concerted motions of all three CBR loops. We show structural renders of the metastable states in Fig. D.3.

The second subsystem identifyed by iVAMPnets contains the C34 and C78 loops (Fig. D.4): The C78 loop shows structural features as described in [1], i.e., loop rearrangements that are governed by different conformations of two valine residues (Val250, Val255) (Tab. D.2). However, iVAMPnets identify another related site that we call C34, a region that is rich in lysine and has been previously described as important for membrane interactions [3]. Its metastable states are described by different conformations of three lysines (Lys 189-191).

Finally, we compare the probabilities of our previous model [1] with the approximate stationary probabilities obtained here (Fig. D.5) and find that both models agree qualitatively. Differences between stationary probabilities may be due to the previous model's

iVAMPnet state	legacy CBR1	legacy CBR2
0	disordered (D)	in (A)
1	Met-in (B)	out (B)
2	disordered (D)	in (A)
3	disordered (D)	in (A)
4	disordered (D)	in (A)
5	disordered (D)	out (B)
6	disordered (D)	n/a
7	top-helix (A)	in (A)
8	top-helix (A)	out (B)
9	site-helix (C)	out (B)
10	disordered (D)	in (A)
11	site-helix (C)*	out (B)

Table D.1: iVAMPnet states of the first subsystem (CBR, shown in Fig. D.3) and their correspondence to our previous model ("legacy") [1]. *State 11, which is structurally similar to state 9, was assigned to legacy state C to incorporate uncertainties of the metastable state assignment in our previous model.

iVAMPnet state	legacy C78
0	C
1	A
2	C
3	C
4	n/a
5	B

Table D.2: iVAMPnet states of the second subsystem (C34 and C78, shown in Fig. D.4) and their correspondence to our previous model ("legacy") [1].

non-optimal subsystem decomposition and differences in metastable state assignments, in particular regarding the CBR1.

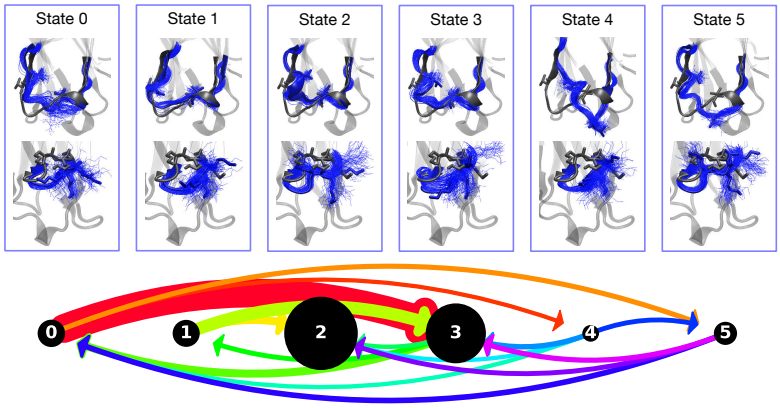


Figure D.4: Metastable states and transition model of the second subsystem of synaptotagmin, which corresponds to the site opposite of the CBR (loops C78 and C34). Each state is depicted by two structural renders, showing details of C78 (top) and C34 (bottom), respectively. The transition model is explained in the caption of Fig. D.3.

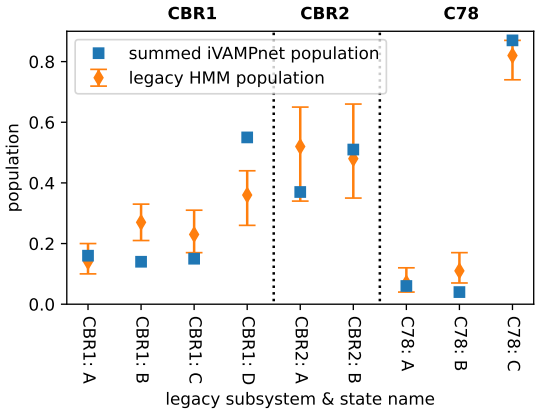


Figure D.5: Comparison of approximated stationary probabilities of iVAMPnet states with estimates of hidden Markov models (HMMs) given in Ref. [1]. iVAMPnet populations are obtained by summing probabilities over all states that are in a certain legacy state (cf. Tabs. D.1,D.2), HMM popluations are obtained from maximum likelihood HMM estimate, the error interval represents HMM model uncertainty (5-95% percentile of Bayesian HMM samples that were obtained using a mixed prior [4, 5], cf. Ref. [1] for details).

D.5 Supplementary Note 5: Counter-example villin: A non-decomposable system

We demonstrate the behavior of iVAMPnets when confronted with a system that cannot be decomposed into subsystems without missing the slowest global processes. Therefore, we employ the method on the villin dataset [6] where several folding and unfolding events are encoded. If studied by a classical VAMPnet, the folding is recovered as the second slowest process. The slowest process describes a transition between a misfolded and the folded state [7]. However, the majority of residues are involved in both of these processes, making it hard to decouple these processes into independent subsystems.

For the analysis, the same hyperparameters are used as for synaptotagmin, but we choose only two states per subsystem. The resolved processes resemble a localized folding of either the left or right helix of the folded structure in each subsystem (Fig. D.6). However, the implied timescales do not converge, expressing non-Markovian behavior. The results can be interpreted as representing a compromise between learning nearly independent processes and approximating slow processes.

Since the independence is strongly enforced, the processes are badly approximated, resulting in unconverged implied timescales. In order to interpret this model, we have furthermore correlated the iVAMPnet eigenfunctions with the ones of a standard global VAMPNet (Fig. D.7). We find that the process found by subsystem 1 has the highest Pearson correlation ($r = -0.79$) with the second global process, which corresponds to peptide folding. However, the second subsystem cannot be clearly assigned to a global process. These results are not surprising since the poor implied timescales convergence and the post-training validation scores (Tab. 1) indicate that the independence approximation does not hold in this example. The system expresses dynamics on the global level that are not or are only poorly approximated by the described iVAMPnets.

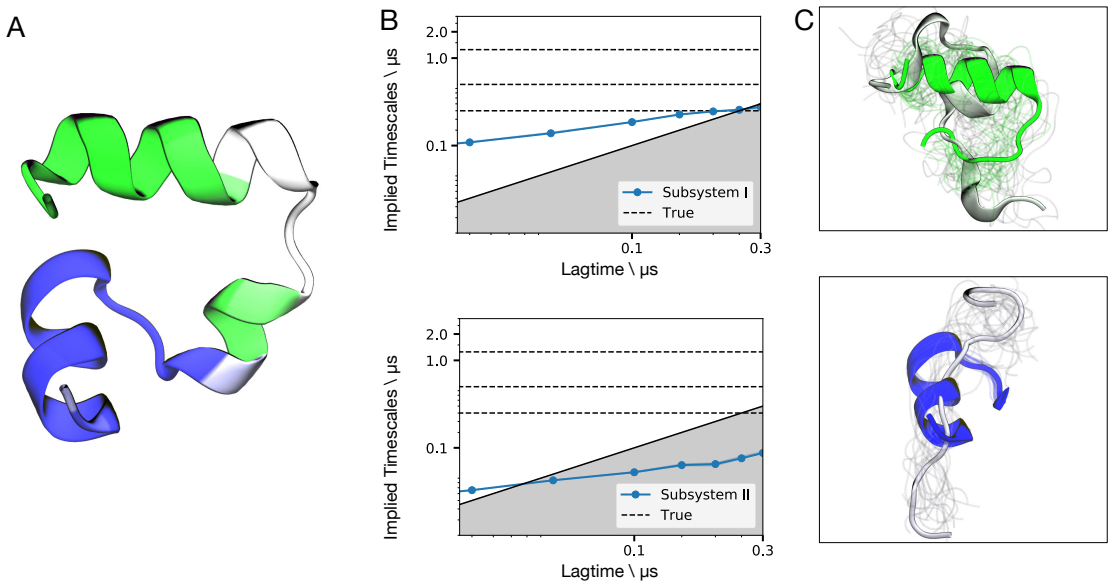


Figure D.6: Counter example to iVAMPNets using villin, trained with independence constraint. **(a)** Subsystem assignment, i.e., masked importance values, are shown as color code on the folded structure. **(b)** Implied timescales of the 2 subsystems, the black dotted lines are reference timescales of a global model trained with a standard VAMPnet. **(c)** 20 representative structures of both extrema of the slowest resolved eigenfunctions for both subsystems. The processes tend to approximate formation of the N- and C-terminal helices, respectively.

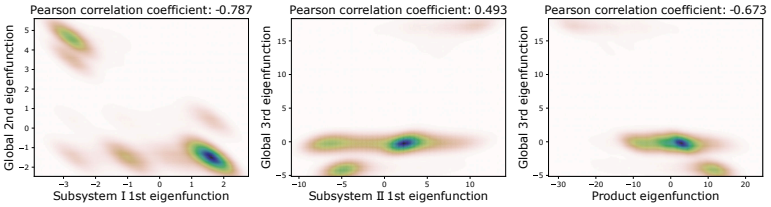


Figure D.7: Interpreting the processes found by the iVAMPnets for villin by correlating them with the global processes found with a standard VAMPnet. Shown are the processes with the highest correlation plotted against each other. The first subsystem correlates with the second global process, which relates to peptide folding. The second subsystem cannot clearly assigned to a global process. However, the product of the two eigenfunctions exhibits significant correlation with the third global eigenfunction.

D.6 Supplementary Note 6: Performance evaluation

In order to evaluate the performance of iVAMPnets, we have modified the 10-cube benchmark system to a variable (even) number of subsystems (N-cube). We generate trajectories of 100,000 data points for each N-cube and train ten instances of both, VAMPnets and iVAMPnets, on it. We recorded the training to be converged when the validation score did not improve by 0.25% compared to the best score so far for 5 training epochs in a row. The results are evaluated first by checking that the global VAMP-E score is indeed converged to its optimum (Fig. D.8a), which is true for all instances of trained iVAMPnets. Regular VAMPnets expectedly fail to scale to larger numbers of subsystems – in this case, they do not converge for 8 subsystems and beyond as that corresponds to a number of states larger or equal to $2^8 = 256$. The performance is now evaluated in terms of elapsed real time for training (Fig. D.8b). We find that expectedly, both methods have increasing time demands for growing numbers of subsystems or states. We note that the elapsed time for VAMPnets with ≥ 8 subsystems could not be evaluated due to the failed training procedure. Furthermore, iVAMPnets slightly outperform VAMPnets, which may be caused by the following features of the benchmark system: a) The N-cube consists of fully independent subsystems, therefore iVAMPnets can find a domain decomposition quickly. b) The number of states per subsystem in iVAMPnets is just 2, i.e., the neural network parameters can be learned from less data (given a domain decomposition) as compared to a VAMPnet that need to be trained on all transitions between 2^N states.

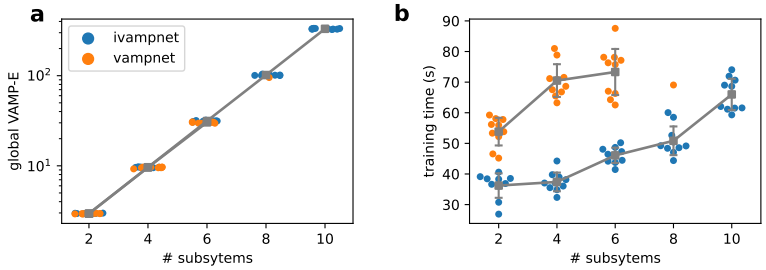


Figure D.8: Performance evaluation of iVAMPnets compared to VAMPnets using an N-cube. **(a)** Global VAMP-score as function of the number of subsystems. **(b)** Training time (in real time) for both methods, as a function of the number of subsystems. Ten instances of VAMPnets and iVAMPnets are trained for each given number of subsystems and drawn as a swarm plot; mean and standard deviations are shown in grey. VAMPnets fail to consistently predict a valid score for 8 subsystems and beyond.

D.7 Supplementary Note 7: MD setups

The used MD data sets were generated with the following properties: The synaptotagmin C2A data set [1] consists of 92 trajectories with a length of 2 µs each, adding up to 184 µs total simulation time. Simulations were conducted with the CHARMM36 force field [8] in explicit solvent and the NPT ensemble at 300K and 1 bar. Trajectories were seeded from a smaller precursive data set which was based on PDB-ID 2R83 [9]. The villin data set [6] consists of a single trajectory of 125 µs length that was started from the unfolded structure. Simulations were performed with the CHARMM22* force field [10] in explicit solvent and the NVT ensemble at 360K.

Bibliography

[1] T. Hempel, N. Plattner, and F. Noé. “Coupling of Conformational Switches in Calcium Sensor Unraveled with Local Markov Models and Transfer Entropy”. *J. Chem. Theory Comput.* 16.4 (2020), pp. 2584–2593.

[2] A. Mardt, L. Pasquali, F. Noé, and H. Wu. “Deep Learning Markov and Koopman Models with Physical Constraints”. *Proc. First Math. Sci. Mach. Learn. Conf.* Ed. by J. Lu and R. Ward. Vol. 107. Proceedings of Machine Learning Research. Princeton University, Princeton, NJ, USA: PMLR, 2020, pp. 451–475.

[3] J. Guillén, C. Ferrer-Orta, M. Buxaderas, D. Pérez-Sánchez, M. Guerrero-Valero, G. Luengo-Gil, J. Pous, P. Guerra, J. C. Gómez-Fernández, N. Verdaguer, and S. Corbalán-García. “Structural Insights into the Ca2+ and PI(4,5)P2 Binding Modes of the C2 Domains of Rabphilin 3A and Synaptotagmin 1”. *Proc. Natl. Acad. Sci.* 110.51 (2013), pp. 20503–20508.

[4] J. D. Chodera, P. Elms, F. Noé, B. Keller, C. M. Kaiser, A. Ewall-Wice, S. Marqusee, C. Bustamante, and N. S. Hinrichs. “Bayesian Hidden Markov Model Analysis of Single-Molecule Force Spectroscopy: Characterizing Kinetics under Measurement Uncertainty”. *ArXiv11081430 Cond-Mat Physicsphysics Q-Bio* (2011). arXiv: 1108.1430 [cond-mat, physics:physics, q-bio].

[5] M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz, and F. Noé. “PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models”. *J. Chem. Theory Comput.* 11.11 (2015), pp. 5525–5542.

[6] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw. “How Fast-Folding Proteins Fold”. *Science* 334.6055 (2011), pp. 517–520.

[7] B. E. Husic and F. Noé. “Deflation Reveals Dynamical Structure in Nondominant Reaction Coordinates”. *J. Chem. Phys.* 151.5 (2019), p. 054103.

[8] J. B. Klauda, R. M. Venable, J. A. Freites, J. W. O’Connor, D. J. Tobias, C. Mondragon-Ramirez, I. Vorobyov, A. D. MacKerell, and R. W. Pastor. “Update of the CHARMM All-Atom Additive Force Field for Lipids: Validation on Six Lipid Types”. *J. Phys. Chem. B* 114.23 (2010), pp. 7830–7843.

[9] K. L. Fuson, M. Montes, J. J. Robert, and R. B. Sutton. “Structure of Human Synaptotagmin 1 C2AB in the Absence of Ca2+ Reveals a Novel Domain Association,” *Biochemistry* 46.45 (2007), pp. 13041–13048.

[10] S. Piana, K. Lindorff-Larsen, and D. E. Shaw. “How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization?” *Biophysical Journal* 100.9 (2011), pp. L47–L49.

Short summary

Computer simulations such as molecular dynamics (MD) provide a possible means to understand protein dynamics and mechanisms on an atomistic scale. The resulting simulation data can be analyzed with Markov state models (MSMs), yielding a quantitative kinetic model that, e.g., encodes state populations and transition rates. However, the larger an investigated system, the more data is required to estimate a valid kinetic model. In this work, we show that this scaling problem can be escaped when decomposing a system into smaller ones, leveraging weak couplings between local domains. Our approach, termed independent Markov decomposition (IMD), is a first-order approximation neglecting couplings, i.e., it represents a decomposition of the underlying global dynamics into a set of independent local ones. We demonstrate that for truly independent systems, IMD can reduce the sampling by three orders of magnitude. IMD is applied to two biomolecular systems. First, synaptotagmin-1 is analyzed, a rapid calcium switch from the neurotransmitter release machinery. Within its C2A domain, local conformational switches are identified and modeled with independent MSMs, shedding light on the mechanism of its calcium-mediated activation. Second, the catalytic site of the serine protease TMPRSS2 is analyzed with a local drug-binding model. Equilibrium populations of different drug-binding modes are derived for three inhibitors, mirroring experimentally determined drug efficiencies. IMD is subsequently extended to an end-to-end deep learning framework called iVAMPnets, which learns a domain decomposition from simulation data and simultaneously models the kinetics in the local domains. We finally classify IMD and iVAMPnets as Markov field models (MFM), which we define as a class of models that describe dynamics by decomposing systems into local domains. Overall, this thesis introduces a local approach to Markov modeling that enables to quantitatively assess the kinetics of large macromolecular complexes, opening up possibilities to tackle current and future computational molecular biology questions.

Kurzfassung

Rechnergestützte Molekulardynamik-Simulationen ermöglichen es, die Dynamik und Mechanismen von Proteinen auf atomarer Ebene nachzuvollziehen. Die resultierenden Simulationsdaten können mithilfe von Markov-Zustandsmodellen (MSMs) analysiert werden, welche die Kinetik beispielsweise mittels Zustandswahrscheinlichkeiten und Übergangsraten quantitativ beschreiben. Der Bedarf an benötigten Daten für MSMs steigt allerdings mit der Größe der untersuchten Systeme stark an. Diese Arbeit zeigt, dass sich das Skalierungsproblem durch die Zerlegung eines Systems in kleinere Teile umgehen lässt. Dafür werden schwache Kopplungen zwischen lokalen Domänen genutzt. Der hier vorgestellte Ansatz, die unabhängige Markov-Zerlegung (IMD), ist eine Näherung erster Ordnung, die Kopplungen vernachlässigt. Die vorliegende globale Dynamik wird in eine Menge von unabhängigen, lokalen Dynamiken zerlegt, welche dann getrennt betrachtet werden. Durch IMD kann die für MSMs erforderliche Datenmenge um drei Größenordnungen reduziert werden. Als Anwendungsbeispiele dienen zwei Proteinsysteme: Den Anfang macht die Analyse von Synaptotagmin-1, einem calciumbindenden Protein in der Neurotransmitter-Exozytose. Innerhalb seiner C2A-Domäne werden lokale Konformationsschalter identifiziert und mit unabhängigen MSMs modelliert. Dies gibt Aufschluss über die calciumabhängige Aktivierung des Proteins. Zweites Anwendungsbeispiel ist ein lokales Modell des katalytischen Zentrums der Serinprotease TMPRSS2, dessen Hemmung mit verschiedenen Medikamenten vergleichend untersucht wird. Für drei Inhibitoren werden Zustandswahrscheinlichkeiten und ein Ratenmodell verschiedener Wirkstoffbindungsmodi geschätzt, welches experimentell ermittelte Wirksamkeiten gut abbildet. Zusätzlich wird IMD zu einer Deep-Learning-Methode erweitert (iVAMPnets). Anhand von Simulationsdaten lernt sie eine Domänenzerlegung und modelliert die lokale Kinetik in diesen Domänen. Abschließend ordnet die Arbeit die entwickelten Methoden als sogenannte Markovsche Feldmodelle in den größeren Kontext verwandter Methoden ein. Zusammenfassend stellt diese Arbeit einen lokalen Ansatz für Markov-Modellierung vor, mit dem große makromolekulare Systeme quantitativ beschrieben werden können. Der Ansatz eröffnet damit neue Möglichkeiten zur Lösung derzeitiger und zukünftiger Probleme der rechnergestützten Molekularbiologie.

List of publications

The following first author manuscripts are re-printed here as part of this cumulative thesis:

1. Tim Hempel, Mauricio J. del Razo, Christopher T. Lee, Bryn C. Taylor, Rommie E. Amaro, and Frank Noé. "Independent Markov Decomposition: Toward Modeling Kinetics of Biomolecular Complexes". *Proceedings of the National Academy of Sciences* 118.31 (2021), e2105230118.

2. Tim Hempel, Nuria Plattner, and Frank Noé. "Coupling of Conformational Switches in Calcium Sensor Unraveled with Local Markov Models and Transfer Entropy". *Journal of Chemical Theory and Computation* 16.4 (2020), pp. 2584–2593.

3. Tim Hempel, Lluís Raich, Simon Olsson, Nurit P. Azouz, Andrea M. Klingler, Markus Hoffmann, Stefan Pöhlmann, Marc E. Rothenberg, and Frank Noé. "Molecular Mechanism of Inhibiting the SARS-CoV-2 Cell Entry Facilitator TMPRSS2 with Camostat and Nafamostat". *Chemical Science* (2021), 10.1039.D0SC05064D.

4. Andreas Mardt*, Tim Hempel*, Cecilia Clementi, and Frank Noé. "Deep Learning to Decompose Macromolecules into Independent Markovian Domains". *Nature Communications* 13.1 (2022), p. 7101.
   * contributed equally

5. Tim Hempel, Simon Olsson, and Frank Noé. "Markov Field Models: Scaling Molecular Kinetics Approaches to Large Molecular Machines". *Current Opinion in Structural Biology* 77 (2022), p. 102458.

Furthermore, the following publications and patent application were co-authored in the course of this doctorate:

6. Christoph Wehmeyer*, Martin K. Scherer*, Tim Hempel*, Brooke E. Husic, Simon Olsson, and Frank Noé. "Introduction to Markov State Modeling with the PyEMMA Software [Article v1.0]". *Living Journal of Computational Molecular*

Science 1.1 (2019), 5965.

* contributed equally

Parts of this manuscript have been re-used in Chapter 1.

7. M. Hoffmann, H. Hofmann-Winkler, J. C. Smith, N. Krüger, P. Arora, L. K. Sørensen, O. S. Søgaard, J. B. Hasselstrøm, M. Winkler, T. Hempel, L. Raich, S. Olsson, O. Danov, D. Jonigk, T. Yamazoe, K. Yamatsuta, H. Mizuno, S. Ludwig, F. Noé, M. Kjolby, A. Braun, J. M. Sheltzer, and S. Pöhlmann. "Camostat Mesylate Inhibits SARS-CoV-2 Activation by TMPRSS2-Related Proteases and Its Metabolite GBPA Exerts Antiviral Activity". *EBioMedicine* (2021), p. 103255.

8. Tim Hempel*, Katarina Elez*, Nadine Krüger*, Lluís Raich, Jonathan H. Shrimp, Olga Danov, Danny Jonigk, Armin Braun, Min Shen, Matthew D. Hall, Stefan Pöhlmann, Markus Hoffmann, and Frank Noé. "Synergistic Inhibition of SARS-CoV-2 Cell Entry by Otamixaban and Covalent Protease Inhibitors: Pre-Clinical Assessment of Pharmacological and Molecular Properties". *Chemical Science* (2021), 10.1039.D1SC01494C.

   * contributed equally

9. Tim Hempel, Katarina Elez, Lluís Raich, Frank Noé, Nadine Krüger, Markus Hoffmann, and Stefan Pöhlman. "Pharmaceutical Composition for Treating COVID-19". International patent application PCT/EP2022/ 069448 (EP21187449), filed July 12, 2022.

10. Moritz Hoffmann, Martin Scherer, Tim Hempel, Andreas Mardt, Brian de Silva, Brooke E Husic, Stefan Klus, Hao Wu, Nathan Kutz, Steven L Brunton, and Frank Noé. "Deeptime: A Python Library for Machine Learning Dynamical Models from Time Series Data". *Machine Learning: Science and Technology* 3.1 (2022), p. 015009.

11. David N. Bernard, Chitra Narayanan, Tim Hempel, Khushboo Bafna, Purva Prashant Bhojane, Myriam Létourneau, Elizabeth E. Howell, Pratul K. Agarwal, and Nicolas Doucet. "Conformational Exchange Divergence along the Evolutionary Pathway of Eosinophil-Associated Ribonucleases". *Structure* (2023), S096921262200497X.

12. Gabor Tóth, Tim Hempel, Krishna Somandepalli, and Shri Narayanan. "Studying Large-Scale Behavioral Differences in Auschwitz-Birkenau with Simulation of Gendered Narratives". *Digital Humanities Quarterly* 16.3 (2022).

Acknowledgments

I thank Prof. Frank Noé for advising my scientific work, for the many fruitful discussions, ample advice, and the many opportunities for developing fresh ideas. It has been a pleasure to work with him and to be part of his group. Furthermore, I thank Prof. Roland Netz for his inspiring lectures and being my second advisor and Prof. Cecilia Clementi for her continuous support and sharing her great physics intuition. I am grateful to Nuria Plattner for introducing me to the field of computational molecular biology and for her thoughtful support during my first years as a PhD student. Furthermore, I thank Dr. Murthy Gudipati (Caltech / NASA JPL) for introducing me to scientific working and for being a role model in general.

What would science be without the many genius individuals that form a research group? Special thanks to Moritz for his generous support with both, mathematics and efficient, bug-free code implementations; to Martin, without whom I would still be waiting for softwares to compile; to Lluís for opening my eyes to biochemistry and protonation states; to Selle for mandatory upgrades to my English writing; to Mohsen for explaining entropy to me over and over again; to Andreas M. for approximating singular functions like no other; and to Katarina for her capabilities as the CPU in a wildly complex project.

Finally, all this work would not have been possible without the fruitful atmosphere that was created by all the other, previous and current, members of the Noé group (in alphabetical order): Aleksander, Andreas K., Atharva, Ben T.-S., Bernat, Brooke, Chris F., Chris W. Esam, Fabian, Fabio, Feliks, Giovanni, Guillermo, Jan, JHP, Jonas, Kasia, Leon, Luca, Luigi, Maaike, Manuel, Mauricio, Michele, Mohamad, Paolo, Robin, Sebastian S., Shreyas, Simon, Terra, Tuan, Yaoyi, and Zeno. I also thank the members of the Clementi group that I was happy to join during my last years (in alphabetical order): Andrea, Clark, David, Félix, Iryna, Klara, Nick, and Wangfei. Additionally, thank you very much Birgit, Carmen, and Katja for your great support with all possible and impossible tasks posed by and to the bureaucratic machinery.

Furthermore, I thank my friends Xander and Michael for the great discussions ranging from science to absurdity; Tom and Manuel for their friendships and the most exciting (sometimes: most lazy) sports experiences; and Lisa for being my heroine in joyful and dire times alike. I thank Sophie for her continued emotional support throughout my last years of high-school, undergrad, and PhD, for listening to enigmatic physics monologues and for making sure that I'll never make myself too comfortable sitting in our yellow armchair (at least not for too long). Ein großer Dank geht an meine Mutter, die trotz gähnender Misserfolge beim Üben des "Kleinen Einmaleins" immer an mich geglaubt, meiner Bildung höchste Priorität eingeräumt und mir ihr Vertrauen geschenkt hat. It's been a pleasure to be surrounded by all of you!

Eigenständigkeitserklärung

Name: Hempel
Vorname: Tim

Ich erkläre gegenüber der Freien Universität Berlin, dass ich die vorliegende Dissertation selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt habe. Die vorliegende Arbeit ist frei von Plagiaten. Alle Ausführungen, die wörtlich oder inhaltlich aus anderen Schriften entnommen sind, habe ich als solche kenntlich gemacht. Diese Dissertation wurde in gleicher oder ähnlicher Form noch in keinem früheren Promotionsverfahren eingereicht.

Mit einer Prüfung meiner Arbeit durch ein Plagiatsprüfungsprogramm erkläre ich mich einverstanden.

Datum:

Unterschrift: