# Building insightful, memory-enriched models to capture long-time biochemical processes from short-time simulations

**Anthony J. Dominic III[a], Thomas Sayer[a], Siqin Cao[b], Thomas E. Markland[c], Xuhui Huang[b], and Andrés Montoya-Castillo[a,1]**

[a]Department of Chemistry, University of Colorado, Boulder, CO 80309, USA; [b]Department of Chemistry, University of Wisconsin-Madison, Madison, WI 53706, USA; [c]Department of Chemistry, Stanford University, Stanford, California, 94305, USA

This manuscript was compiled on April 28, 2023

**The ability to predict and understand the complex molecular motions occurring over diverse timescales ranging from picoseconds to seconds and even hours occurring in biological systems remains one of the largest challenges to chemical theory. Markov State Models (MSMs), which provide a memoryless description of the transitions between different states of a biochemical system, have provided numerous important physically transparent insights into biological function. However, constructing these models often necessitates performing extremely long molecular simulations to converge the rates. Here we show that by incorporating memory via the time-convolutionless generalized master equation (TCL-GME) one can build a theoretically transparent and physically intuitive memory-enriched model of biochemical processes with up to a three order of magnitude reduction in the simulation data required while also providing a higher temporal resolution. We derive the conditions under which the TCL-GME provides a more efficient means to capture slow dynamics than MSMs and rigorously prove when the two provide equally valid and efficient descriptions of the slow configurational dynamics. We further introduce a simple averaging procedure that enables our TCL-GME approach to quickly converge and accurately predict long-time dynamics even when parameterized with noisy reference data arising from short trajectories. We illustrate the advantages of the TCL-GME using alanine dipeptide, the human argonaute complex, and FiP35 WW domain.**

Markov State Models | Master equation | Biomolecular dynamics

Biomolecules, such as proteins, dynamically change conformations to perform their functions and thus play a critical role in processes such as protein misfolding and aggregation and protein-ligand recognition. Therefore, investigating biomolecular dynamics is essential for discovering next generation therapeutics, developing novel antibiotic targets, and elucidating protein folding mechanisms that underlie diseases such as Alzheimer's, Parkinson's, and Cystic Fibrosis (1). Indeed, all-atom molecular dynamics (MD) computer simulations can offer insight at resolutions beyond standard experimental setups. However, since small atomic motions such as vibrations occur on the order of femtoseconds, whereas the complex motions at the heart of large conformational changes that drive processes such as protein folding and allostery span timescales from microseconds to seconds, a direct atomistic simulation of such long-timescale motions is only feasible for relatively small biological systems.

Markov state models (MSMs) are a powerful approach that have emerged to tackle this grand challenge (2–12). Currently, widely used open-source libraries offer robust implementations for constructing MSMs (13–15). MSMs benefit from massive parallelism by exploiting many short molecular dynamics simulations to capture the long-time configurational dynamics that reveal the mechanisms of biomolecular processes (16). This is accomplished by partitioning configuration space into a set of states: distinct structures whose component configurations interconvert on a faster timescale than with those belonging to different structures. Identifying the slowest interconverting structures, however, remains a formidable problem (17–25). This difficulty arises from the fact that, to perform a perfect partitioning, one needs detailed knowledge of the full free energy landscape of a complex condensed phase system. Instead, one is generally limited to a set of states that evolve on slow timescales but are not optimally partitioned (16, 26). With such a set of configurations, an MSM then provides a discrete-time kinetic description of the interstate conversion, enforcing an effective separation of timescales by requiring transitions between states have no dependence on the history of the system. In this memory-less, or Markovian, limit the rate constants in the kinetic scheme are time-independent. This kinetic description provides an *approximation* to the true dynamics and its accuracy depends on the extent of timescale separation. For a sufficiently accurate ('valid') MSM, the maximum resolution in time (minimum time step) allowed by the approximate description is termed the 'Markovian lag time'. Formally, the intrastate relaxation establishes a *lower bound* to the lag time (16), which is the minimum simulation time required for MD data to parameterize the model.

Ultimately, what one would want is a handful of states

## Significance Statement

Developing a mechanistic understanding of complex biomolecular processes occurring over long timescales presents a formidable challenge. While state-of-the-art techniques like Markov State models are a vital tool in decoding these processes they require a substantial amount of simulation data to construct an accurate model. Here we introduce an approach that goes beyond previous Markovian (memoryless) theories which dramatically reduces the amount of simulation data required to construct a simple and interpretable model of biomolecular processes based on physically transparent time-dependent rates. By deriving a rigorous bound for the simulation times required to construct non-Markovian models of these processes we show that such models provide a much more data efficient approach to understand the dynamics of complex biomolecular systems.

[1]To whom correspondence should be addressed. E-mail: Andres.MontoyaCastillocolorado.edu

that provide chemical interpretability for understanding complex biomolecular mechanisms. However, algorithms designed to maximize this timescale separation usually produce many, physically obscure states. This is because downfolding to a biologically intuitive space subsumes slower interstate dynamics of the many-state space into the intrastate dynamics of the reduced space (27), increasing the lag time. For example, to model the millisecond folding of the NTL9 peptide using the available simulation data, Pande and coworkers required an MSM containing 2,000 states (with a lag time of 12 ns) (28), while recent work on the RNA Polymerase (RNAP) II backtracking necessitated MSMs consisting of 800 states to reach Markovianity within the affordable trajectory (29). Therefore there is a balance to be drawn: one wants to coarse-grain aggressively to facilitate interpretability, yet this generally leads to long lag times, which result in both poor temporal resolution and the need to perform longer MD simulations.

Recent work has demonstrated that one can employ non-Markovian theories to resolve the tensions at the heart of the MSM, increasing the resolution to be equal to the MD time step (30–34), while simultaneously using only a fraction of the data in the models' construction (35). Of these, the GME, recently used in its time-convolution form as a quasi-MSM (qMSM), provides a particularly useful tool. Indeed, qMSMs have proven useful in tackling important problems such as the gate opening motion of a bacterial RNAP (36) and the mechanism of messenger RNA recognition and inhibition via the RNA-induced silencing complex (37). Like MSMs, GMEs are most efficient when there is a separation of timescales between intra- and interstate dynamics. Unlike MSMs, GMEs encode the intrastate dynamics into a time-dependent friction function—a memory kernel—removing the approximation of perfect timescale separation. It is this explicit description of the non-Markovianity that allows the improved resolution in time. Yet, the time-nonlocal GME formulation precludes simple interpretation of the dynamics in terms of 'states and rates', which are typically used to describe the mechanisms of biological processes. This motivates the question: is it possible to combine the interpretability of the MSM with the improved accuracy, resolution, and efficiency of GMEs?

In this work, we employ a time-convolutionless (TCL) GME approach that, like the qMSM, encodes the non-Markovian dynamics associated with intrastate motions but, unlike the qMSM, conserves the chemically intuitive nature of MSMs through the action of a generalized non-Markovian rate matrix. We show this easy, accurate, and efficient GME-based approach can capture the biomolecular dynamics of systems of varying complexity, with the resulting dynamics constituting an improvement that combines the advantages, while removing the limitations, of both qMSM and MSM approaches. Indeed, not only does the TCL-GME approach perform just as well as the qMSM on systems that can be exhaustively sampled, but in more difficult cases where all methods struggle to treat statistically underconverged MD data, the TCL-GME can be systematically improved in a manner that has no apparent analogue in the qMSM (or MSM) case. We achieve this through a simple averaging protocol that leverages the onset of Markovian behavior to tame the deleterious effect of noise. Upon reformulating the TCL-GME in discrete-time (38), our averaging procedure provides a simple and robust scheme to capture the complex dynamics of biomolecular motions, even in cases that suffer from poor temporal resolution. Finally, in the extreme case where our averaging procedure includes the entire non-Markovian region, our TCL-GME reduces to a high-resolution version of the analogous MSM, recapitulating its identity as the non-Markovian generalization of the conventional MSM and fully elucidating the source of improvement over the traditional time-local approach. We demonstrate that our discrete-time method remains robust even when benchmarked against MD data that extends into the microsecond regime: two orders of magnitude longer than the time required to parameterize the model in question. The strict improvement of our time-local approach is epitomized by its ability to converge an computational sensitive experimental observable (the folding time) using less than half of the data required by the traditional MSM.

## Connecting Markovian and non-Markovian Evolution

Whether one wants to directly use a long MD trajectory or many short MD simulations to elucidate complex biomolecular processes, the first task is to find the states that will provide one with the basis of a mechanistic interpretation. The second task is to construct an accurate and efficient description of the dynamics of such configurations. As we mentioned in the Introduction, below we do not consider how one identifies these configurational basins (the interested reader can see, for instance, Refs. (18–21, 23–25, 39)), but rather focus on the second problem: given a set of configurations whose dynamics one can only afford for only short times, how does one construct a dynamical framework to accurately and efficiently capture the dynamics of these configurations over all time?

To characterize the time-dependent transitions connecting states, it is natural to focus on their equilibrium time-correlation functions,

$$\mathcal{C}_{k,j}(t) = \pi_j^{-1} \int \mathrm{d}\boldsymbol{p}_0 \int \mathrm{d}\boldsymbol{q}_0 \; f_{\mathrm{eq}}(\boldsymbol{q}_0, \boldsymbol{p}_0) \chi_k(\boldsymbol{q}_t) \chi_j(\boldsymbol{q}_0), \qquad [1]$$

where $f_{\mathrm{eq}}(\boldsymbol{q}_0, \boldsymbol{p}_0) = f_{\mathrm{eq}}(\boldsymbol{q}_t, \boldsymbol{p}_t) = e^{-\beta \mathcal{H}(\boldsymbol{q},\boldsymbol{p})}/Z$ is invariant to time evolution, the MD Hamiltonian $\mathcal{H}$ is dependent on the coordinates ($\boldsymbol{q}_t = e^{-i\mathcal{L}t}\boldsymbol{q}_0$) and momenta ($\boldsymbol{p}_t = e^{-i\mathcal{L}t}\boldsymbol{p}_0$) of all atoms in the system at time $t$, and $\mathcal{L} = i\{\mathcal{H}, ...\}_{\mathrm{PB}}$ is the Poisson bracket that generates the evolution of the system. Here, $\{\chi_k\}$ are mutually orthogonal indicator functions that define the continuous sets of configurations that compose each state, and $\pi_j = \int \mathrm{d}\boldsymbol{p} \int \mathrm{d}\boldsymbol{q} \; \frac{e^{-\beta \mathcal{H}(\boldsymbol{q},\boldsymbol{p})}}{Z} \chi_j(\boldsymbol{q})$ is the equilibrium probability of state $j$ with $Z$ being the canonical partition function of the system. Since the states are mutually disjoint, $\mathcal{C}(0) = \mathbf{1}$. These correlation functions, together the transition probability matrix (TPM), correspond to the conditional probability of finding the biomolecular complex in configuration $k$ at time $t$ given that it started in configuration $j$ at $t = 0$. We now turn to both Markovian (MSMs) and non-Markovian (GMEs) descriptions of the dynamics of the TPM, $\mathcal{C}(t)$.

**MSMs and qMSMs.** After configuration space has been partitioned into non-overlapping states (22), to obtain a valid Markovian description of the TPM dynamics, the MSM framework requires one to identify the smallest timescale $\tau_L$ such that the TPM satisfies the Chapman-Kolmogorov condition (18, 40),

$$\mathcal{C}[(n+1)\tau_L] = e^{\mathcal{M}\tau_L} \mathcal{C}(n\tau_L). \qquad [2]$$

Here, $\mathcal{M}$ is a time-independent rate matrix and $\tau_L$ is defined to be the Markovian lag time. In practice, Eq. 2 is rearranged such that $\tau_L$ is found by identifying the onset of a plateau in the implied timescale (ITS), defined as

$$\mathrm{ITS}(t) = -t[\log \mathcal{C}(t)]^{-1}. \quad [3]$$

This timescale is associated with the time taken for degrees of freedom within the aggregated states to achieve equilibrium and thus for the systems to become memoryless, or Markovian. Once $\tau_L$ is identified, the configurational dynamics can be predicted at integer multiples of $\tau_L$. In other words, $\tau_L$ defines the interval at which a given (non-Markovian) biomolecular process can be viewed as Markovian. Consequently, the resulting dynamics are discontinuous (40), thus obscuring the observations of dynamical processes which may occur on the interval $[n\tau_L, (n+1)\tau_L]$. Furthermore, Eq. (2) implies $\tau_L$ sets the lower bound on MD simulation time required to parameterize the MSM that describes $\mathcal{C}(t)$ (20). There is, however, no guarantee that intrastate equilibration will occur within an affordable timescale to perform MD (16).

Recent work has shown that it is possible to employ a GME approach to account for the effect of memory (non-Markovian) behavior at early times, allowing one to construct a quasi-Markov State Model (qMSM), given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{C}(t) = \dot{\mathcal{C}}(0)\mathcal{C}(t) + \int_0^t \mathrm{d}s\, \mathcal{K}(s)\mathcal{C}(t-s). \quad [4]$$

We note Eq. 4 does not contain an "inhomogeneous term" (analogous to the random force in the language of the Generalized Langevin Equation (41, 42)) because the GME is parameterized with equilibrium MD simulations, which is consistent with the correlation functions of interest given by Eq. 1 (35, 43, 44). In Eq. 4, the potentially complex intrastate dynamics are encoded into the time dependent memory kernel $\mathcal{K}$ (35). Crucially, $\mathcal{K}$ decays to zero on a characteristic time-scale $\tau_K$, termed the kernel cutoff time, enabling one to approximate the upper limit of the integral in Eq. (4) as $\min\{\tau_K, t\}$. It has been further shown that $\tau_K \leq \tau_L$, illustrating that the qMSM approach strictly improves upon the MSM. It does this by reducing the amount of simulation time needed to capture the exact dynamics, while simultaneously giving access to the dynamical events occurring between multiples of $\tau_L$. Indeed, the qMSM offers remarkable accuracy, temporal resolution, and often requires much less MD simulation time to fully construct the generator of the dynamics, i.e., the memory kernel $\mathcal{K}(t)$ (35). The qMSM has been profitably applied to, for example, understand the significance of the $\beta$-lobe of RNA polymerase during transcriptional initiation (36), and elucidate the mechanisms used by the RNA-induced silencing complex to recognize and target mRNA molecules in a sequence specific manner (45).

Unfortunately, the qMSM is not without its problems. First, evaluation of a convolution integral becomes computationally cumbersome as the dimension of the TPM increases. Second, constructing $\mathcal{K}$ requires the first and second derivatives of $\mathcal{C}$ (46), giving rise to numerical instabilities which we will analyze in a later section. Third, from a qualitative perspective, the qMSM approach obfuscates the physical interpretation of the MSM in terms of "states and rates". Specifically, the MSM provides a physically intuitive rate matrix, $\mathcal{M}$, whose diagonals can be interpreted as the likelihood of remaining in a particular state, and whose off-diagonals describe the probability of making a transition from one state to another. In contrast, the memory kernel appears under a convolution integral in the equation of motion for the TPM, Eq. 4, and therefore cannot be understood separately from its cumulative effect over the *history* of the TPM. Hence, the qMSM does not appear to offer a simple way to interpret the memory kernel matrix elements in terms of instantaneous transition rates, e.g., where a number twice as large can be immediately identified as taking half as long to move between two states in a given chemical scheme. These complications motivate the search for an alternative method that accurately and efficiently captures the exact dynamics in a robust, accurate, and intuitive manner.

**The TCL-GME.** For a non-Markovian theory, such as the qMSM, to be interpreted in terms of rates one would want to write it in a time-local form, comparable to Eq. (2). For this reason, we perform the formally exact rewriting of Eq. (4) as a time-convolutionless (TCL) GME (47–49),

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{C}(t) = \mathcal{R}(t)\mathcal{C}(t), \quad [5]$$

where $\mathcal{R}$ is the time-local generator that encodes the non-Markovian dynamics arising from imperfect timescale separation between intra- and interstate dynamics, and can be understood as a generalized time-dependent rate matrix. Furthermore, the matrix elements of the time-local generator plateau at a characteristic timescale, $\tau_R$ (38, 49), allowing one to separate the time over which non-Markovian evolution takes place ($0 \leq t < \tau_R$) and when Markovian evolution begins,

$$\mathcal{C}(t \geq \tau_R) = e^{\mathcal{R}_\infty (t-\tau_R)}\mathcal{C}(\tau_R), \quad [6]$$

where $\mathcal{C}(\tau_R) = \exp_{\rightarrow}[\int_0^{\tau_R} \mathrm{d}s\, \mathcal{R}(s)]\mathcal{C}(0)$ is the value of the TPM at $\tau_R$ given by the action of the time-ordered propagator on the initial condition, $\mathcal{C}(0) = \mathbf{1}$, and $\mathcal{R}_\infty \equiv \mathcal{R}(t \geq \tau_R)$ is the long-time limit of the time-local generator. $\mathcal{R}_\infty$ is the time-independent rate matrix that encodes the *true* Markovian evolution of $\mathcal{C}(t)$ beyond $\tau_R$ and elucidates the connection with Eq. (2).

Since the two timescales, $\tau_L$ and $\tau_R$, determine the minimal amount of simulation data required to fully construct the MSM and TCL-GME, respectively, it would be profitable to derive a relationship connecting the two quantities. In Appendix A, we analytically demonstrate that

$$\mathrm{ITS}(t \geq \tau_R) = -\left(\mathcal{R}_\infty + \frac{\Lambda}{t}\right)^{-1}, \quad [7]$$

where $\Lambda = \int_0^{\tau_R}[\mathcal{R}(s) - \mathcal{R}_\infty]\,\mathrm{d}s$ quantifies the deviation that intrastate motions cause on otherwise Markovian interstate transition rates. Comparing this to Eq. (3) allows us to state that

$$\tau_R \leq \tau_L. \quad [8]$$

Importantly, Eq. (8) demonstrates that the only cases where an MSM can be as data-efficient as the TCL-GME, albeit at the cost of a lower temporal resolution, is when $\Lambda = 0$. This inequality thus enforces a new lower bound on the amount of required simulation time and is one of the central results to the paper, demonstrating that *the TCL-GME always provides a description that is more data-efficient or, at worst, as data-efficient, as the MSM while retaining a high temporal*

*resolution*. What remains to be shown is the relative accuracy and efficiency of the TCL-GME approach in comparison with the qMSM. We will achieve this by comparing the performance of each dynamical approach on three different protein systems of varying levels of complexity: alanine dipeptide (35), the human argonaute complex (45), and the FiP35 WW domain (35, 50).

## All-atom Protein Systems

In what follows, we apply the TCL-GME to three systems of varying complexity—alanine dipeptide, argonaute, and FiP35 WW domain—and compare these predicted dynamics to those calculated by both the MSM and qMSM. Here, as previously stated, we do not consider the specifics of how to construct the reduced space but rather restrict our attention to their dynamics. Firstly, for alanine dipeptide we consider a 4-state model with metastable states corresponding to the molecule's free energy projected onto the backbone torsional angles $\{\psi, \phi\}$, as constructed in Ref. (35). Secondly, for argonaute, we use another 4-state model from structures corresponding to local minima in the free energy landscape of the first two slowest modes, as constructed in Ref. (45). Finally, for FiP35 WW domain, we use two reduced models: the first contains 3 states and its construction is detailed in Appendix B; the second contains 4 states corresponding to a folded state composed of two $\beta$-hairpins, an unfolded state, and structures corresponding to both on- and off-pathways, and its construction is outlined in Ref. (35). To clearly benchmark each method while illustrating its advantages and disadvantages, we show only one of the time-dependent conditional probabilities for each protein system. The full time-dependent conditional probability matrices are available in the Supporting Information.

**Alanine Dipeptide.** We begin our analysis of the TCL-GME and illustrate the utility of the inequality in Eq. (8) using a simple test system, alanine dipeptide. After obtaining TPM dynamics from MD simulation, we construct an MSM as discussed in Ref. (35), and we construct both the qMSM and TCL-GME as described in the Materials and Methods section. In Fig. 1(a) we identify the values of $\tau_R$, $\tau_K$, and $\tau_L$ using a root mean square error (RMSE) analysis (see Appendix F) that quantifies the deviation of the dynamics predicted as a function of $\tau_L$, $\tau_R$, and $\tau_K$ from the reference dynamics. We use a convergence threshold of 5% of the final value in the RMSE, which leads to graphical accuracy in the resulting dynamics This corresponds to quantitative agreement between the predicted dynamics and the MD data (open circles) as shown in Figs. 1(b),(c). For the qMSM and the TCL-GME, this leads to $\tau_R = \tau_K = 1.5$ ps, while for the MSM the lag time at the same error is $\tau_L = 10$ ps. The results in Fig. 1(b) show the dynamics that would result if one could only use TPM data, obtained from the MD, for the first 1.5 ps; such a choice of $\tau_L$ leads the MSM to severely overestimate the equilibration rate. In contrast, Fig. 1(c) shows how a valid MSM is able to capture the exact dynamics, albeit with severely reduced temporal resolution. The drawback of the finite resolution is visible at earlier times, where the (negative) curvature of the MD data is neglected by the MSM but captured by the GMEs. Together, the results of Fig. 1 show that the TCL-GME suffers no loss of performance with respect to the qMSM, with both GMEs able to make accurate high resolution predictions using only 15% of the MD data required to construct a valid MSM.

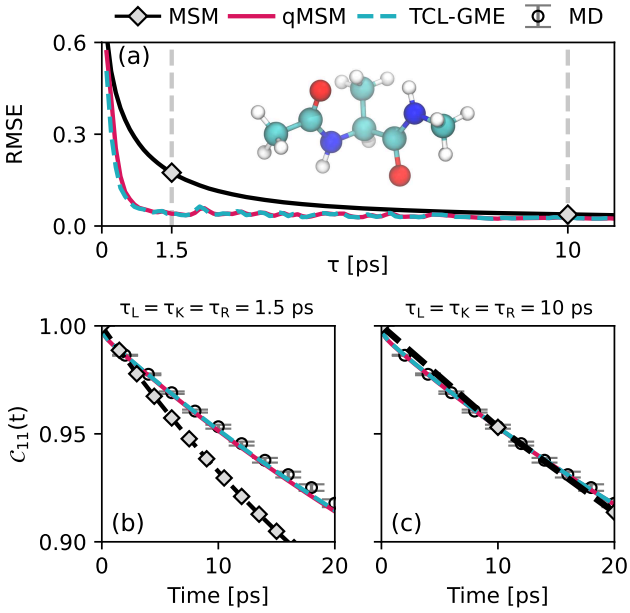


**Fig. 1.** Application of the TCL-GME to alanine dipeptide with comparisons to the MSM and qMSM (a) Root mean square error (RMSE) curves for the MSM, qMSM, and TCL-GME quantifying the deviation from the MD data (open circles) as the model is parameterized with increasing amounts of data (see Appendix F). Vertical lines show the errors associated with cutoffs ($\tau$) of 1.5 ps and 10 ps. Alanine dipeptide is shown (2 residues). (b) State 1 TPM dynamics, $\mathcal{C}_{11}(t)$, computed with MSM, qMSM, and TCL-GME approaches parameterized with 1.5 ps of MD data, i.e., $\tau_L = \tau_K = \tau_R = 1.5$ ps. (c) State 1 TPM dynamics computed with $\tau_L = \tau_K = \tau_R = 10$ ps. The 4-state TPMs parameterized with $\tau_K = \tau_R = 1.5$ ps and $\tau_L = 10$ ps are shown in Fig. S1. MD error bars were obtained using a bootstrapping approach as discussed in Ref. (35).

**Argonaute.** Will the simplistic form of Eq. (6) maintain a comparable level of performance to the qMSM for a much more complicated system? To address this, we consider the target recognition of human argonaute 2 complex (37, 51). It is challenging to obtain sufficient MD sampling to model the dynamics of this complex process, which involves coupled conformational changes of messenger RNA, microRNA, and the Argonaute protein. In fact, the ITS curves shown in Fig. 2 do not plateau over the available time window, demonstrating that the available TPM time is not sufficient to construct a valid MSM. That is to say, constructing an MSM is unaffordable at the same level of dimensionality reduction as the faithful qMSM (45).

Owing to the statistical noise that arises from averaging over limited MD data to construct the TPM (45), the numerically extracted $\mathcal{K}(t)$ and $\mathcal{R}(t)$ in Fig. 3(a) and (c) also display noise that makes it difficult to graphically identify their cutoff times, $\tau_K$ and $\tau_R$, respectively. To illustrate how both GMEs behave as the cutoff time $\tau$ is increased, we display the dynamics predicted from each method using representative cutoff choices of $\tau_R, \tau_K \in \{25, 35, 45, 55\}$ ns in Fig. 3(b),(d). Interestingly, the qMSM and TCL-GME perform similarly, with $\mathcal{K}$ and $\mathcal{R}$ predicting dynamics within the MD error bars using cutoff times of 35 ns. We emphasize that the ultimate goal is to identify the value of $\tau_R$ or $\tau_K$ such that the predicted dynamics match the MD reference data (open circles) exactly. Disappointingly, neither GME exhibits stability with respect to increasing $\tau_R$ or $\tau_K$, and the resulting RMSE curves do not monotonically converge towards zero (see Fig. S2). For

example, when we parameterize either model with the longer value of $\tau_K = \tau_R = 45$ ns, the resulting dynamics do not lead to an equilibrium value. This suggests that this truncation of the memory kernel or time-local generator fails to recover detailed balance.

This lack of controlled convergence can be rationalized by recalling that constructing the GME requires time derivatives of the MD data (See Materials and Methods, Eq. (14)). This is true for both $\mathcal{K}$ and $\mathcal{R}$. One might hypothesize that the noise in these under-converged MD data is sufficient to compromise the stability of both GME approaches for argonaute. Since TPMs at longer times—like other equilibrium time correlation functions (6, 52)—are constructed from averaging over less MD data, TPMs at longer times are beset by worse statistical errors. Hence, working with the hypothesis that the fluctuations at later times correspond to noise from statistically under-converged dynamics, we posit a method which averages the noise in $\mathcal{R}$ at long times. In fact, during the qMSM approach, truncation at $\tau_K$ equates to replacing $\mathcal{K}$ with its long-time average. However, while $\mathcal{K}(t \to \infty) \to 0$ for dissipative problems that equilibrate, we can only estimate it for $\mathcal{R}$.

Visually, Fig. 3(c) suggests that $\mathcal{R}(t)$ starts to oscillate around its long-time limit around $t = 10$ ns. Thus we introduce an averaging scheme where at $\tau_R$ we replace $\mathcal{R}$ with $\langle \mathcal{R}_\infty \rangle$, its time average over the interval $[t_r, \tau_R]$. Here, we choose $t_r$ to be the time where the time-local generator appears to have plateaued (See Appendix D). We identify $t_r = 10$ ns and show the corresponding $\mathcal{R}_{22}$ matrix element for $\tau_R = 30$ ns in Fig. 3(e). As Fig. 3(f) shows, with this simple adjustment the TCL-GME converges to the reference dynamics within 55 ns, which strictly improves upon both the MSM and qMSM constructed from the same data. Moreover, the convergence of the TCL-GME with increasing values of $\tau_R$ is monotonically decreasing (see Fig. S3).

A closer look at Fig. 3(f) reveals that the averaging scheme approaches the reference dynamics from below, but does not actually obtain perfect agreement within these 150 ns. To remedy this, one could average $\mathcal{R}(t)$ for longer to get a better estimate for $\mathcal{R}(\tau_R)$. However, this would run counter to our objective of working with the minimal possible MD data.

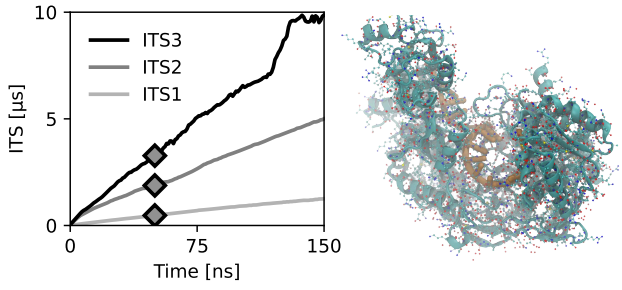


**Fig. 2.** Demonstration that the massive spatial and temporal scales of the argonaute protein present a challenge to MSMs. **Left**: Implied timescales (ITS) plot of Eq. 3, for the three non-unitary eigenvalues, whose plateau time corresponds to the Markovian lag time, $\tau_L$. Diamonds show the choice of $\tau_L$ in Fig. 3, but one can appreciate that no choice for this window of MD data would be satisfactory. Using the $\langle \mathcal{U} \rangle$-GME approach (discussed in this section), Markovianity is found to require $\sim 1200$ times as much simulation data. **Right**: Rendering of the argonaute protein containing the mRNA strand used to obtain the MD data. The protein itself is composed of 831 residues.

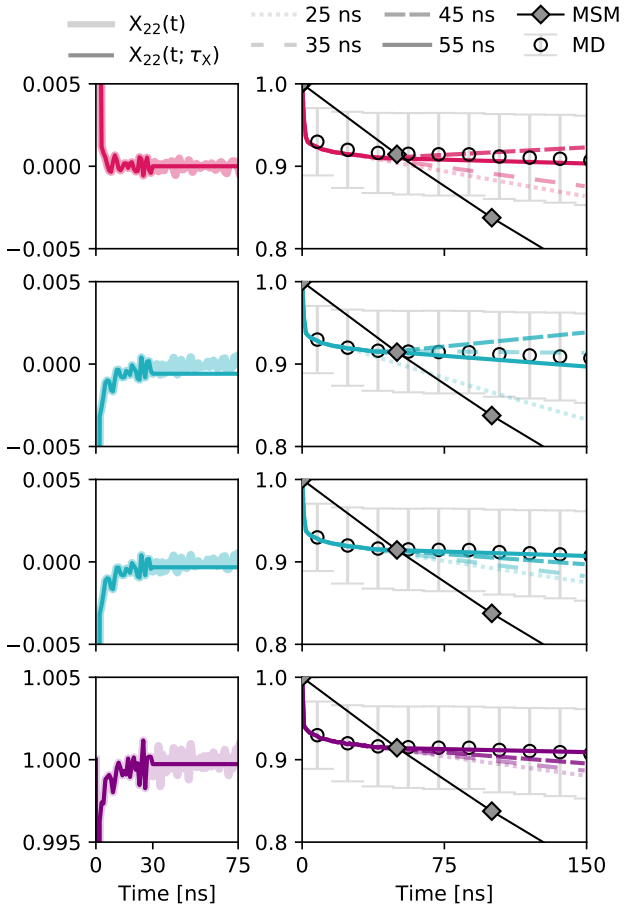


**Fig. 3.** Instability of the qMSM and TCL-GME in the case of the argonaute protein and demonstration of the robustness of our $\langle \mathcal{U} \rangle$-GME approach. (a) The transparent line shows the state 2 memory kernel $\mathcal{K}_{22}(t)$ as a function of time. From the RMSE [see Fig. S2(a)], we observe that $\mathcal{K}(t)$ converges by 35 ns. The solid line shows the replacement of $\mathcal{K}_{22}(t)$ with zero after this time. (b) Time-dependent conditional probability of starting in state 2 and remaining in state 2 (state 2 dynamics) predicted using the qMSM with $\tau_K \in \{25, 35, 45, 55\}$ ns, where increasing transparency corresponds to decreasing values of $\tau_K$. (c) Similar to (a), the transparent line shows the state 2 time-local generator $\mathcal{R}_{22}(t)$ as a function of time, and the solid line shows the replacement of $\mathcal{R}(t)$ with $\mathcal{R}(\tau_R)$ after $\tau_R = 30$ ns. (d) State 2 dynamics predicted using the TCL-GME with $\tau_R \in \{25, 35, 45, 55\}$ ns, where increasing transparency corresponds to decreasing values of $\tau_R$. (e) Like (c), the transparent line shows $\mathcal{R}_{22}(t)$ as a function of time. Here, the solid line is instead illustrating the replacement of $\mathcal{R}(t)$ with its time average over the window $[20, 30]$ ns after$\tau_R = 30$ ns, i.e., $(t_r, \tau_R) = (20, 30)$ ns. (f) Dynamics predicted using the $\langle \mathcal{R} \rangle$-GME. (g) The transparent line shows the propagator $\mathcal{U}_{22}(t)$ as a function of time, and the solid line shows the replacement of $\mathcal{U}(t)$ with its average over the window $[20, 30]$ ns after $\tau_R = 30$ ns. (h) Dynamics predicted using the $\langle \mathcal{U} \rangle$-GME. In (b), (d), (f), and (h) we show an MSM parameterized with $\tau_L = 50$ ns. The MD data and error bars were computed using the bootstrapping approach (see Ref.(45) for details).

Additionally, as one can appreciate from Eq. (6), error in the estimation of $\langle \mathcal{R}_\infty \rangle$ is exponentiated when predicting the GME dynamics. To this end, we propose an alternative route to employ the TCL-GME formalism without requiring any time derivatives or exponentiation of noisy data (38). This simply requires re-casting Eq. (6) as

$$\mathcal{C}(t) = \mathcal{U}(t, t_0)\, \mathcal{C}(t_0). \qquad [9]$$

That is, we now work directly with the time-dependent propagator (53), $\mathcal{U}(t, t_0)$, whose construction is detailed in Ap-

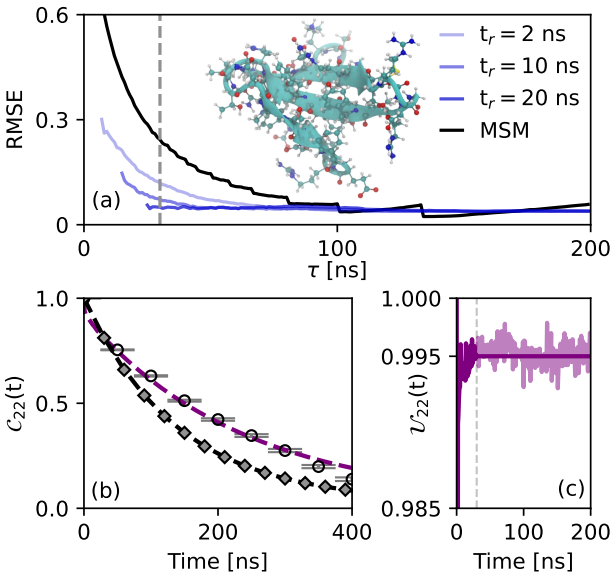


**Fig. 4.** Ability of our $\langle\mathcal{U}\rangle$-GME to accurately predict the dynamics of the FiP35 WW domain. (a) RMSE curves for the MSM and the $\langle\mathcal{U}\rangle$-GME as a function of $\tau_L$ and $\tau_R$, while varying choices of $t_r$ to illustrate convergence. The structure of the FiP35 WW domain is shown (35 residues). (b) TPM dynamics ($\mathcal{C}_{22}(t)$) computed using $\langle\mathcal{U}\rangle$-GME and MSM approaches with $\tau_R = 25$ ns ($\ell = 5$ ns) and $\tau_L = 25$ ns. (c) The propagator $\mathcal{U}_{22}(t)$ as a function of time, showing that $\mathcal{U}$ has been replaced with its average at 25 ns.

pendix E. This obviates integration of Eq. (5), and so the noise in the data is never exponentiated during our calculations. Moreover, this method has shown to be robust with respect to low resolution dynamical data in quantum dynamical problems (38). Importantly for the protein folding problem, both the time-local interpretability and frugality that result from the plateau at $\tau_R$ are unaffected by this manipulation.

Here we extend the protocol proposed in Ref. (38) by combining the direct calculation of $\mathcal{U}$ with the aforementioned averaging scheme. This results in our most direct and noise resilient TCL-GME formulation. We identify $t_r$ to be 10 ns and, in Fig. 3(g)–(h), we show the results of this $\langle\mathcal{U}\rangle$-GME. Here, with only minimal adjustments to the original formulation, the $\langle\mathcal{U}\rangle$-GME monotonically converges to the MD data within 55 ns, maintaining the strict improvements of the TCL-GME over both MSM and qMSM approaches.

With the convergent and stable $\langle\mathcal{U}\rangle$-GME dynamics obtained above, we can now determine the true lag time required for a valid MSM description of the dynamics of the 4 states used to elucidate mRNA recognition in the argonaute complex in Ref. (45). To do this, we employ the $\langle\mathcal{U}\rangle$-GME to predict the TPM dynamics at long times and use Eq. (3) to obtain the ITS plot (Fig. S4). We observe that the ITS curves only plateau by $t \sim 60$ µs, indicating that $\tau_L$ is 1200 times larger than the MSM constructed Ref. (45). By comparison the time-local generator cutoff used in our $\langle\mathcal{U}\rangle$-GME, $\tau_R \sim 50$ ns, is more than 3 orders of magnitude smaller, demonstrating that our approach provides a highly compact and efficient means to fully capture the short- as well as long-time dynamics of complex biomolecular systems.

**FiP35 WW-domain.** The $\langle\mathcal{U}\rangle$-GME method requires two convergence parameters: $t_r$, the beginning of the averaging window, and $\tau_R$, the total amount of MD simulation time required to parameterize the model (see Appendix E). This begs an important practical question: how does one choose $t_r$ when the onset of the plateau in $\mathcal{U}$ is hidden under the noise? After all, one might expect to observe a lack of convergence when $t_r$ is chosen to be too early. However, by considering a 4-state model of FiP35 WW domain, we find that this is not the case. In this system, where the plateau is not visually obvious (see Fig. 4(c)), we observe that for every choice of $t_r$, there is a value of $\tau_R$ capable of accurately capturing the reference dynamics. In Fig. 4(a) we demonstrate that the $\tau_R$ required for the $\langle\mathcal{U}\rangle$-GME to provide accurate dynamics merely increases as $t_r$ is reduced to zero. Indeed, since we know from Eq. (8) that $\tau_R$ is bounded above by $\tau_L$, if $t_r$ is given the extreme value of zero then the $\langle\mathcal{U}\rangle$-GME reduces to the MSM, with the important distinction that it is able to capture the dynamics between MSM points (see Fig. S5; we also give the mathematical justification for this result in Appendix E). In this sense, the $\langle\mathcal{U}\rangle$-GME parameterized with $t_r = 0$ constitutes a higher-resolution MSM. The practical implication of this is that while one may make a poor choice of $t_r$ to begin averaging from, one will only pay for this in the length of MD data required to construct the model, $\tau_R$, and not in the final accuracy of the $\langle\mathcal{U}\rangle$-GME dynamics.

The best, earliest choice of $\tau_R$ is therefore parametrically dependent on $t_r$, but well defined. Since all choices of $t_r$ converge to the same RMSE value, $\tau_R$ is robustly identified by a common convergence threshold. To identify the optimal $(t_r, \tau_R)$ pair, we simply find the minimum of the plot of $\tau_R$ as a function $t_r$. Choosing a value of 5% error as converging to the MD dynamics within visual accuracy (see Appendix F), for these FiP35 WW domain data we identify $t_r = 20$ ns, $\tau_R = 25$ ns, and $\tau_L = 200$ ns, as shown in Fig. 4(a). For comparison, we display the dynamics predicted by both the MSM and $\langle\mathcal{U}\rangle$-GME when parameterized using only these 30 ns of MD data in Fig. 4(b). In Fig. 4(c), we show the replacement of $\mathcal{U}$ with its average $\langle\mathcal{U}\rangle$ (obtained over the averaging interval of $[20, 25]$ ns). We observe that MSM dynamics predicted using only 25 ns of the MD data set overestimates the equilibration rate, as was the case with alanine dipeptide and the argonaute complex, whereas the $\langle\mathcal{U}\rangle$-GME parameterized with the same amount of reference data accurately captures the MD data until $\sim 375$ ns. The small deviation that starts at $\sim 375$ ns disappears at longer times, where the $\langle\mathcal{U}\rangle$-GME correctly captures the long-time limit (see Fig. S6). Thus, our analysis shows that accurate predictions of the dynamics from the $\langle\mathcal{U}\rangle$-GME require only 15% of the MD data needed to construct a valid MSM.

We now consider the ability of the $\langle\mathcal{U}\rangle$-GME to capture the long-time dynamics through a different, experimentally accessible measure: the folding time of the protein. For this, we will consider a 3-state model of FiP35 WW domain (for construction details, see Appendix B) with states one, two, and three corresponding to misfolded, unfolded, and folded structures of the protein, respectively (50). Here, we compute the folding time using the mean first passage time (MFPT) procedure outlined in Appendix G. First, we use the reference dynamics (Fig. S7) to compute the folding time to be $\tau_{\text{ref}} = 18.65$ µs (Fig. S8), which is taken to be the *exact* result for this model, which is in reasonable agreement with the experimentally measured value of $14 \pm 1.5$ µs (54). In particular, if the clustering algorithm does not correctly identify configurations with the

folded, unfolded, and misfolded states, this may cause the folding time to appear artificially long. Hence, we focus not on the deviation from the experimental value but rather on the internal consistency between the reference dynamics and the predictions from the $\langle\mathcal{U}\rangle$-GME and the MSM approaches. To obtain the $\langle\mathcal{U}\rangle$-GME predictions of the MFPT, we first identify $t_r = 50$ ns. As described in Appendix G, we compute the MFPTs corresponding to increasing values of $\tau_R$ and $\tau_L$ and observe that both the $\langle\mathcal{U}\rangle$-GME and MSM approaches converge to the reference result at long times (see Fig. S8). We also find that the MSM continuously underestimates $\tau_{\text{ref}}$ and appears to continue increasing at times beyond 1000 ns (see Fig. S8). In contrast, the $\langle\mathcal{U}\rangle$-GME remains within 8% of the reference value for the duration of available MD data. Indeed, to converge within 5% error, the $\langle\mathcal{U}\rangle$-GME requires data up to 168 ns, whereas the MSM does not reach this threshold until 452 ns, suggesting that the $\langle\mathcal{U}\rangle$-GME provides, even in the estimation of folding times, a more efficient means to capture the long-time dynamics of complex biomolecular systems. This is in agreement with previous works demonstrating that a purely Markovian process fails to faithfully capture barrier crossing phenomena (31, 55).

## Conclusion

In this work, we have developed and applied the $\langle\mathcal{U}\rangle$-GME and demonstrated that it is an accurate, chemically intuitive, and systematically improvable approach to modeling non-Markovian biomolecular dynamics. While previous work had exploited the memory of the MSM's intrastate motions to construct an exact qMSM that could significantly reduce the computational cost required to efficiently predict protein dynamics at long times, it eluded a simple and intuitive chemical interpretation and, as we show here, is highly sensitive to statistical noise in the reference TPM dynamics from which it must be constructed. Here, we have abandoned the time-nonlocal qMSM by moving to a time-convolutionless formulation which admits a simple formal integration, elucidating the analytical connection between GMEs and MSMs and permits a simple interpretation. In particular, not only does this allow the time-local generator to be interpreted as a time-dependent rate matrix, it also allows for systematic improvement in regimes of noisy data. Specifically, we have identified that for cases where the reference TPM suffers from statistical noise (e.g., the argonaute system), a straightforward averaging scheme allows our time-convolutionless approaches (both $\mathcal{R}(t)$ and $\mathcal{U}(t)$) to uniformly converge to the reference dynamics. In contrast, the time-nonlocal approach displays instabilities with increasing simulation time that have no comparable solution without resorting to manipulations of the qMSM formalism such as the introduction of an integrative form of the GME (56). Furthermore, using alanine dipeptide, FiP35 WW Domain, and argonaute, we have demonstrated that the time-local GME can accurately and efficiently capture short-, intermediate-, and long-time dynamics with no loss of performance. Not only does this approach require an equivalent amount of data as the qMSM, the $\langle\mathcal{U}\rangle$-GME requires minimal numerical and physical complexity by eliminating the need for both time-convolution integrals and numerical time derivatives of potentially noisy data. By providing a rigorous and physically transparent method to capture the non-Markovian dynamics of a given set of states, we expect the $\langle\mathcal{U}\rangle$-GME to provide a robust scaffold to construct novel methods to find optimal configuration clusters and offer a framework to investigate the mechanisms of complex biomolecular conformational changes.

## Materials and Methods

**A. Rigorous Connection of MSM with TCL-GME.** Here, we derive Eq. (7) and Eq. (8) from the main text, which rigorously connect the MSM to the TCL-GME. We begin by considering some time $t$ that is strictly greater than $\tau_R$ and re-writing $\mathcal{C}(t)$ as

$$\begin{aligned}\mathcal{C}(t) &= \exp_{\rightarrow}\left(\int_0^t \mathcal{R}(s)\,\mathrm{d}s\right)\mathcal{C}(0)\\ &= \exp\Big(\mathcal{R}_\infty(t-\tau_R)\Big)\exp_{\rightarrow}\left(\int_0^{\tau_R}\mathcal{R}(s)\,\mathrm{d}s\right)\\ &= \exp\Big(\mathcal{R}_\infty(t-\tau_R)\Big)\mathcal{U}_{nM}(\tau_R,0)\end{aligned} \quad [10]$$

where we have used the fact that the initial condition is the identity matrix, $\mathcal{C}(0) = \mathbf{1}$ and have introduced $U_{nM}(\tau_R, 0)$ as the propagator over the non-Markovian region. This is equivalent to Eq. (6) in the main text. We insert the above result into the implied timescale equation, defined in Eq. (3), to obtain the result in the main text,

$$\mathrm{ITS}(t) = -\left(\mathcal{R}_\infty + \frac{\Lambda}{t}\right)^{-1}, \quad [7]$$

where

$$e^{\Lambda} \equiv e^{-\mathcal{R}_\infty\tau_R}U_{nM}(\tau_R,0). \quad [11]$$

Equation Eq. (11) is exact and easy to calculate given the framework presented here for obtaining the non-Markovian propagator; it can be interpreted as the total deviation in the propagation due to non-Markovian behavior. Keeping in mind that the MSM lag time is taken to be the minimum time-scale associated with the onset of a plateau in an ITS plot, we see see that the right-hand-side of Eq. (7) does not necessarily stabilize for times immediately after $\tau_R$. This allows us to conclude the inequality presented in the main text, that

$$\tau_R \leq \tau_L. \quad [8]$$

To further simplify its interpretation, one can neglect the effect of time-ordering in the definition of the non-Markovian propagator, which yields the following, modified expression for $\Lambda$,

$$\Lambda \approx \int_0^{\tau_R}[\mathcal{R}(s) - \mathcal{R}_\infty]\,\mathrm{d}s. \quad [12]$$

Here, it is clear that $\Lambda$ approximately corresponds to the integral deviation between the time-local generator over its non-Markovian variation, and its long-time limit.

**B. TPM Construction.** The transition probability matrix (TPM) is computed from the transition count matrix (TCM). We first computed the TCM from the MD trajectories. For each lag time $\tau$, the raw TCMs ($T^{\text{raw}}$) were first counted from transition pairs between frames at $t$ and $t+\tau$: $T_{ij}^{\text{raw}}(\tau) = \langle\chi_i(t+\tau)\chi_j(t)\rangle$, where $\chi_i(t)$ is the indicator function that determines if the frame at time $t$ is in state $i$. Here, $t = 0$, $\Delta t$, $2\Delta t$, ...,$(N_{\text{traj}} - 1)\Delta t - \tau$, where $\Delta t$ is the saving interval of trajectories, and $N_{\text{traj}}$ is the length of trajectories. Normally,

detailed balance requires that the TCM be symmetric, i.e., $T_{ij} = T_{ji}$. However, since the raw TCMs are normally not symmetric, we further symmetrize the raw TCMs to enforce detailed balance: $T_{ij}(\tau) = (T_{ij}^{\text{raw}}(\tau) + T_{ji}^{\text{raw}}(\tau))/2$ (57). Finally, we calculated TPMs by column-normalizing the TCM: $\mathcal{C}_{ij}(\tau) = T_{ij}(\tau)/\sum_j T_{ij}(\tau)$.

In our TPM construction, the raw TCM was directly counted from the macrostate models. The 4-state model of alanine dipeptide was constructed with a splitting-and-lumping approach. We first split all the available MD conformations into 1000 microstates using the K-Centers clustering algorithm (58–60). Then we lumped the 1000 microstates into 4 macrostates via the PCCA+ (Perron Cluster Cluster Analysis) (61, 62), with the lag time of 2 ps.

We constructed the 3-state model of the FiP35 WW domain using tICA (Time-lagged Independent Component Analysis) (57, 63), K-Centers clustering (58–60) and PCCA+ (Perron Cluster Cluster Analysis) (61, 62) lumping from MD trajectories provided by D. E. Shaw research. We first performed tICA with pairwise distances between all $\alpha$ carbon atoms of the peptide with a lag time of 10 ns. Then we used the K-Centers algorithm to generate a 1000-state model based on the top three tICs (Time-lagged Independent Components) from tICA. Finally, we performed the PCCA+ clustering to generate the 3-state model based on the 1000-state TPM computed at the lag time of 10 ns.

We constructed the 4-state model of the argonaute using spectral oASIS (64), tICA, APLoD clustering, and PCCA (61, 65). We employed spectral oASIS to reduce the number of input features, followed by tICA for the dimensionality reduction. Then we grouped the conformations into 81 clusters from the APLoD clustering algorithm, based on the top 4 tICs from tICA. Finally, we used the PCCA+ algorithm to group the 81 microstates into four macrostates.

**C. qMSM Construction.** To solve the integro-differential equation in Eq. (4), we must first construct the memory kernel, $\mathcal{K}(t)$, as a function of time directly from the TPM data. We follow Ref. (46) and derive the classical analogue of the self-consistent expansion of the memory kernel

$$\mathcal{K}(t) = \mathcal{K}^{(1)}(t) + \int_0^t \mathrm{d}\tau\, \mathcal{K}^{(3)}(t-\tau)\mathcal{K}(\tau), \qquad [13]$$

where

$$\begin{aligned} \mathcal{K}^{(1)}(t) &= \ddot{\mathcal{C}}(t) - \{\dot{\mathcal{C}}(0), \dot{\mathcal{C}}(t)\} + \dot{\mathcal{C}}(0)\mathcal{C}(t)\dot{\mathcal{C}}(0) \\ \mathcal{K}^{(3)}(t) &= \dot{\mathcal{C}}(0)\mathcal{C}(t) - \dot{\mathcal{C}}(t) \end{aligned} \qquad [14]$$

are the projection-free auxiliary kernels.

To compute both $\dot{\mathcal{C}}$ and $\ddot{\mathcal{C}}$, and to thus compute $\mathcal{K}^{(1)}(t)$ and $\mathcal{K}^{(3b)}(t)$, we numerically differentiate the TPM data, $\mathcal{C}(t)$. With these auxiliary kernels, we compute $\mathcal{K}$ according to Eq. (13) using the discretization procedure in Ref. (66). For completeness, we summarize the algorithm. At the initial time and first timestep, $\mathcal{K}(t_0) = \mathcal{K}^{(1)}(t_0)$ and

$$\begin{aligned} \mathcal{K}(t_1) = \left[\mathbf{1} - \tfrac{\Delta t}{2}\mathcal{K}^{(3)}(t_0)\right]^{-1} \Big[&\mathcal{K}^{(1)}(t_1) \\ &+ \tfrac{\Delta t}{2}\mathcal{K}^{(3)}(t_1)\mathcal{K}(t_0)\Big]. \end{aligned} \qquad [15]$$

For all subsequent times ($n \geq 2$),

$$\begin{aligned} \mathcal{K}(t_n) = \left[\mathbf{1} - \tfrac{\Delta t}{2}\mathcal{K}^{(3)}(t_0)\right]^{-1} \Big[&\mathcal{K}^{(1)}(t_n) \\ &+ \tfrac{\Delta t}{2}\mathcal{K}^{(3)}(t_n)\mathcal{K}(t_0) + \Delta t \sum_{j=1}^{n-1} \mathcal{K}^{(3)}(t_{n-j})\mathcal{K}(t_j)\Big]. \end{aligned} \qquad [16]$$

Here $\mathbf{1}$ is the identity matrix and we employ equally spaced time intervals, such that $\Delta t \equiv t_{j+1} - t_j$.

Once we construct $\mathcal{K}(t)$, we employ Heun's method (second-order accurate with respect to $\Delta t$) to integrate Eq. (4) and obtain $\mathcal{C}(t)$. Then we identify an *appropriate* memory kernel cutoff time, $\tau_K$, by applying the RMSE analysis in Appendix C. We approximate the upper limit of the integral in Eq. (4) with $\tau_K$, enabling us to predict the dynamics for times beyond the duration of the MD simulation.

**D. TCL-GME Construction.** To reap the benefits of the time-local formalism, we first calculate $\mathcal{R}(t)$ from the TPMs obtained from MD simulation. We do this by rearranging Eq. (5) via matrix inversion to obtain

$$\mathcal{R}(t) = \dot{\mathcal{C}}(t)[\mathcal{C}(t)]^{-1}, \qquad [17]$$

where we calculate $\dot{\mathcal{C}}$ by numerically differentiating the TPM data. As we have discussed, the matrix elements of $\mathcal{R}$ plateau on a timescale, $\tau_R$, associated with the conclusion of non-Markovian evolution, allowing us to set $\mathcal{R}(t > \tau_R) = \mathcal{R}_\infty \equiv \mathcal{R}(\tau_R)$. With this definition, we can describe the dynamics after the onset of Markovian evolution, as shown in Eq. (6).

Once we find $\mathcal{R}(t)$, we employ Heun's method to integrate Eq. (5) and obtain $\mathcal{C}(t)$. Similar to the discussion in Appendix C, we identify an *appropriate* generator cutoff time, $\tau_R$, using the RMSE analysis discussed in Appendix F.

**E. $\langle\mathcal{U}\rangle$-GME Construction.** We first formally integrate the TCL-GME in Eq. (5) to obtain

$$\mathcal{C}(t + \Delta t) = \mathcal{U}(t + \Delta t, t)\mathcal{C}(t), \qquad [18]$$

where we have defined $\mathcal{U}(t+\Delta t, t) \equiv \exp_{\rightarrow}[\int_t^{t+\Delta t} \mathrm{d}s\, \mathcal{R}(s)]$ with the "+" subscript denoting the chronological time-ordering of the exponential, as above. We then compute the value of $\mathcal{U}(t + \Delta t, t)$ through direct matrix inversion

$$\mathcal{U}(t + \Delta t, t) = \mathcal{C}(t + \Delta t)[\mathcal{C}(t)]^{-1}, \qquad [19]$$

as introduced in Ref. (38). Because $\mathcal{R}$ becomes constant at $\tau_R$, the propagator $\mathcal{U}$ also becomes a constant. Hence, we define $\mathcal{U}_\infty(\Delta t) \equiv \exp[\mathcal{R}_\infty \Delta t]$. We compute the dynamics beyond $\tau_R$ according to

$$\mathcal{C}(\tau_R + n\Delta t) = [\mathcal{U}_\infty(\Delta t)]^n \mathcal{C}(\tau_R). \qquad [20]$$

As discussed in our analysis of the argonaute complex, we developed and implemented a simple averaging scheme capable of taming noise arising from statistically underconverged MD estimates of the TPM. We begin by applying the RMSE stability analysis in Appendix F to determine a valid generator cutoff time; here, we denote this cutoff by $t_r$. We then introduce another parameter $\ell$ that represents the number of *high quality* TPMs after $t_r$ and denote the corresponding time as $t_{r+\ell}$. This number is, of course, limited by data availability.

To predict the dynamics beyond $t_{r+\ell}$, we compute the time average of $\mathcal{U}$ on the time interval $[t_r, t_{r+\ell}]$ using

$$\langle \mathcal{U} \rangle = \frac{1}{\ell} \sum_{n=r}^{r+\ell-1} \mathcal{U}(t_{n+1}, t_n). \quad [21]$$

Because our $\langle \mathcal{U} \rangle$-GME requires at least $r + \ell$ data points to circumvent the instabilities imposed by noise in biomolecular systems, we generalize our the definition of the generator cutoff time to be $\tau_R = t_{r+\ell}$, *representing not the generator cutoff but rather the minimum amount of data needed to accurately predict the true TPM dynamics.* Ultimately, we recommend that the user performs a rigorous stability analysis with respect to the choices of $r$ and $\ell$.

It can be seen by equating expressions Eq. (6) and Eq. (2) given the same first time step ($\tau = \tau_L = \tau_R$),

$$\begin{aligned} \exp \mathcal{M}\tau &= \exp_{\rightarrow} \left( \int_0^\tau \mathcal{R}(s)\, \mathrm{d}s \right) \\ &\approx \exp \left( \int_0^\tau \mathcal{R}(s)\, \mathrm{d}s \right), \end{aligned} \quad [22]$$

where the right-hand side of Eq. (22) uses the explicit form of the propagator (see Appendix A for details). If this time-ordering of the exponential can be neglected, then we can identify $\mathcal{M} \approx \langle \mathcal{R} \rangle$. The practical implication of this is that, if we can replace $\langle \exp \mathcal{R}\tau \rangle$ with $\exp \langle \mathcal{R} \rangle \tau$ in the $\langle \mathcal{U} \rangle$-GME, we will obtain exact agreement with the MSM parameterized by the same $\tau$ (at integer multiples of $\tau$). The requirement for this to be true is that $\mathcal{U} \sim \mathbf{1}$, which we show to be satisfied by panel (c) of Fig. 4. Since $\mathcal{R}$ and therefore $\mathcal{U}$ are formally exact before cutoff (by construction they return the reference dynamics (38)), the dynamics between these MSM points is also accessible to the $\langle \mathcal{U} \rangle$-GME. This explains why Fig. 4(c) shows that $\lim_{t_r \to 0}(\tau_R) = \tau_L$.

**F. RMSE Analysis.** To determine values of $\tau_x \in \{\tau_L, \tau_R, \tau_K\}$, we find the lowest time by which the time-averaged root mean squared error (RMSE), given by

$$\mathrm{RMSE}(\tau_x) = \left( \frac{1}{N_t} \sum_{i=0}^{N_t} \sum_{j,k}^{n} \left[ \mathcal{C}_{jk}^{\mathrm{MD}}(i) - \mathcal{C}_{jk}(i; \tau_x) \right]^2 \right)^{1/2}, \quad [23]$$

becomes and stays sufficiently small. We identify $\tau_x$ to be this minimum amount of time. The RMSE quantifies the error associated with the dynamics predicted by a method (i.e., MSM, qMSM, TCL-GME, and $\langle \mathcal{U} \rangle$-GME) as a function of $\tau_x$ by comparing it pointwise to the reference dynamics obtained from MD over the length of the trajectory, $N_t$, which we take to be the 'exact' result. Here, $n$ corresponds to the total number of macrostates, i.e. we sum over all elements of the $\mathcal{C}(t)$ matrix, not just the representative elements we display. In the absence of noise, these error curves are expected to monotonically tend towards zero. In practice, however, this is not the case [See SI Fig 2(a)]. Therefore, the user must determine an acceptable threshold for a particular application. In our results, we choose the RMSE to be $\sim 5\%$, which results in graphical agreement between the reference and GME or MSM dynamics.

**G. MFPT method.** We apply our newly developed $\langle \mathcal{U} \rangle$-GME to compute folding times for FiP35 WW Domain. To do so, we consider a 3-state model where states one, two, and three are defined to be the misfolded, folded, and unfolded structures, respectively. To employ Meyer's mean-first passage time (MFPT) method (67, 68), we construct the time-dependent MFPT matrix, $M$, as

$$M_{ij}(t) = t + \sum_{k \neq i} M_{ik}(t) \mathcal{C}_{kj}(t). \quad [24]$$

The element $M_{32}$ then corresponds to the folding time in this problem.

Practically, one solves Eq. (24) as a system of linear equations (69). To solve for the MFPT corresponding to passage to state 3, the folded state, we consider the row 3 MFPT matrix elements and obtain the following system of equations

$$\begin{aligned} M_{31}(\tau) &= \tau + M_{31}(\tau)\mathcal{C}_{11}(\tau) + M_{32}(\tau)\mathcal{C}_{21}(\tau) \\ M_{32}(\tau) &= \tau + M_{31}(\tau)\mathcal{C}_{12}(\tau) + M_{32}(\tau)\mathcal{C}_{22}(\tau), \end{aligned} \quad [25]$$

We recast the system in terms of matrices and obtain the final form by matrix inversion,

$$\begin{bmatrix} M_{31} \\ M_{32} \end{bmatrix} = \tau \begin{bmatrix} 1 - \mathcal{C}_{11} & \mathcal{C}_{21} \\ \mathcal{C}_{12} & 1 - \mathcal{C}_{22} \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \quad [26]$$

As the dynamics approach equilibrium, the inverse matrix on the right-hand-side of Eq. (26) becomes constant. In practice, we define the folding time to be when $M_{32}/\tau$ is within 5% of $M_{32}(\tau_{\mathrm{final}})/\tau_{\mathrm{final}}$ for the rest of time.

## Acknowledgements

A.M.C. acknowledges the start-up funds from the University of Colorado, Boulder. X.H. acknowledges the Hirschfelder Professorship Fund. This work was supported by National Science Foundation (Grant No. CHE-2154291 to T.E.M.).

## References.

1. TK Chaudhuri, S Paul, Protein-misfolding diseases and chaperone-based therapeutic approaches. *FEBS J.* **273**, 1331 (2006).
2. CR Schwantes, RT McGibbon, VS Pande, Perspective: Markov models for long-timescale biomolecular dynamics. *J. Chem. Phys.* **141**, 090902 (2014).
3. W Wang, S Cao, L Zhu, X Huang, Constructing Markov State Models to elucidate the functional conformational changes of complex biomolecules. *WIREs Comput. Mol. Sci.* **8**, e1343 (2018).
4. VS Pande, K Beauchamp, GR Bowman, Everything you wanted to know about Markov State Models but were afraid to ask. *Methods* **52**, 99 (2010).
5. BE Husic, VS Pande, Markov State Models: From an Art to a Science. *J. Am. Chem. Soc.* **140**, 2386 (2018).
6. GR Bowman, VS Pande, F Noé, *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation.* Vol. 797, (2013).
7. NV Buchete, G Hummer, Coarse master equations for peptide folding dynamics. *J. Phys. Chem. B* **112**, 6057 (2008).
8. RD Malmstrom, CT Lee, AT Van Wart, RE Amaro, Application of molecular-dynamics based markov state models to functional proteins. *J. Chem. Theory Comput.* **10**, 2648 (2014).
9. X Huang, GR Bowman, S Bacallado, VS Pande, Rapid equilibrium sampling initiated from nonequilibrium data. *PNAS* **106**, 19765 (2009).
10. F Morcos, et al., Modeling conformational ensembles of slow functional motions in pin1-WW. *PLoS Comput. Biol.* **6**, e1001015 (2010).
11. BW Zhang, et al., Simulating Replica Exchange: Markov State Models, Proposal Schemes, and the Infinite Swapping Limit. *J. Phys. Chem. B* **120**, 8289 (2016).
12. AC Pan, B Roux, Building Markov state models along pathways to determine free energies and rates of transitions. *J. Chem. Phys.* **129**, 064107 (2008).
13. S Doerr, MJ Harvey, F Noé, G De Fabritiis, HTMD: High-Throughput Molecular Dynamics for Molecular Discovery. *J. Chem. Theory Comput.* **12**, 1845 (2016).
14. MK Scherer, et al., PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.* **11**, 5525 (2015).
15. MP Harrigan, et al., MSMBuilder: Statistical Models for Biomolecular Dynamics. *Biophys. J.* **112**, 10 (2017).

16. WC Swope, JW Pitera, F Suits, Describing protein folding kinetics by molecular dynamics simulations. 1. Theory. *J. Phys. Chem. B* **108**, 6571 (2004).
17. K Röder, DJ Wales, The Energy Landscape Perspective: Encoding Structure and Function for Biomolecules. *Front. Mol. Biosci.* **9** (2022).
18. JH Prinz, et al., Markov models of molecular kinetics: Generation and validation. *The J. Chem. Phys.* **134**, 174105 (2011).
19. F Nüske, BG Keller, G Pérez-Hernández, AS Mey, F Noé, Variational approach to molecular kinetics. *J. Chem. Theory Comput.* **10**, 1739 (2014).
20. KA Konovalov, IC Unarta, S Cao, EC Goonetilleke, X Huang, Markov State Models to Study the Functional Dynamics of Proteins in the Wake of Machine Learning. *JACS Au* **1**, 1330 (2021).
21. J Lu, E Vanden-Eijnden, Exact dynamical coarse-graining without time-scale separation. *J. Chem. Phys.* **141**, 044109 (2014).
22. A Kai-Hei Yik, Y Qiu, IC Unarta, S Cao, X Huang, A Step-by-step Guide on How to Construct quasi-Markov State Models to Study Functional Conformational Changes of Biological Macromolecules. *ChemRxiv* (2022).
23. ZF Brotzakis, M Parrinello, Enhanced Sampling of Protein Conformational Transitions via Dynamically Optimized Collective Variables. *J. Chem. Theory Comput.* **15**, 1393 (2019).
24. J Rogal, E Schneider, ME Tuckerman, Neural-Network-Based Path Collective Variables for Enhanced Sampling of Phase Transformations. *Phys. Rev. Lett.* **123**, 245701 (2019).
25. H Klem, GM Hocky, M McCullagh, Size-and-Shape Space Gaussian Mixture Models for Structural Clustering of Molecular Dynamics Trajectories. *J. Chem. Theory Comput.* **18**, 3218 (2021).
26. N Singhal, VS Pande, Error analysis and efficient sampling in Markovian state models for molecular dynamics. *J. Chem. Phys.* **123**, 204909 (2005).
27. G Hummer, A Szabo, Optimal Dimensionality Reduction of Multistate Kinetic and Markov-State Models. *J. Phys. Chem. B* **119**, 9029 (2015).
28. VA Voelz, GR Bowman, K Beauchamp, VS Pande, Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1-39). *J. Am. Chem. Soc.* **132**, 1526 (2010).
29. LT Da, et al., Bridge helix bending promotes RNA polymerase II backtracking through a critical and conserved threonine residue. *Nat. Commun.* **7**, 11244 (2016).
30. OF Lange, H Grubmüller, Collective Langevin dynamics of conformational motions in proteins. *J. Chem. Phys.* **124**, 214903 (2006).
31. C Ayaz, et al., Non-Markovian modeling of protein folding. *PNAS* **118**, e2023856118 (2021).
32. H Vroylandt, L Goudenège, P Monmarché, F Pietrucci, B Rotenberg, Likelihood-based non-Markovian models from molecular dynamics. *PNAS* **119**, e117586119 (2022).
33. C Ayaz, L Scalfi, BA Dalton, RR Netz, Generalized Langevin equation with a nonlinear potential of mean force and nonlinear memory friction from a hybrid projection scheme. *Phys. Rev. E* **105**, 054138 (2022).
34. I Horenko, C Hartmann, C Schütte, F Noe, Data-based parameter estimation of generalized multidimensional Langevin processes. *Phys. Rev. E* **76**, 016706 (2007).
35. S Cao, A Montoya-Castillo, W Wang, TE Markland, X Huang, On the advantages of exploiting memory in Markov state models for biomolecular dynamics. *The J. Chem. Phys.* **153**, 014105 (2020).
36. I Christy Unarta, et al., Role of bacterial RNA polymerase gate opening dynamics in DNA loading and antibiotics inhibition elucidated by quasi-Markov State Model. *PNAS* **118**, e2024324118 (2021).
37. G Meister, Argonaute proteins: Functional insights and emerging roles. *Nat. Rev. Genet.* **14**, 447 (2013).
38. T Sayer, A Montoya-Castillo, Compact and complete description of non-Markovian dynamics. *J. Chem. Phys.* **158**, 014105 (2023).
39. A Mardt, L Pasquali, H Wu, F Noé, VAMPnets for deep learning of molecular kinetics. *Nat. Commun.* **9**, 5 (2018).
40. JD Chodera, N Singhal, VS Pande, KA Dill, WC Swope, Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J. Chem. Phys.* **126**, 155101 (2007).
41. R Zwanzig, *Nonequilibrium Statistical Mechanics*. (Oxford University Press), (2001).
42. W Coffey, YP Kalmykov, JT Waldron, *The Langevin Equation: With Applications in Physics, Chemistry and Electrical Engineering*. (World Scientific) Vol. 14, 2 edition, (2004).
43. A Montoya-Castillo, DR Reichman, Approximate but accurate quantum dynamics from the Mori formalism. II. Equilibrium time correlation functions. *J. Chem. Phys.* **146**, 084110 (2017).
44. A Kelly, A Montoya-Castillo, L Wang, TE Markland, Generalized quantum master equations in and out of equilibrium: When can one win? *J. Chem. Phys.* **144**, 184105 (2016).
45. L Zhu, et al., Critical role of backbone coordination in the mRNA recognition by RNA induced silencing complex. *Commun. Biol.* **4**, 1345 (2021).
46. A Montoya-Castillo, DR Reichman, Approximate but accurate quantum dynamics from the Mori formalism: I. Nonequilibrium dynamics. *J. Chem. Phys.* **144**, 184104 (2016).
47. MT Ama, H Mori, Statistical-Mechanical Theory of the Boltzmann Equation and Fluctuations in μ, Space. *Prog. Theor. Phys.* **56**, 1073 (1976).
48. S Chaturvedil, F Shibata, Time-Convolutionless Projection Operator Formalism for Elimination of Fast Variables. Applications to Brownian Motion. *Z. Physik B* **35**, 297 (1979).
49. HP Breuer, F Petruccione, *The Theory of Open Quantum Systems*. (Oxford University Press), pp. 444–447 (1985).
50. DE Shaw, et al., Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* **330**, 341 (2010).
51. E Elkayam, et al., The structure of human argonaute-2 in complex with miR-20a. *Cell* **150**, 100 (2012).
52. MP Allen, DJ Tildesley, *Computer Simulation of Liquids*. (Oxford University Press, New York), 2 edition, (2017).
53. A Fetter L., J Walecka D., *Quantum Theory of Many-Particle Systems*. (McGraw-Hill), pp. 53–56 (1971).
54. F Liu, et al., An experimental survey of the transition between two-state and downhill protein folding scenarios. *PNAS* **105**, 2369 (2007).
55. J Kappler, JO Daldrop, FN Brünig, MD Boehle, RR Netz, Memory-induced acceleration and slowdown of barrier crossing. *J. Chem. Phys.* **148** (2018).
56. S Cao, Y Qiu, M Kalin, X Huang, Integrative Generalized Master Equation: A Theory to Study Long-timescale Biomolecular Dynamics via the Integrals of Memory Kernels. *ChemRxiv 10.26434/chemrxiv-2022-0n9ld* (2022).
57. Y Naritomi, S Fuchigami, Slow dynamics of a protein backbone in molecular dynamics simulation revealed by time-structure based independent component analysis. *J. Chem. Phys.* **139**, 215102 (2013).
58. W Wang, T Liang, FK Sheong, X Fan, X Huang, An efficient Bayesian kinetic lumping algorithm to identify metastable conformational states via Gibbs sampling. *J. Chem. Phys.* **149**, 072337 (2018).
59. DS Hochbaum, DB Shmoys, A best possible heuristic for the k-center problem. *Math. operations research* **10**, 180 (1985).
60. JH Peng, W Wang, YQ Yu, HL Gu, X Huang, Clustering algorithms to analyze molecular dynamics simulation trajectories for complex chemical and biological systems. *Chin. J. Chem. Phys.* **31**, 404 (2018).
61. P Deuflhard, M Weber, Robust Perron cluster analysis in conformation dynamics. *Linear Algebr. Its Appl.* **398**, 161 (2005).
62. S Röblitz, M Weber, Fuzzy spectral clustering by PCCA+: Application to Markov state models and data classification. *Adv. Data Analysis Classif.* **7**, 147 (2013).
63. CR Schwantes, VS Pande, Improvements in Markov State Model construction reveal many non-native interactions in the folding of NTL9. *J. Chem. Theory Comput.* **9**, 2000 (2013).
64. F Litzinger, et al., Rapid Calculation of Molecular Kinetics Using Compressed Sensing. *J. Chem. Theory Comput.* **14**, 2771 (2018).
65. S Liu, L Zhu, FK Sheong, W Wang, X Huang, Adaptive partitioning by local density-peaks: An efficient density-based clustering algorithm for analyzing molecular dynamics trajectories. *J. Comput. Chem.* **38**, 152 (2017).
66. WC Pfalzgraff, A Montoya-Castillo, A Kelly, TE Markland, Efficient construction of generalized master equation memory kernels for multi-state systems from nonadiabatic quantum-classical dynamics. *J. Chem. Phys.* **150**, 244109 (2019).
67. CD Meyer, An Alternative Expression for the Mean First Passage Matrix. **22**, 41–47 (1978).
68. A Kells, Z Mihálka, A Annibale, E Rosta, Mean first passage times in variational coarse graining using Markov state models. *J. Chem. Phys.* **150**, 134107 (2019).
69. CH Jensen, D Nerukh, RC Glen, Calculating mean first passage times from Markov models of proteins. *AIP Conf. Proc.* **940**, 150 (2007).