Independent Markov Decomposition: Towards modeling kinetics of biomolecular complexes

Tim Hempel\textsuperscript{a,b}, Mauricio J. del Razo\textsuperscript{a,c,d,1}, Christopher T. Lee\textsuperscript{e,1}, Bryn C. Taylor\textsuperscript{f,†}, Rommie E. Amaro\textsuperscript{a,b,1}, and Frank Noé\textsuperscript{a,b,1}

\textsuperscript{a}Freie Universität Berlin, Department of Mathematics and Computer Science, Berlin, Germany; \textsuperscript{b}Freie Universität Berlin, Department of Physics, Berlin, Germany; \textsuperscript{c}University of Amsterdam, Van’t Hoff Institute for Molecular Sciences and Korteweg-de Vries Institute for Mathematics, Amsterdam, The Netherlands; \textsuperscript{d}Dutch Institute for Emergent Phenomena, Amsterdam, The Netherlands; \textsuperscript{e}University of California, San Diego, Department of Mechanical and Aerospace Engineering, San Diego, CA, USA; \textsuperscript{f}University of California, San Diego, Biomedical Sciences Graduate Program, San Diego, CA, USA; \textsuperscript{g}Rice University, Department of Chemistry, Houston, TX, USA; \textsuperscript{1}Contributed equally to this work.

In order to advance the mission of \textit{in silico} cell biology, modeling the interactions of large and complex biological systems becomes increasingly relevant. The combination of molecular dynamics (MD) and Markov state models (MSMs) have 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 increase 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 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.

Conflict of interest statement: The authors declare no conflict of interest.

1MJR, CTL, and BCT contributed equally to this work.

To whom correspondence should be addressed. E-mail: ramaro@ucsd.edu, frank.noé@fuberlin.de

This manuscript was compiled on March 24, 2021.
We first describe IMD for discrete-state MSMs before generalizing it to time series with continuous descriptors.

**Markov State models.** An MSM consists of a discretization of molecular state space into a disjoint set of states \( \{S_1, ..., S_n\} \) 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 \( i \)th state or not by using indicator functions:

\[
\chi_i(x) = \begin{cases} 
1 & x \in S_i \\
0 & \text{otherwise.}
\end{cases}
\]

The vector \( \chi = [\chi_1, ..., \chi_n]^T \) is thus a “one-hot encoding” that maps the continuous state \( x \) to the MSM discretization. For this or any other choice of features \( \chi \) we can compute the instantaneous and time-lagged correlation matrices \( C_{00} = \sum_i \chi(x_i)\chi(x_i)^T \) and \( C_{0r} = \sum_i \chi(x_i)\chi(x_{i+r})^T \), 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:

\[
\mathbf{P}(\tau) = C_{00}^{-1}C_{0r}.
\]

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 we always start in state 0 at time 0. This vector can be evolved in time using the transition matrix, until it converges to the equilibrium distribution \( \pi = \lim_{t \to \infty} \mathbf{p}_t \):

\[
\mathbf{p}_{t+r}^\top = \mathbf{p}_t^\top \mathbf{P}(\tau).
\]

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:

\[
R_n(P) = \left\| C_{00}^{-1/2}C_{0r}C_{rr}^{-1/2} \right\|_n^n
\]

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, and we can employ correlation matrix estimators that provide \( C_{00} = C_{rr} \) and symmetric \( C_{0r} \) (detailed balance). In this special case, VAMP becomes the VAC mentioned above, and the variational score simply becomes \( R_n(P) = \| \mathbf{P}(\tau) \|_n^n \). 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 transitions in structure. For instance, the proposed validation score works even with limited data \((41)\). VAMP finds that a Markovian model that best approximates the high-dimensional continuous dynamics maximizes the VAMP-n score:

\[
R_n(P) = \left\| C_{00}^{-1/2}C_{0r}C_{rr}^{-1/2} \right\|_n^n
\]
processes of the molecular process under investigation (40, 43). For a fixed state space discretization, optimizing the VAC results in the MSM estimator (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 \( \chi \) with deep neural networks, leading us to VAMPnets (45).

**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. 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” (\( \alpha, \beta = 0^\circ \)) or “open” (\( \alpha, \beta = 90^\circ \)). 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 \( p_{A,t+i} = p_{A,t} \) and \( p_{B,t+i} = p_{B,t} \) (Fig. 2b). As the MSMs \( A \) and \( B \) are statistically independent, the probability distribution of the entire system follows Eq. (3) with

\[
P_1 = p_{A,t} \otimes p_{B,t}
\]

\[
P(\tau) = p_{A}(\tau) \otimes p_{B}(\tau),
\]

where \( \otimes \) is the Kronecker product (46) (see SI Appendix, Markov operators). The vector \( p_1 \) 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 \( P(\tau) \) is the 4 \( \times \) 4 transition matrix between these combinatorial states whose transition probabilities are simply products of the individual transition events in subsystems \( A \) and \( B \) (Fig. 2c, d).

The power of this approach is apparent when comparing Figures 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 exponentially 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 \( m^{2N} \) transitions, while a decomposition into independent MSMs reduces this to a polynomial number of \( N m^2 \) transitions which can be scaled to large systems.

The above example trivially generalizes to \( N \) systems with \( P(\tau) = \bigotimes^N p_i(\tau) \). We note that it is customary to dismiss variables of the full state space \( \Gamma \) (Fig. 2a) that are assumed to average quickly, e.g., solvent degrees of freedom. Thus the modeled space \( \Omega \) in practice only encompasses the variables of interest, e.g., internal coordinates of a protein system.

**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 which quantifies decomposition validity.

To this end we come back to the variational approach Eq. (4). Conveniently, matrix norms follow simple rules when applied to a Kronecker product (SI Appendix, Markov operators). 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 molecu-

---

**Fig. 2.** Operator decomposition and discretization on a test molecule. a) A test molecule is decomposed into two subsystems (blue and red). The two angles \( \alpha \) and \( \beta \) span subspaces \( A \) and \( B \) corresponding to the two subdomains, respectively. The space \( \Gamma \) is composed of all system degrees of freedom. The space \( \Omega \) 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 \( \Omega \) are given as the tensor product \( P_A \otimes P_B \). b) The molecule has metastable states at \( \alpha, \beta = 0^\circ \) as shown in Fig. 2a, where \( A \) and \( B \) are 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 which quantifies decomposition validity.

---

**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. 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” (\( \alpha, \beta = 0^\circ \)) or “open” (\( \alpha, \beta = 90^\circ \)). 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 \( p_{A,t+i} = p_{A,t} \) and \( p_{B,t+i} = p_{B,t} \) (Fig. 2b). As the MSMs \( A \) and \( B \) are statistically independent, the probability distribution of the entire system follows Eq. (3) with

\[
P_1 = p_{A,t} \otimes p_{B,t}
\]

\[
P(\tau) = p_{A}(\tau) \otimes p_{B}(\tau),
\]

where \( \otimes \) is the Kronecker product (46) (see SI Appendix, Markov operators). The vector \( p_1 \) 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 \( P(\tau) \) is the 4 \( \times \) 4 transition matrix between these combinatorial states whose transition probabilities are simply products of the individual transition events in subsystems \( A \) and \( B \) (Fig. 2c, d).

The power of this approach is apparent when comparing Figures 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 exponentially 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 \( m^{2N} \) transitions, while a decomposition into independent MSMs reduces this to a polynomial number of \( N m^2 \) transitions which can be scaled to large systems.

The above example trivially generalizes to \( N \) systems with \( P(\tau) = \bigotimes^N p_i(\tau) \). We note that it is customary to dismiss variables of the full state space \( \Gamma \) (Fig. 2a) that are assumed to average quickly, e.g., solvent degrees of freedom. Thus the modeled space \( \Omega \) in practice only encompasses the variables of interest, e.g., internal coordinates of a protein system.

**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 which quantifies decomposition validity.

To this end we come back to the variational approach Eq. (4). Conveniently, matrix norms follow simple rules when applied to a Kronecker product (SI Appendix, Markov operators). 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 molecu-
lar 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):

$$R_n(P) = \prod_{i=1}^N R_n(P_i). \quad [6]$$

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. Note that Eq. (6) is a necessary but not a sufficient condition for Markov independence. Significant deviations from equality in Eq. 6 indicate that the assumption of independence is invalid. However, in particular if separate MSMs $P_i$ can probe the same molecular features it is possible to satisfy Eq. (6) even though the subsystem MSMs are not statistically independent. Eq. (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. (6).

In practice, we want to perform an IMD because often we cannot compute the global MSM $P$ due to limited sampling (Fig. 1), and we consequently do not know $R_n(P)$. Therefore, we choose to only check the equality of Eq. (6) on pairs of subsystems $A, B$, i.e., $R_n(P_{A,B}) = R_n(P_A) \cdot R_n(P_B)$. We then search over possible partitions of the molecular system into subsystems by evaluating the graph of pairwise dependencies $d(A, B)$:

$$d(A, B) = |R_n(P_{A,B}) - R_n(P_A) \cdot R_n(P_B)| \quad [7]$$

In practice, computing $P_{A,B}$ 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. (6) and (7) is purely a result of the chosen basis set of MSMs (Eq. (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 carrying out an MSM discretization and estimation for each subsystem. When considering more general features $\chi$, 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 easiest described by “stacking” observable feature vectors rather than by defining an MSM discretization on the combinatorial space. 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 $\Psi_{A,B} = (\psi^A_1, \psi^B_1, \psi^A_2, \psi^B_2, \ldots)$. The transfer operator that describes 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. Please 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

$$d(A, B) = |R_n(K_A) + R_n(K_B) - R_n(K_{A,B})| \quad [8]$$

where $K$, the Koopman operator, takes the place of the transition matrix $P$. See SI Appendix, VAMP score decomposition for the derivation. Please also note that working in observable space entirely rules out discretization artifacts.

Results

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 which 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 both experimental measurements of ion channel properties sets the stage for computational simulations to provide molecular details and mechanistic insights (53). While it is possible to fit a phenomenological MSM using data from electrophysiological experiments, atomistic modeling remains out of reach due to the long timescale 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, we exploit the independence of individual subunits or gates and partition accordingly (Fig. 3a1). This produces four matrices $P_i \in \mathbb{R}^{2 \times 2}$, $1 \leq i \leq 4$ that describe individual gate opening and closing. As derived above, the Kronecker product of subsystem transition matrices yields a transition matrix $P \in \mathbb{R}^{16 \times 16}$ of the full ion channel (Fig. 3a2). The 16 states enumerate all possible combinations of open and closed gates of the full ion channel, a state space referred to as $S$ 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. 3b). In Ref. (54), channel symmetry is used to define 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
Fig. 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 (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.

full system configuration \( s \in \mathcal{S} \) 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) \( P_{\text{empirical}} = \Pi^{-1} \chi^T \Pi \), \( \chi \in \mathbb{R}^{5 \times 5} \) with \( \Pi = \text{diag}(\pi) \) the diagonal matrix of the stationary distribution \( \pi \) in full space and in empirical space \( \Pi_{\text{c}} = \text{diag}(\chi^T \pi) \).

Choosing parameters \( \alpha \) and \( \beta \) 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, empirical 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. 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).

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 IMD sampling 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. 3d we present the percentage of connected IMD estimated on an ensemble of ten realizations of the Markov chain and compare to a classical MSM. It 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. The higher computational efficiency of IMD is evident from the much faster convergence of implied timescales as a function of simulation length (Fig. 3d), showing a reduction in sampling by three orders of magnitude, from tens of seconds to tens of milliseconds (Fig. 3d). For example, ionic conductance is reasonably approximated with 100 ms of sampling and the IMD approach (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. S1, modeling of weakly coupled systems is possible in a robust fashion.

Finding Independent Markov Partition 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 development of viable partitions which produce independent subsystems.

In this section we demonstrate how the dependency defined in Sec. 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. 4a). In our setup, this patch contains a dimer of channels which 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. 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. 4d). Further, reducing the coupling strength within a channel does not alter our qualitative results (Fig. 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.

Finding Independent Markov Partition 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. 5). This partition contains the conformational switches defined in our last study (15): In particular, the C78 switch (Fig. 5, blue) emerges as an independent cluster in the Fruchterman-Reingold projection, confirming our previous results. However, even though two conformational switches in the Calcium Binding Region (CBR), CBR-1 and 2 together (Fig. 5, green and red), have low dependency to the other protein residues, 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. S2).

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. This development comes with an increasing number of (metastable) states that demands vast amounts of sampling time and hampers our attempts to rigorously model protein dynamics. We believe that the classical MSM method is
reaching 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 ones. 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 decomposed MSMS 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 have shown 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 which 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 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 estimation without knowing the dependency graph structure a priori.

In order to optimally partition a system in practical applications, 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 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). For future work, in particular for larger biomolecular complexes, it will however be desirable to incorporate experimental knowledge about size and properties of “protein sectors” (32).

An aspect excluded in this conceptual study is the discretiza-
in-silico. We anticipate that IMD will have a broad application basis for our interest in modeling large biological systems is by using Hempel established methods and software of the MD MSM community, toy potassium ion channel were obtained from the Hodgkin-Huxley because this approach comes with all the computational cost. Because this approach comes with all the data efficient and might be easier to apply than classical global applications (4, 62). 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 may reduce discretization artifacts compared to classical MSMs. However, implications of the discretization error should be discussed when applying the dependency score as it is unclear how the error propagates to joint space probability estimates.

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 is 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 biology.

Materials and Methods

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,

\[ I_K = G_K(V_m - V_K) = \frac{g_K n^n (V_m - V_K)}{1 + \frac{g_K n^n (V_m - V_K)}{1 + \frac{g_K n^n (V_m - V_K)}{1 + \frac{g_K n^n (V_m - V_K)}}}} \]

where \( I_K \) is the current, \( G_K \) is the conductance, \( g_K \) 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 ODE,

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

where \( \alpha_n \) and \( \beta_n \) are the kinetic rates for 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 \( \alpha_n \) and \( \beta_n \),

\[ \alpha_n(V_m) = \frac{0.01(10 - V_m)}{\exp\left(\frac{10 - V_m}{10}\right) - 1}, \]

\[ \beta_n(V_m) = 0.125 \exp\left(-\frac{V_m - 80}{80}\right). \]

The term \( n^4 \) is the joint probability that the four independent subunits of the tetrameric potassium ion channel are concomitantly open. Thus \( \alpha_n \) and \( \beta_n \) 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 (63) to serve as the ground truth for later comparison with IMD 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, Secs. Toy systems, Dimer model.

Generally, a transition matrix describing a (full) test system (possibly including couplings) is chosen, akin to \( P(\tau) \) in Eq. 5. Time series are generated using the Markov chain sampler implemented in pyEMMA/msmtools (64). Subsequently, full system states are mapped to individual subsystem states, yielding subsystem trajectories which are parallel in time. Estimation of subsystem transition matrices \( P_i(\tau) \) in Eq. 5 is followed by assembly of a full system transition matrix. The latter is utilized to extract full system observables such as implied timescales.

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 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.,

\[ d = \frac{R_n(1) + R_n(2) - R_n(1, 2)}{\min(R_n(1), R_n(2))} \in [0, 1] \]

with \( R_n(x) \) being the VAMP-n-score of residue \( x \). The VAMP-n-score is computed with PyEMMA (64).

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.

Data availability. The code that implements our discrete models and reproduces the presented results can be found in our GitHub repository. The molecular dynamics data set of Synaptotagmin C2A is available upon request.

ACKNOWLEDGMENTS. TH thanks Moritz Hoffmann and Andreas Mardt (FU Berlin) for fruitful discussions. We acknowledge

Fig. 5. Dependency-network between residues of Syt-1 C2A depicted using a standard graph layout (Fruchterman-Rheingold 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 = 8). 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. S2).
funding from Deutsche Forschungsgemeinschaft (SFB/TRR 186, Project A12; SFB1114, Project A04), the Berlin Mathematics center MATHE+ (BIMoS), the Bundesministerium für Bildung und Forschung BMBF (BIFOLD), and the European Commission (ERC CoG 772230 "ScaleCell"). BCT and REA acknowledge support from a Hartwell Foundation Postdoctoral Fellowship. REA acknowledges funding from NIH Grant R01 GM132826. CTL acknowledges support from a Hartwell Foundation Postdoctoral Fellowship. REA acknowledges funding from NIH Grant R01 GM132826.


