Mechanism beyond Markov models: How and why to use non-Markovian analysis of trajectory data

Ernesto Suárez

2017

University of Pittsburgh
Department of Computational & Systems Biology

MSM in Science

Number of scientific publications per year

2016: 20,700

2015: 20,500

.

.

.

2000: 4,600

MSM in Computational Chemistry

JACS JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
Communication
pubs.acs.org/JACS

Emergence of Glass-like Behavior in Markov State Models of Protein Folding Dynamics

Jeffrey K. Weber,[†] Robert L. Jack,[‡] and Vijay S. Pande*[,†]

[†]Department of Chemistry, Stanford University, Stanford, California 94305, United States
[‡]Department of Physics, University

JACS JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
ARTICLE
pubs.acs.org/JACS

Markov State Model Reveals Folding and Functional Dynamics in Ultra-Long MD

Thomas J. Lane,[†] Greg

[†]Department of Chemistry a

CrossMark click for updates
PNAS PLUS
PNAS

Multiensemble Markov models of molecular thermodynamics and kinetics

Hao Wu[a,1], Fabian Paul[a,1], Christoph Wehmeyer[a], and Frank Noé[a,2]

[a]Department of Mathematics and Computer Science,

Edited by David Baker, University of Washington, Sea

JACS JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
Article
pubs.acs.org/JACS
Terms of Use

Complex RNA Folding Kinetics Revealed by Single-Molecule FRET and Hidden Markov Models

Bettina G. Keller,*[,†] Andrei Kobitski,[‡] Andres Jäschke,[§] G. Ulrich Nienhaus,[‡,∥] and Frank Noé*[,⊥]

[†]Freie Universität Berlin, Institute of Chemistry and Biochemistry, Takustr. 3, 14195 Berlin, Germany
[‡]Institute of Applied Physics, Center for Functional Nanostructures, and Institute of Toxicology and Genetics, Karlsruhe Institute of Technology (KIT), Wolfgang-Gaede-Str. 1, 76131 Karlsruhe, Germany
[§]Institute of Pharmacy and Molecular Biotechnology, Heidelberg University, 69120 Heidelberg, Germany
[∥]Department of Physics, University of Illinois at Urbana−Champaign, Urbana, Illinois 61801, United States
[⊥]Freie Universität Berlin, Institute of Mathematics, Arnimallee 6, 14195 Berlin, Germany

Markov State Models (MSM)

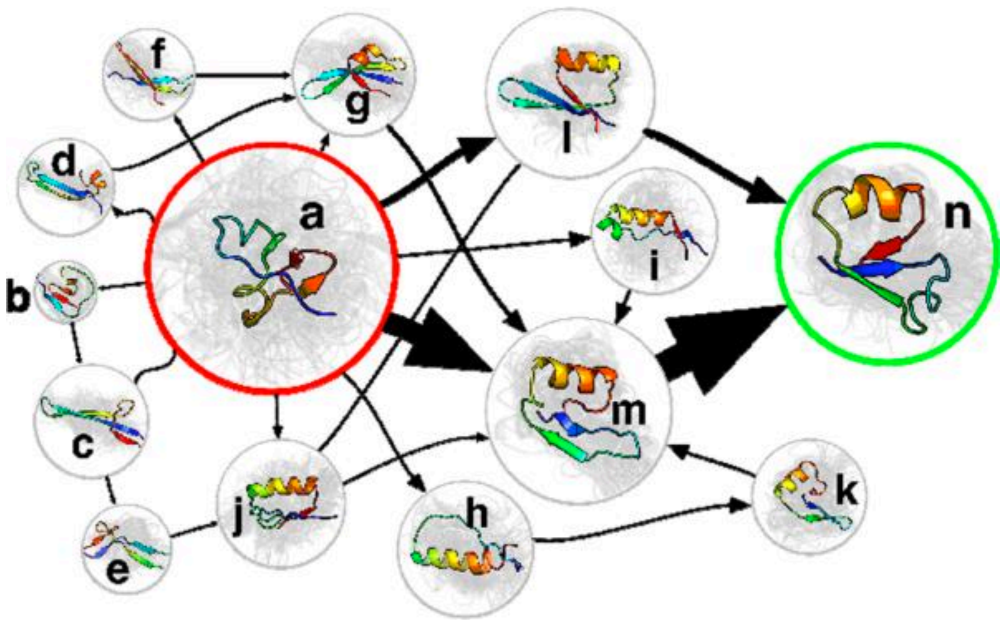
Widely used to analyze and interpret molecular trajectories. The final goal is to infer long time behavior.

Main assumption: The Markov property

$$k_{ij}(\tau) = P\{X_{t+\tau} = j \mid X_t = i\}$$

Regular simulations are Markovian in their full continuous phase spaces. However any discrete partition of the phase space generates non-Markovian trajectories.

Learning process in MSM



$$\begin{pmatrix} \hat{k}_{11} & \hat{k}_{12} & \hat{k}_{13} & \cdots \\ \hat{k}_{21} & \hat{k}_{22} & \hat{k}_{23} & \\ \hat{k}_{31} & \hat{k}_{32} & \hat{k}_{33} & \\ \vdots & & & \ddots \end{pmatrix}$$

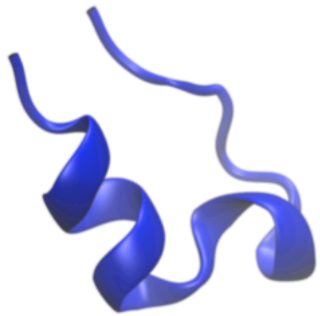
$$\hat{k}_{ij}(\tau) = c_{ij} / c_i$$

$$\mathbf{K}^T \mathbf{p} = \mathbf{p}$$

Biased for kinetics

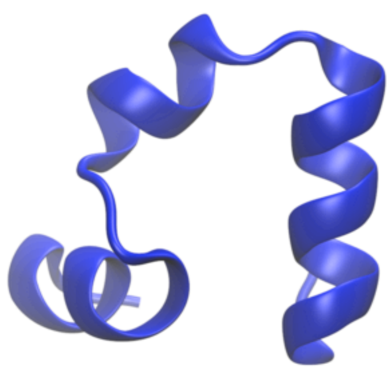
Voelz et al., J. Am. Chem. Soc., 2010, 132(5), pp 1526-1528

Protein Models

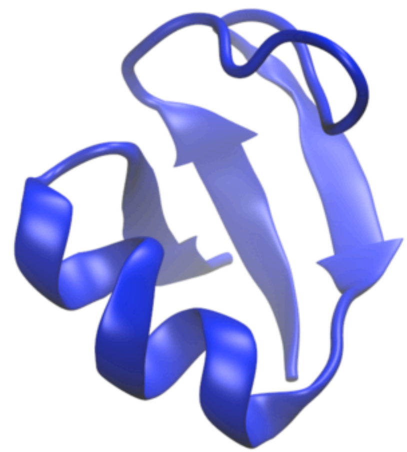


Trp-cage 208μs

Chignolin 106μs



Villin 125μs



NTL9 1100μs

Shaw et al., Science 2011, 334(6055), pp. 517-520

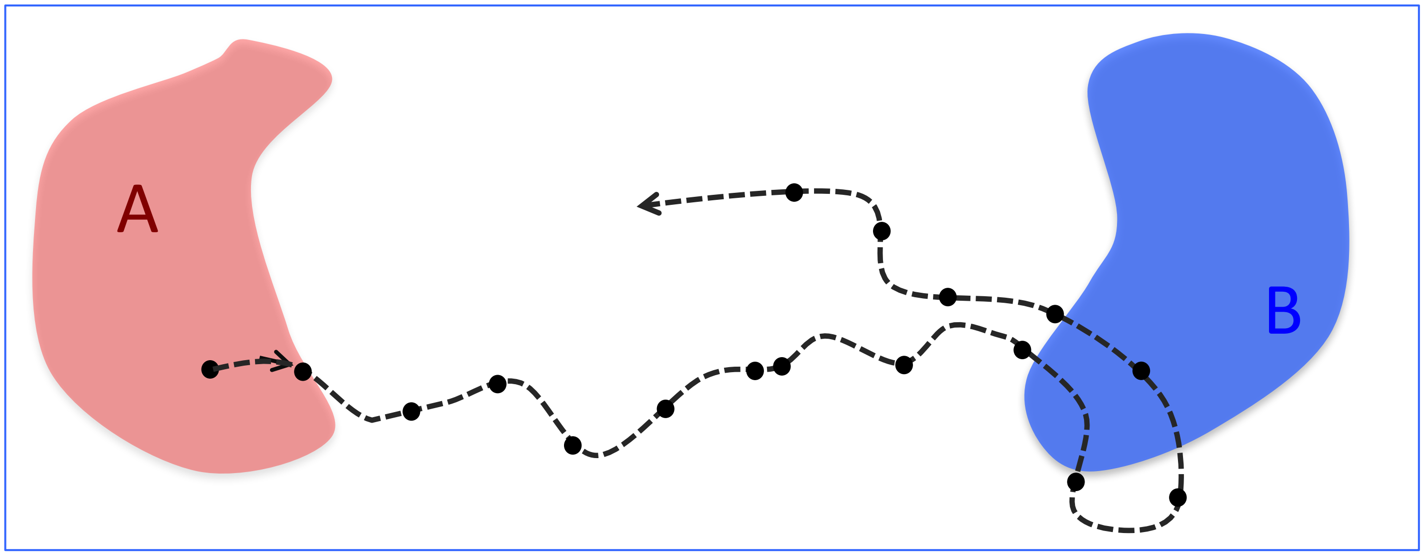
MSM Analysis: Standard Recipe

1. Divide the space in “Markovian” regions
2. Estimate parameters and select a lag time
3. Analysis

Estimating kinetic properties

Mean First Passage Time (MFPT)

Mean First Passage Time (MFPT)

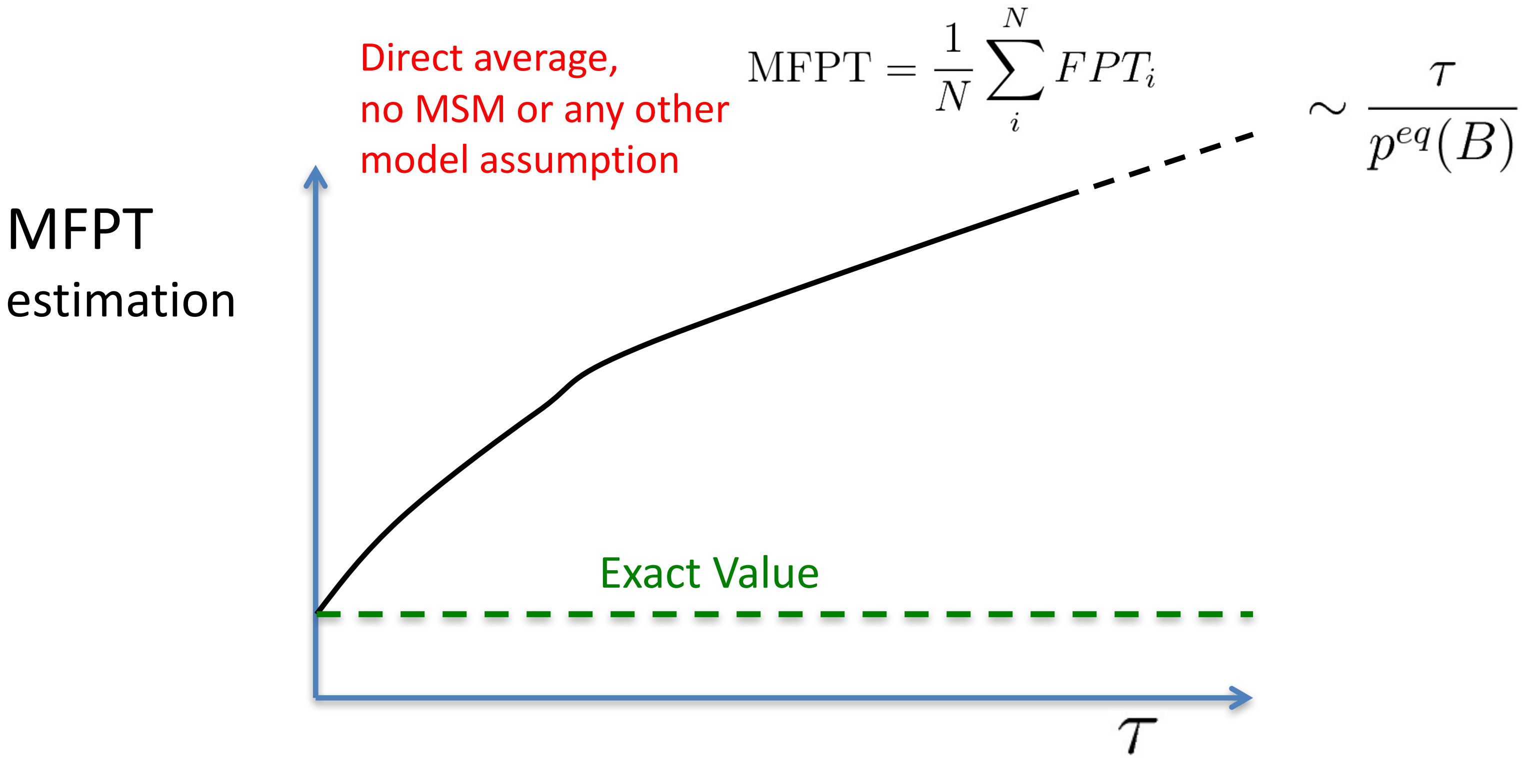


τ = The lag-time >> integration time step δt

$$\mathrm{MFPT} = \frac{1}{k_{AB}}$$

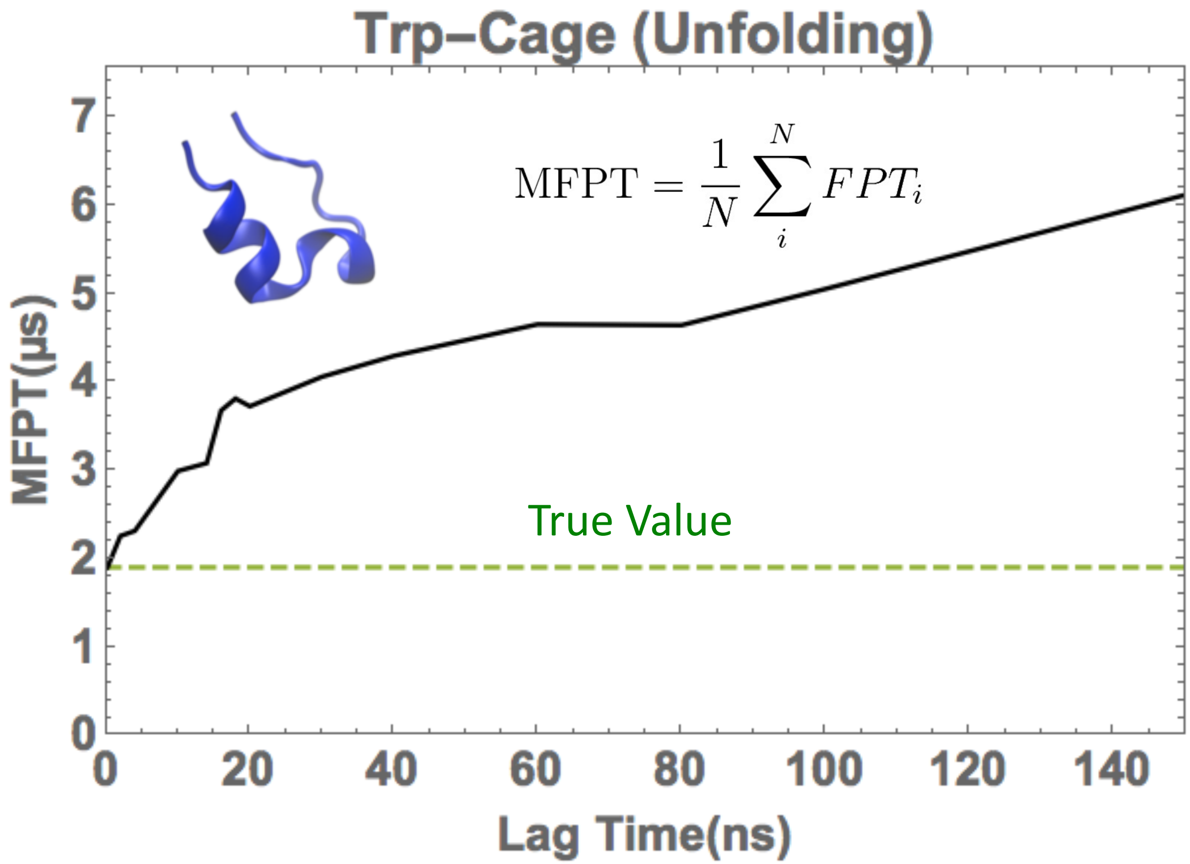
NOT MSM

Mean First Passage Time (MFPT)



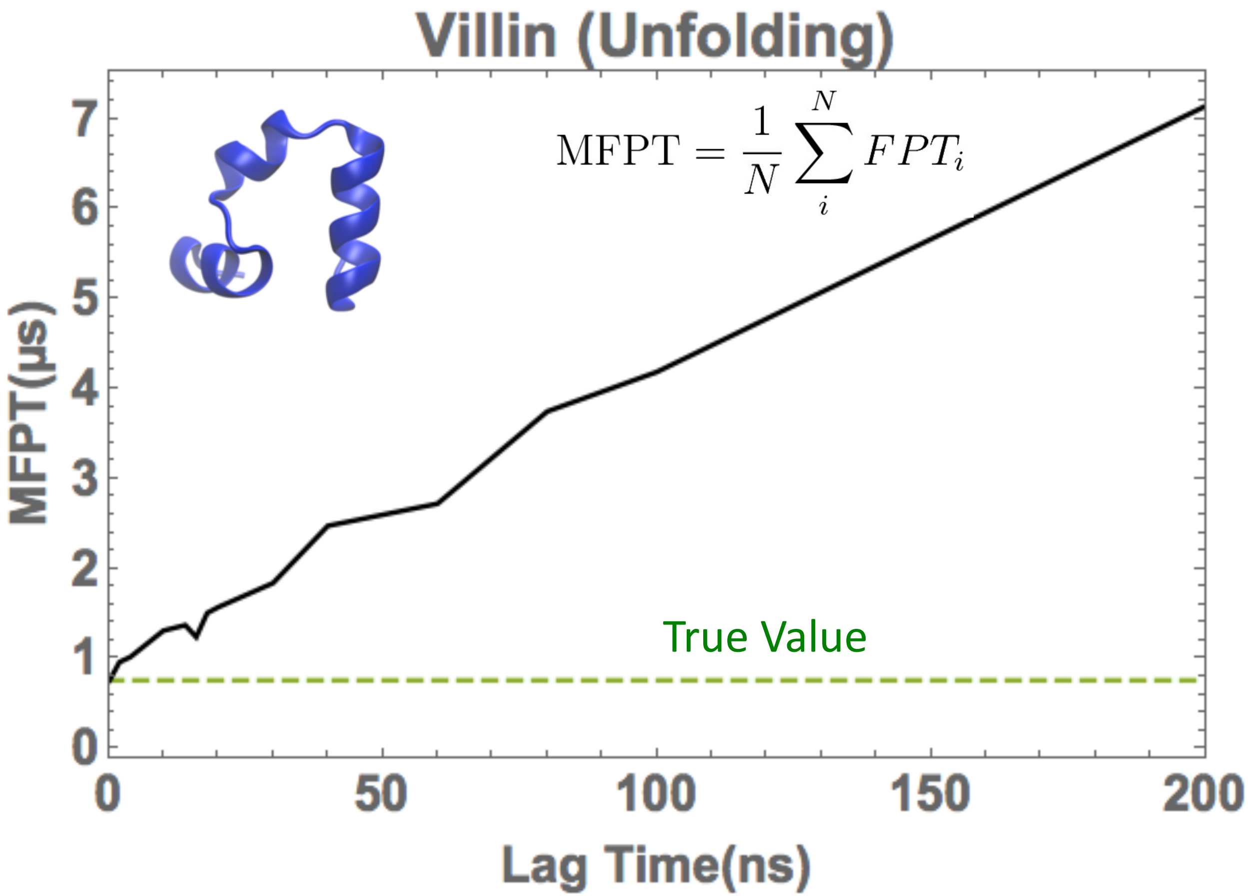
Our estimation of the kinetic properties are lag-time dependent while the thermodynamic properties are the same for every lag-time.

MFPT vs lag-time



No MSM or any other model assumption

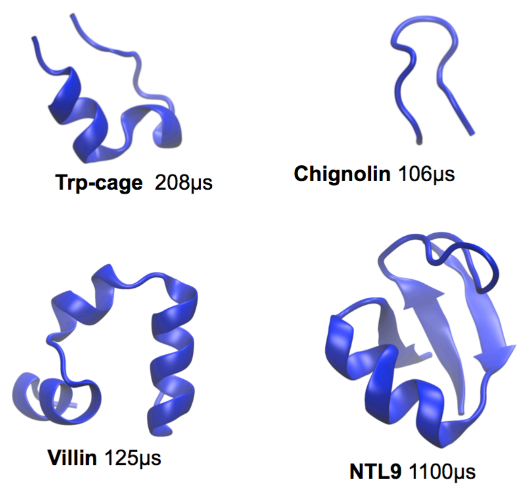
MFPT vs lag-time



No MSM or any other model assumption

Data choices

Long MD Simulations



Full data set

- Markovian
- Non-Markovian

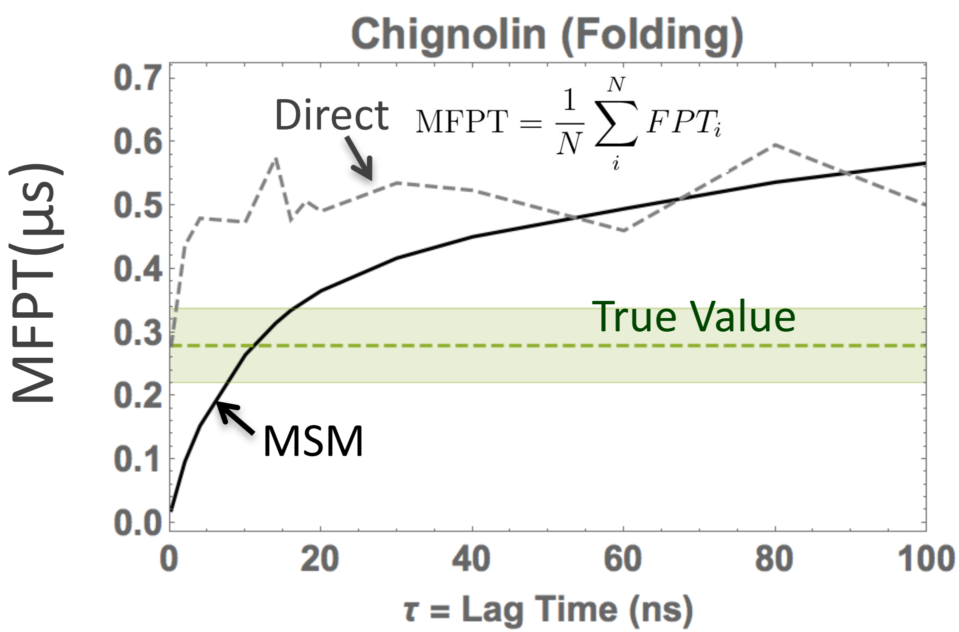
Reduced data set (< 5%)

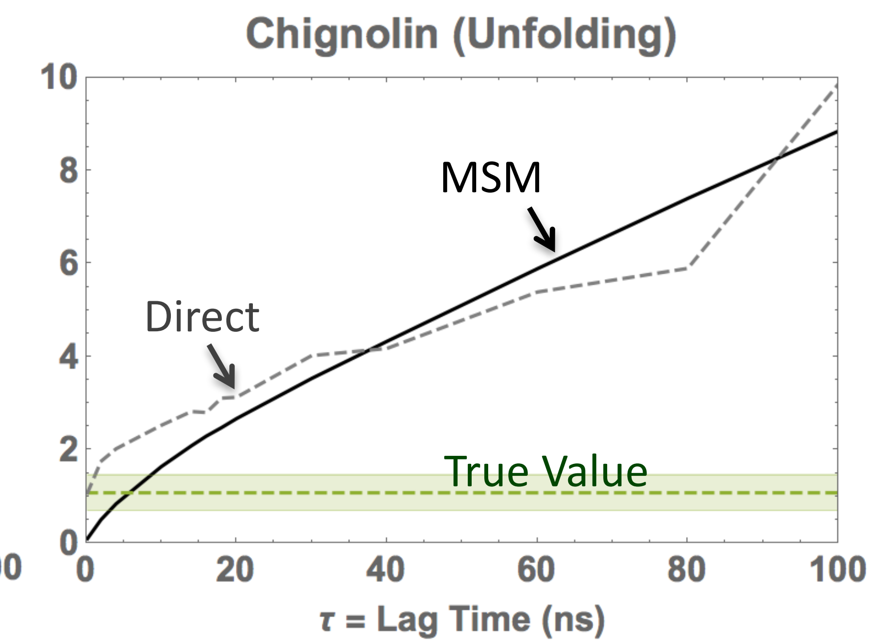
- Non-Markovian

Markov State Models

Markov MFPT vs lag-time

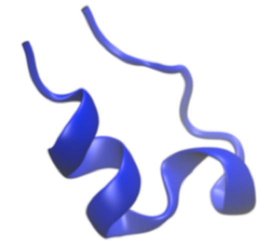
Chignolin

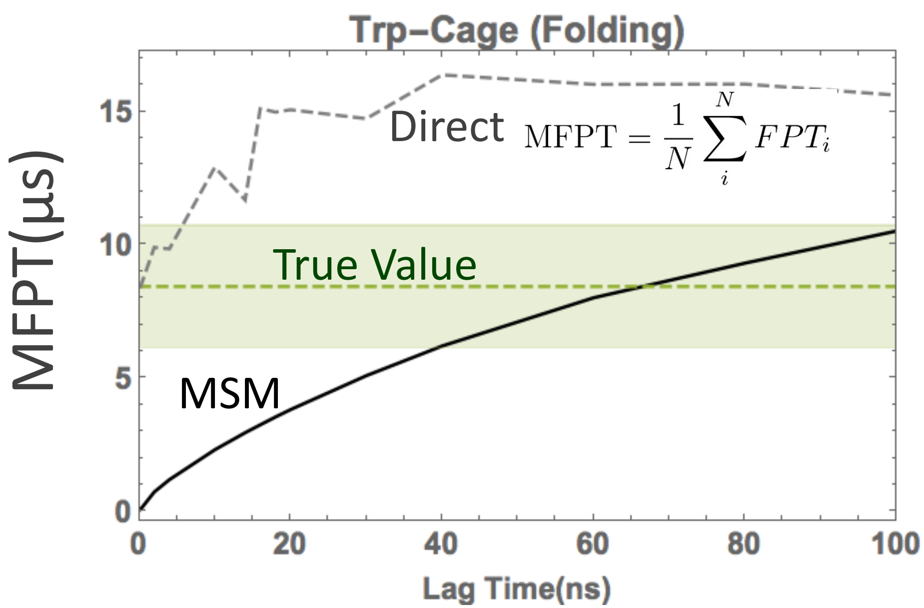


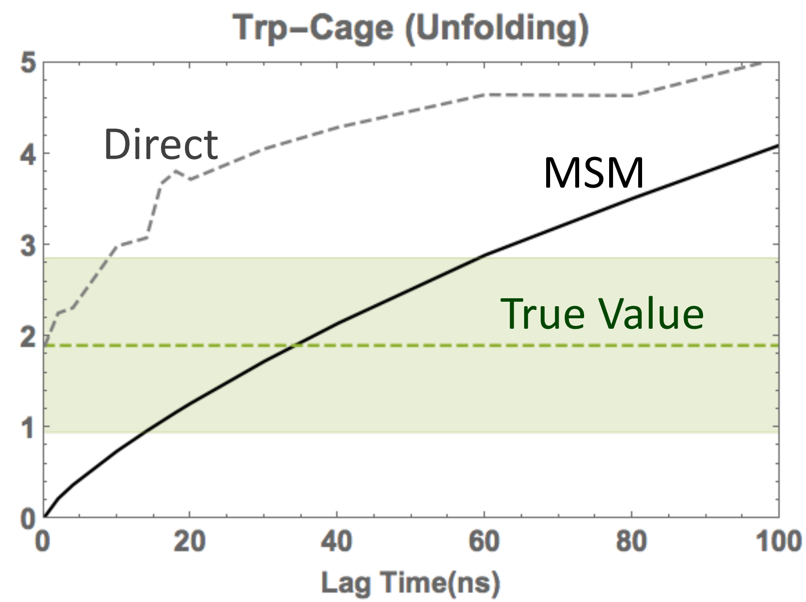


Markov MFPT vs lag-time

Trp-cage

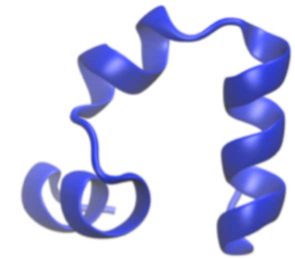


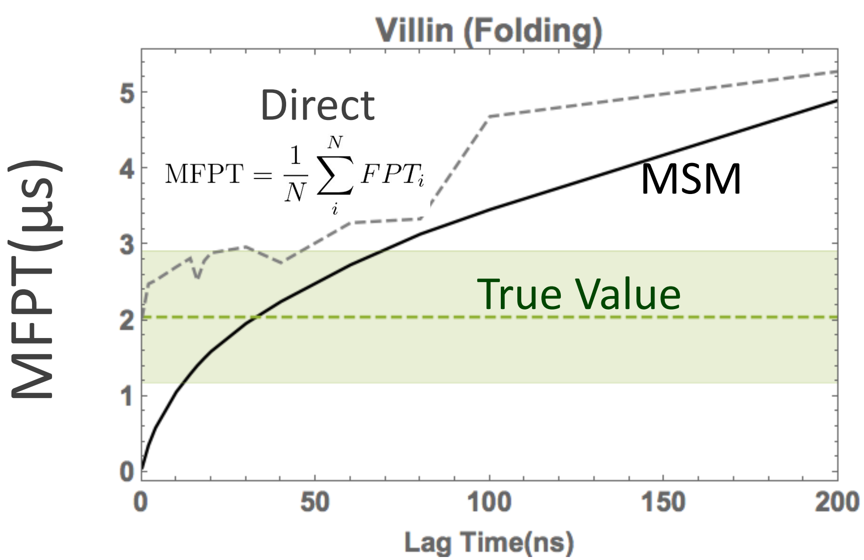


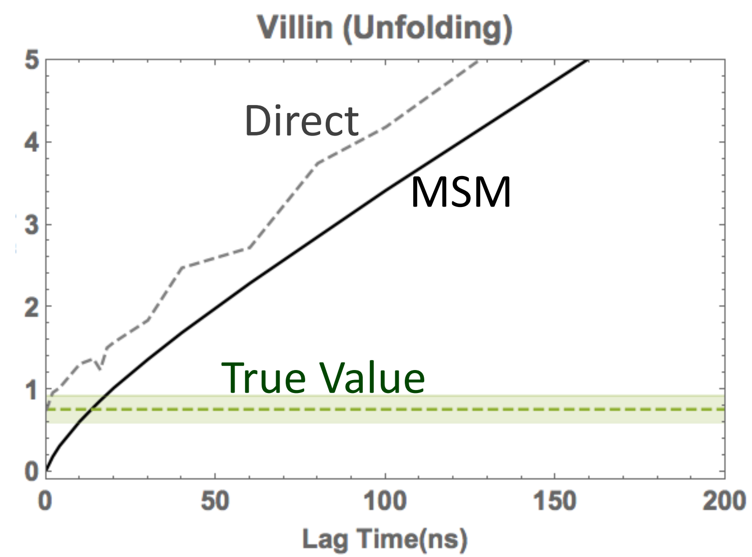


Markov MFPT vs lag-time

Villin

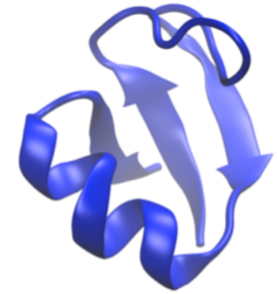


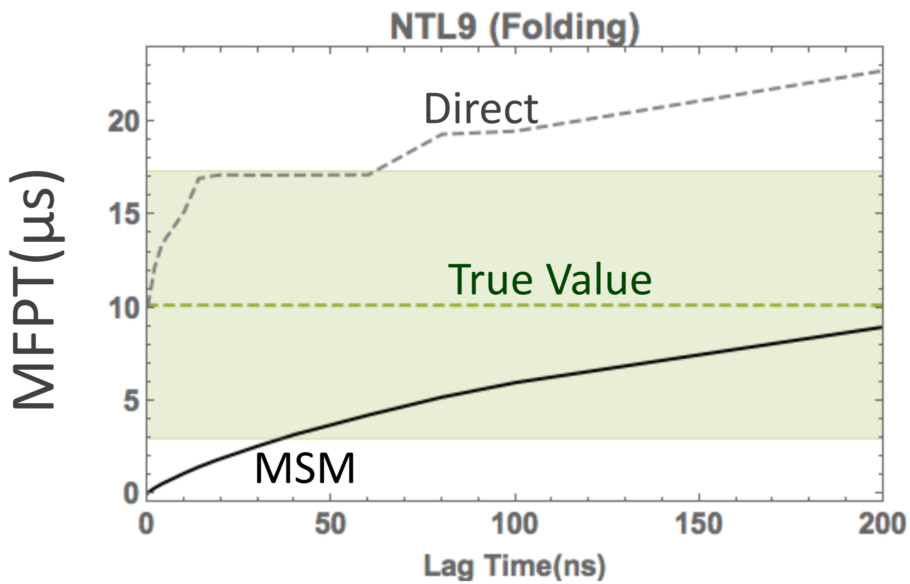


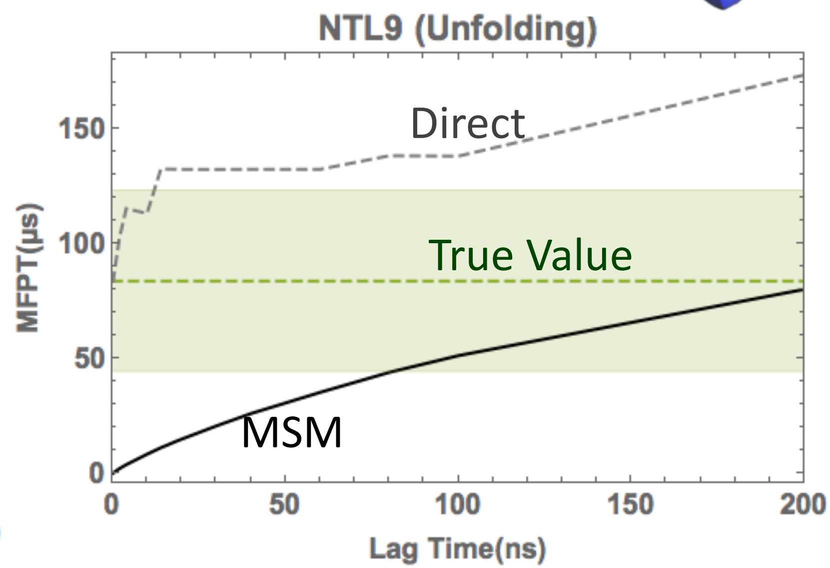


Markov MFPT vs lag-time

NTL9

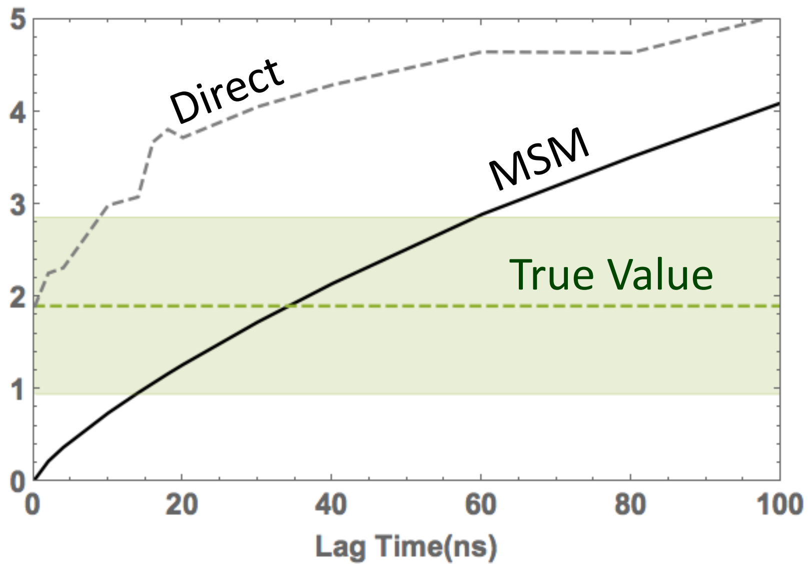






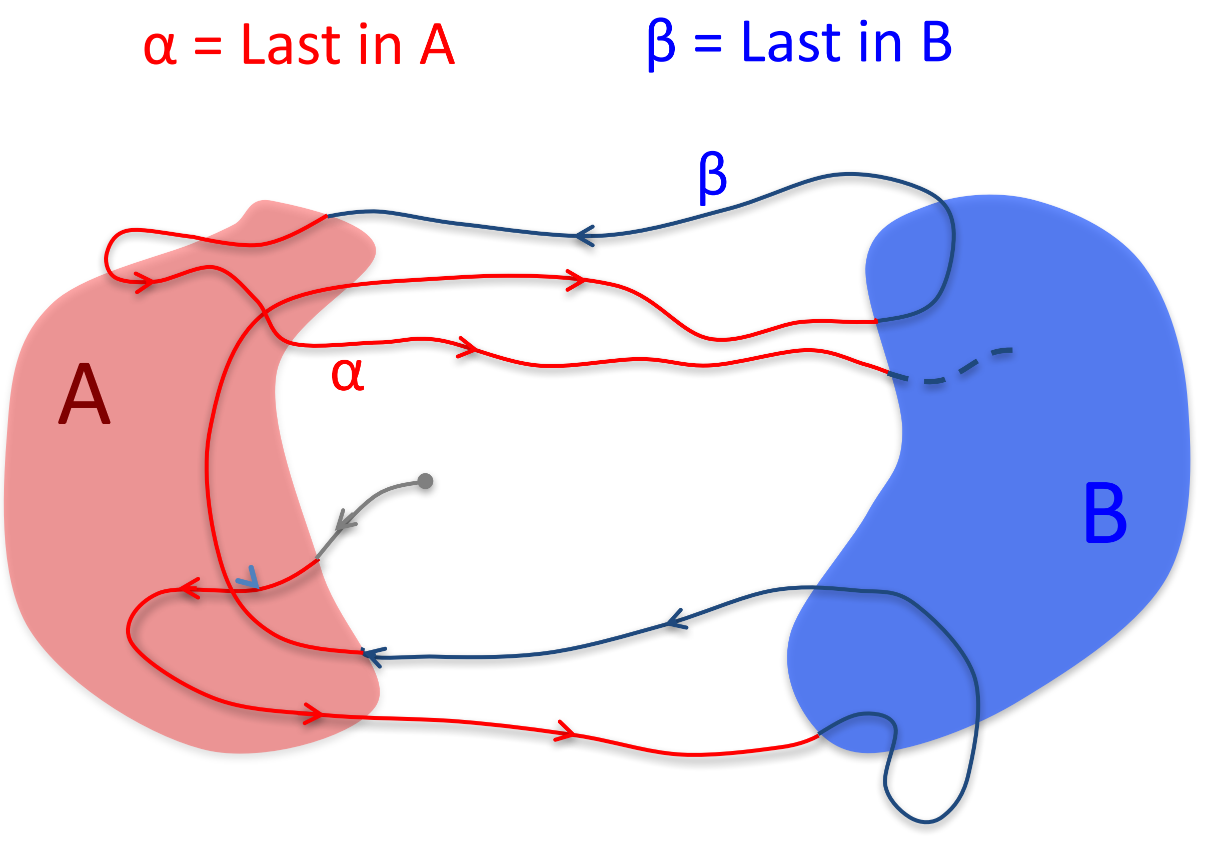
MSM Analysis

- Biased for kinetic properties
  - Discretization error ↑MFTP
  - Markov error ↓MFPT



Non-Markovian Analysis

Beyond Markov: Color



Suarez et al., J. Chem. Theory Comput., 2014, 10 (7), pp 2658–2667
Vanden-Eijnden et al., J. Chem. Phys., 2009, 131(4), pp 44120

Beyond Markov: Color

$$k_{ij}(\tau) = P\{X_{t+\tau} = j | X_t = i\}$$

$$\Downarrow$$

$$k_{ij}^{\mu\nu}(\tau) = P\{X_{t+\tau} = j, L_{t+\tau} = \nu | X_t = i, L_{t+\tau} = \mu\} \quad \mu, \nu = \alpha, \beta$$

Suarez et al., J. Chem. Theory Comput., 2014, 10 (7), pp 2658–2667
Vanden-Eijnden et al., J. Chem. Phys., 2009, 131(4), pp 44120

Beyond Markov: Color

Biased kinetics

N x N

$$\left[\begin{array}{c:c:c} k_{11} & k_{12} & k_{13} \\ \hdashline k_{21} & k_{22} & k_{23} \\ \hdashline k_{31} & k_{32} & k_{33} \end{array}\right] \Longrightarrow$$

Unbiased kinetics

2N x 2N

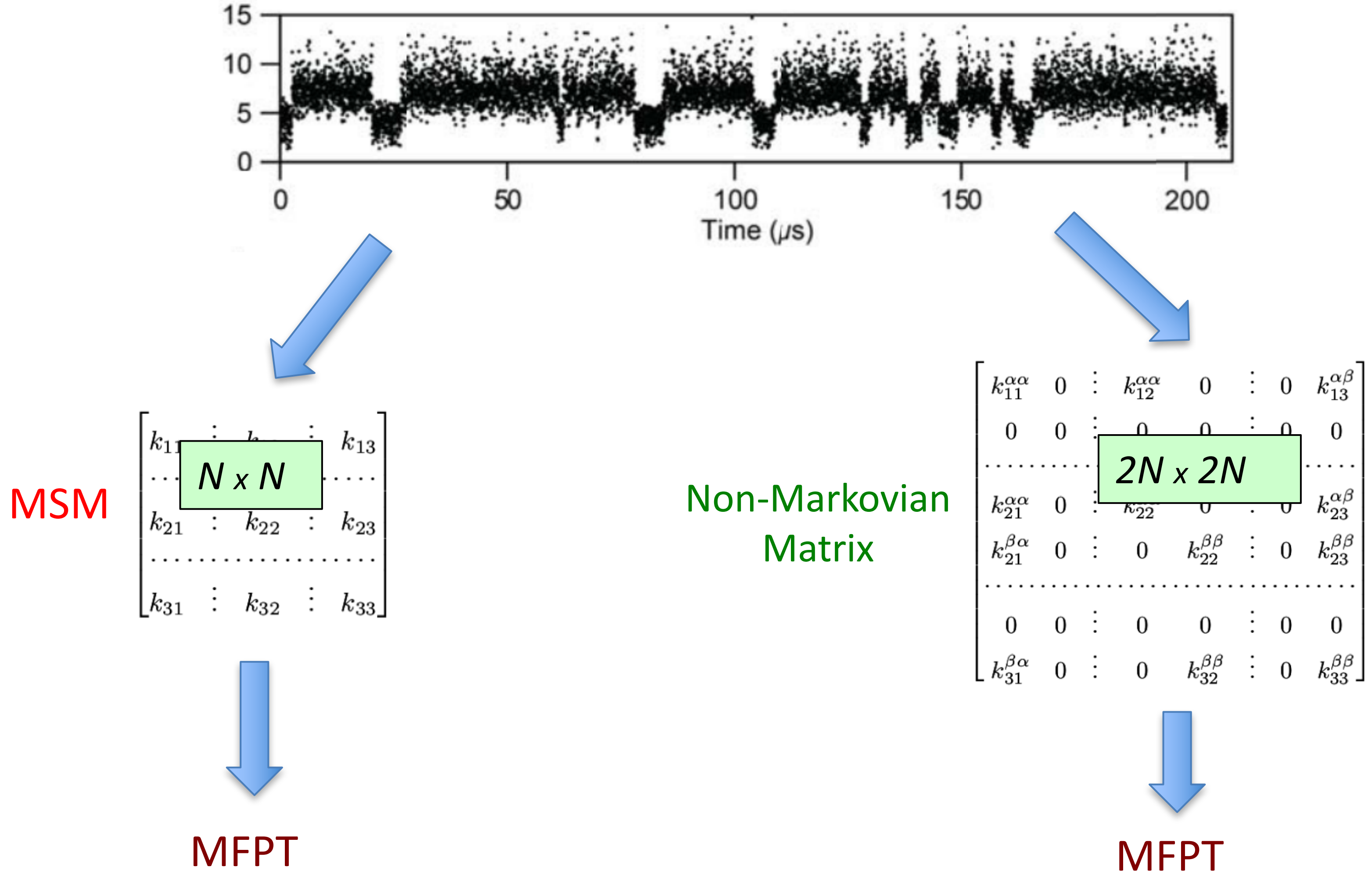
$$\left[\begin{array}{cc:cc:cc} k_{11}^{\alpha\alpha} & 0 & k_{12}^{\alpha\alpha} & 0 & 0 & k_{13}^{\alpha\beta} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hdashline k_{21}^{\alpha\alpha} & 0 & k_{22}^{\alpha\alpha} & 0 & 0 & k_{23}^{\alpha\beta} \\ k_{21}^{\beta\alpha} & 0 & 0 & k_{22}^{\beta\beta} & 0 & k_{23}^{\beta\beta} \\ \hdashline 0 & 0 & 0 & 0 & 0 & 0 \\ k_{31}^{\beta\alpha} & 0 & 0 & k_{32}^{\beta\beta} & 0 & k_{33}^{\beta\beta} \end{array}\right]$$

Example with 3 bins. *A* is defined as bin 1 and *B* as bin 2

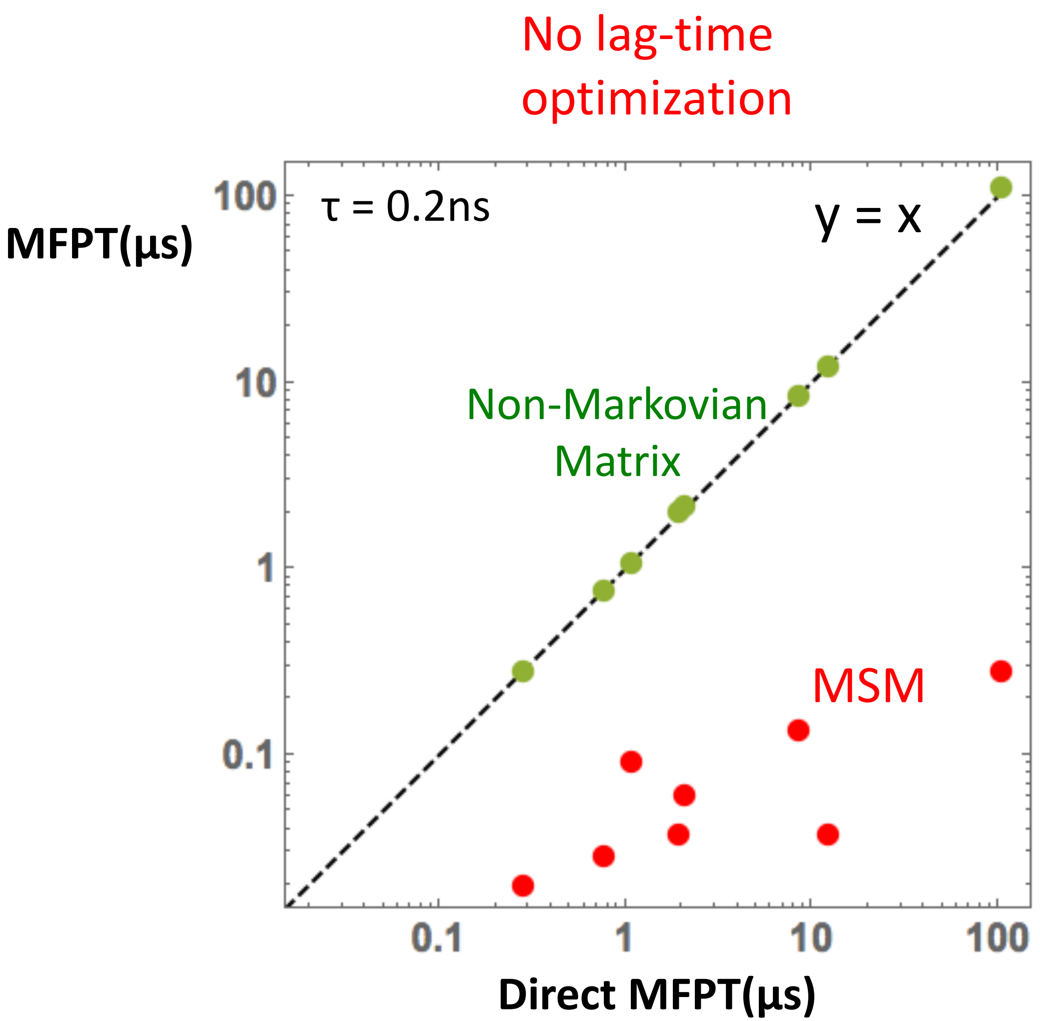
$$\mathcal{K}^T \mathbf{p}^\mu = \mathbf{p}^\mu \qquad p_i^{\mathrm{eq}} = p_i^\alpha + p_i^\beta$$

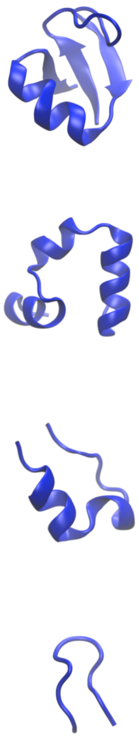
Suárez et al., J. Chem. Theory Comput., 2014, 10 (7), pp 2658–2667

MSM vs Non-Markovian Analysis



MSM vs Non-Markovian Analysis





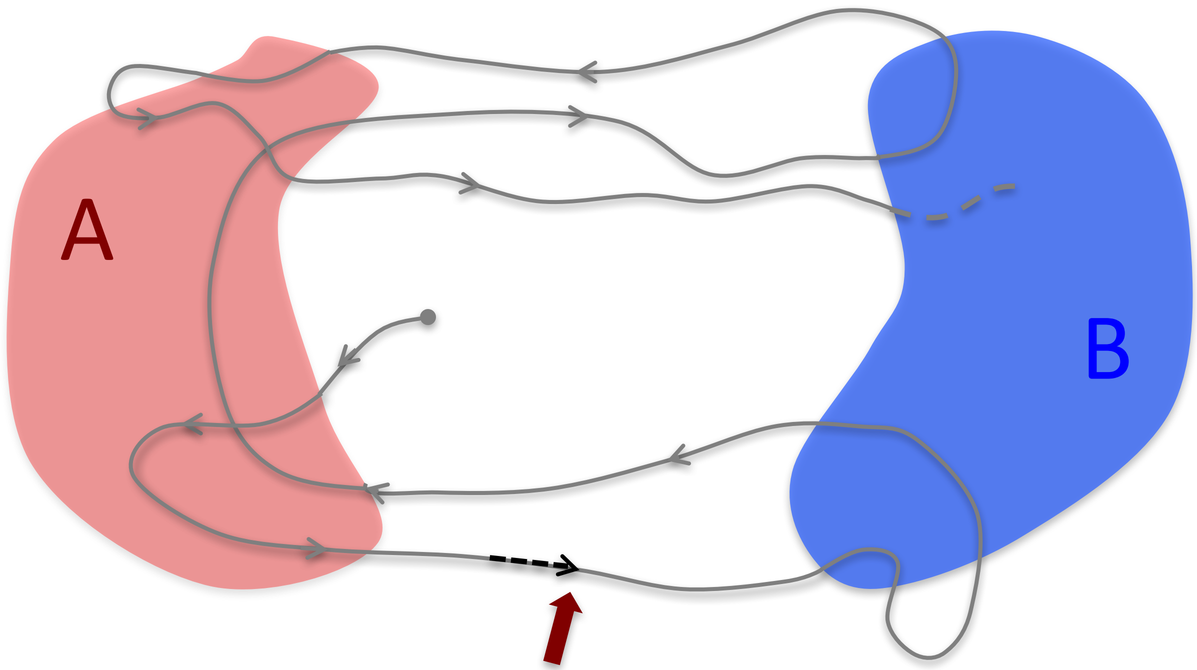
Non-Markovian Analysis

With sufficient history (color) information we get

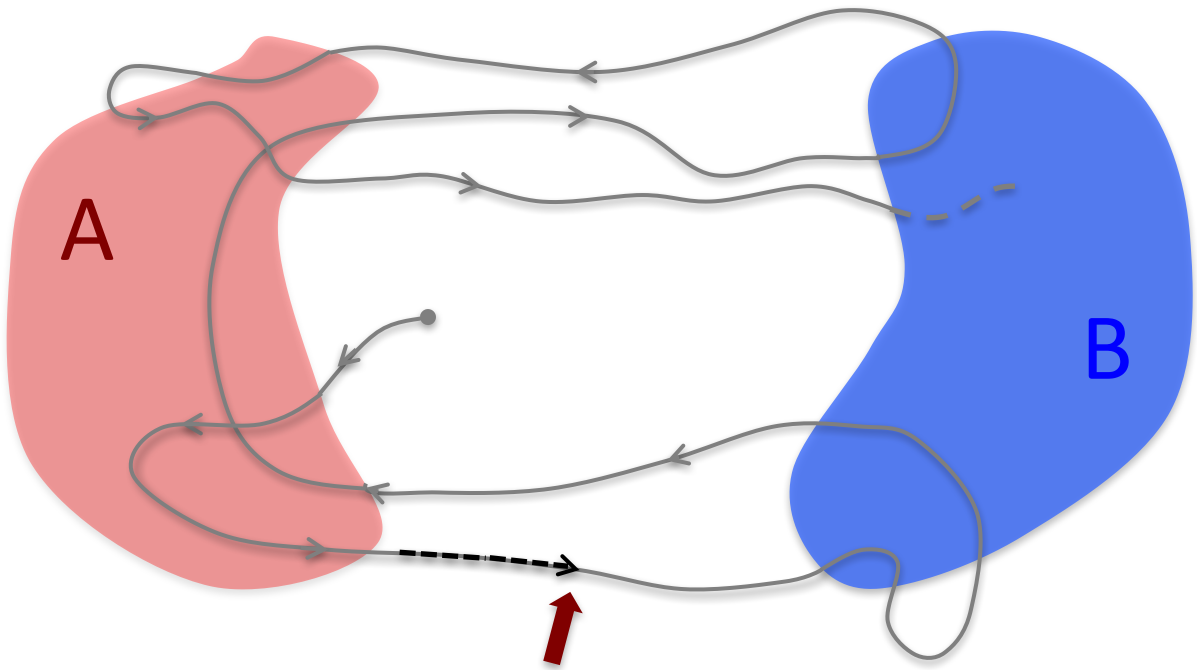
- Unbiased thermodynamics (populations)
- Unbiased MFPT ($\tau \longrightarrow 0$)

Suarez et al., J. Chem. Theory Comput., 2014, 10 (7), pp 2658–2667

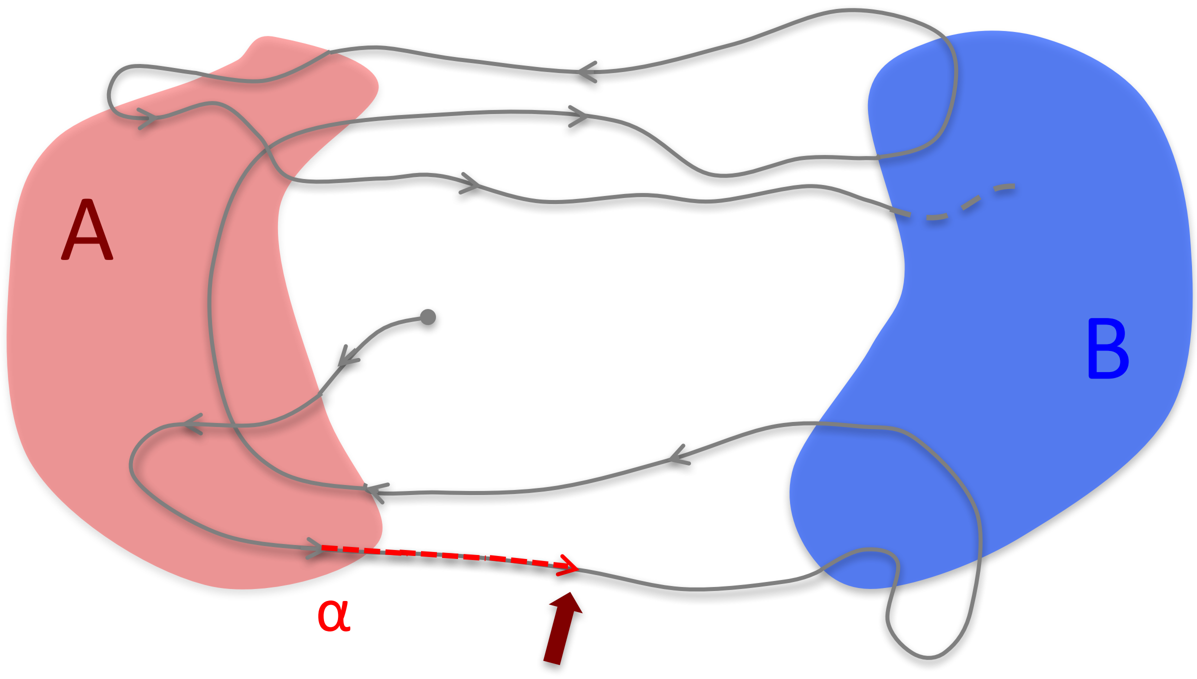
Limited color/history info



Limited color/history info



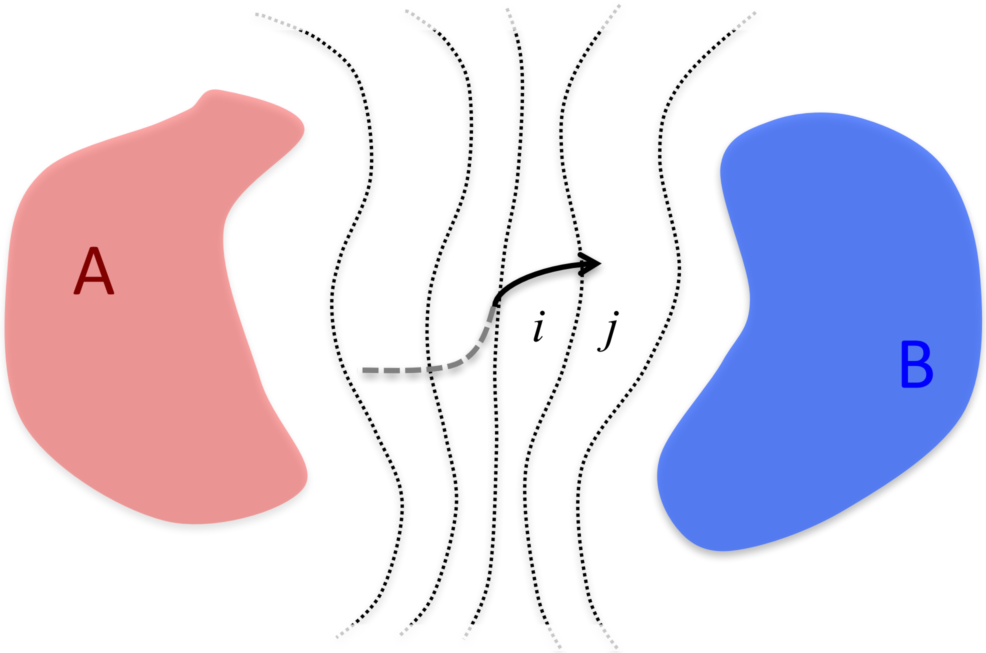
Limited color/history info



Other non-Markovian Analyses

Markov + Color

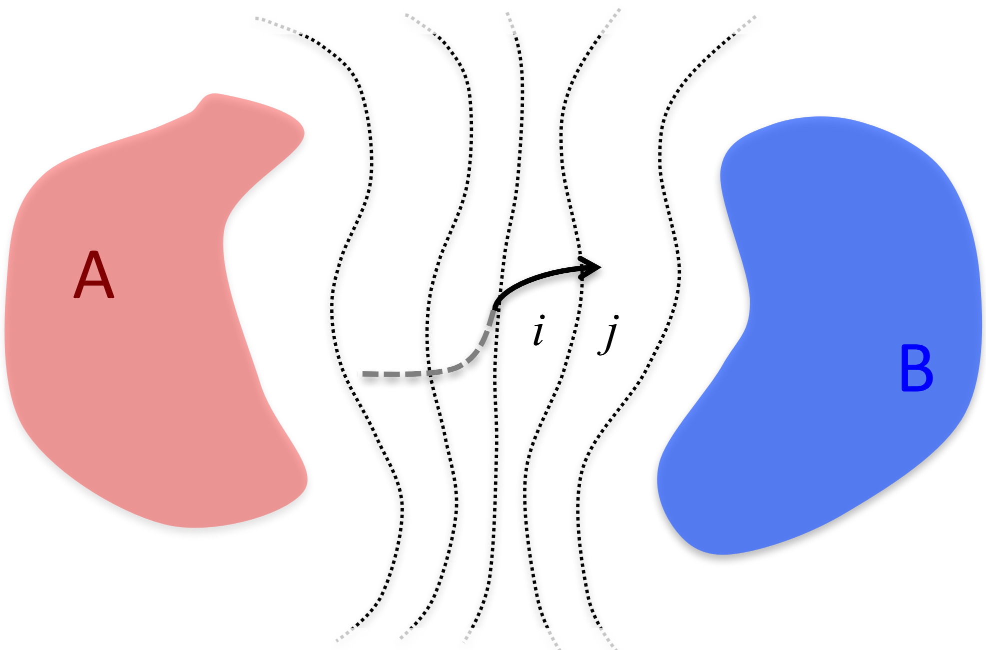
When examining a given time point of the trajectory for estimating a labeled rate, the α or β label are assigned if possible given the amount of history. Otherwise the label is assigned stochastically assuming a Markov behavior.



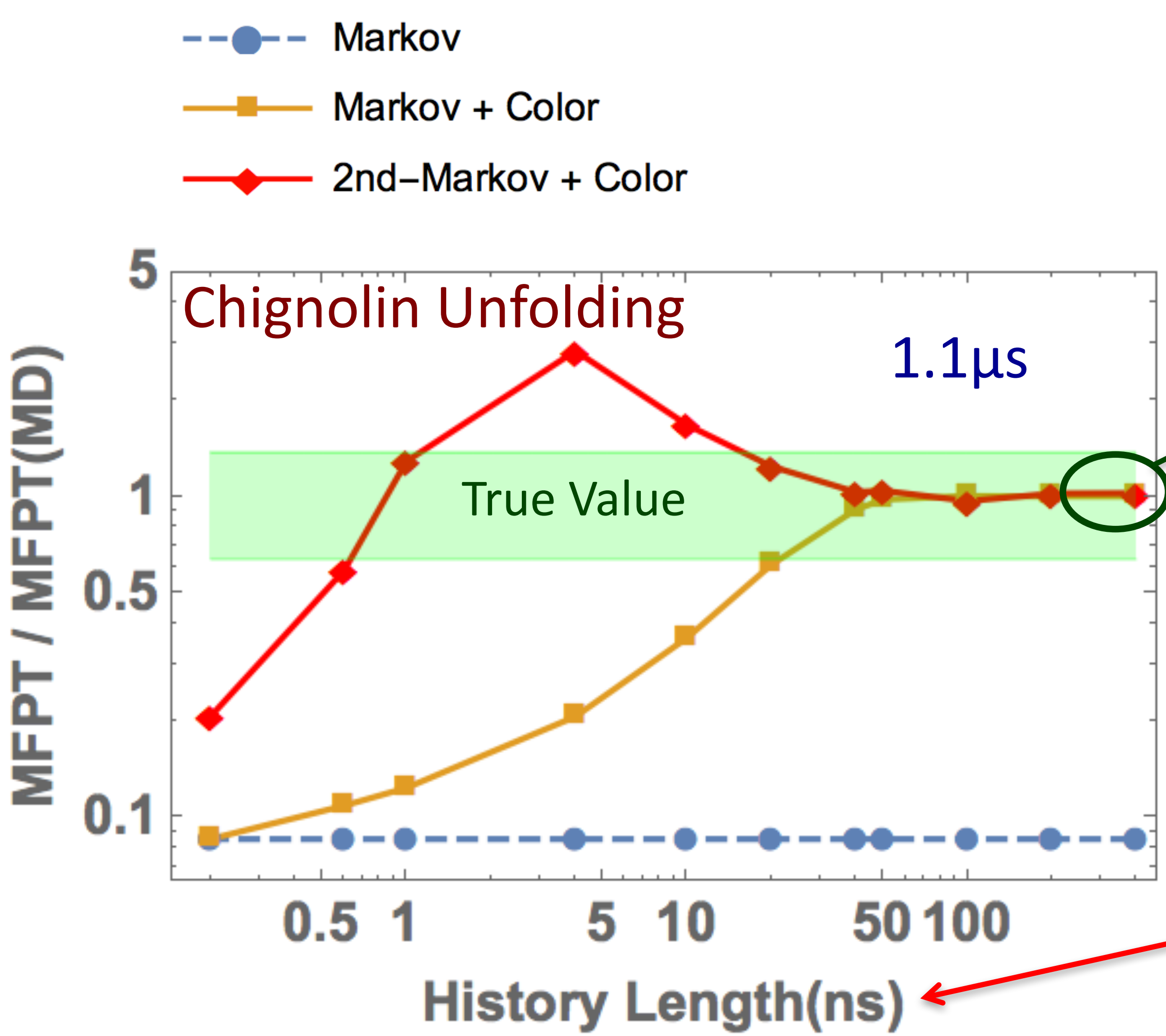
Other non-Markovian Analyses

2nd Order Markov + Color

When examining a given time point of the trajectory for estimating a labeled rate, the α or β label are assigned if possible given the amount of history. Otherwise the label is assigned stochastically assuming a 2nd-Order Markov model.



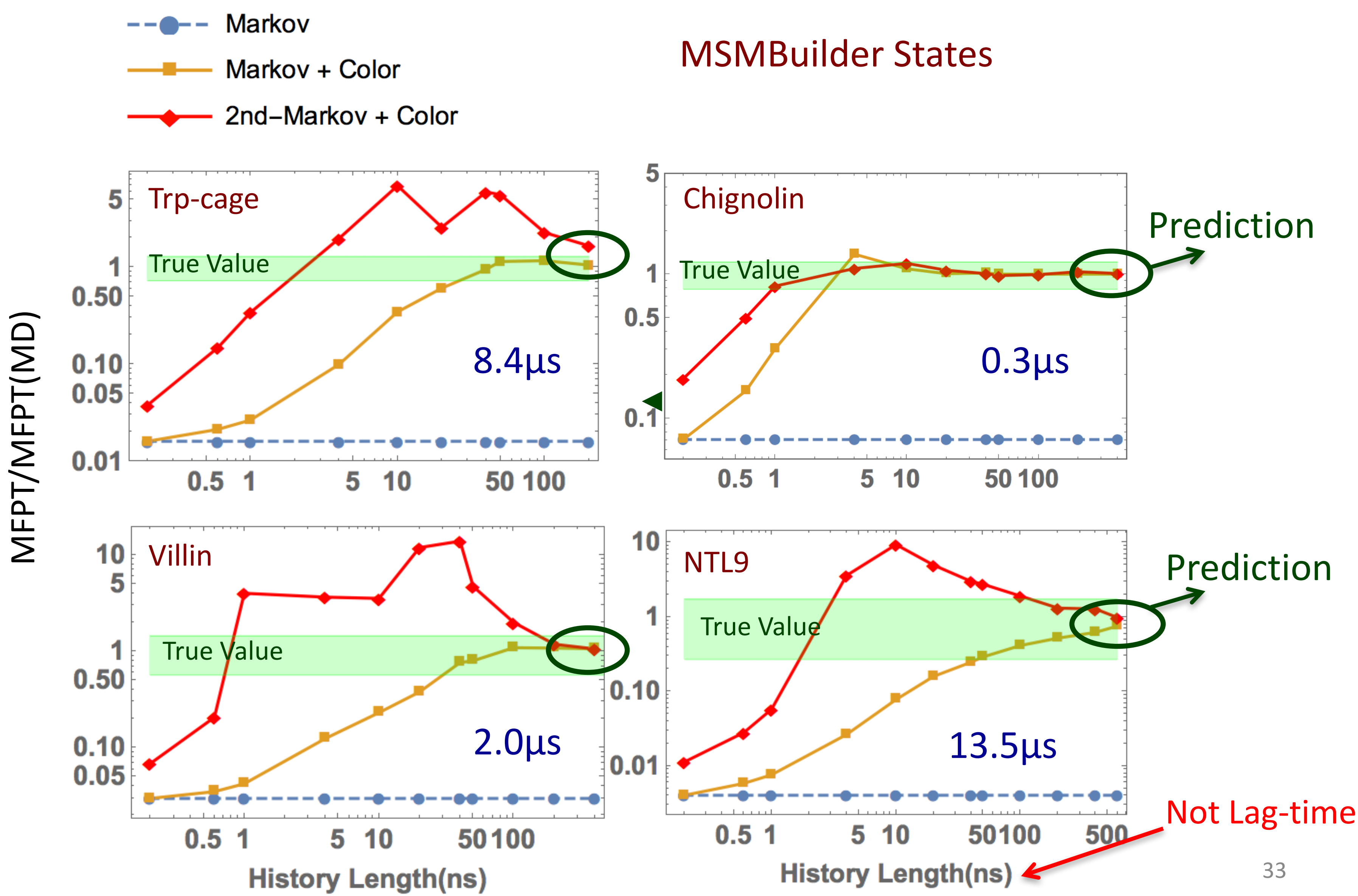
Non-Markovian Analyses



Our prediction

Not Lag-time

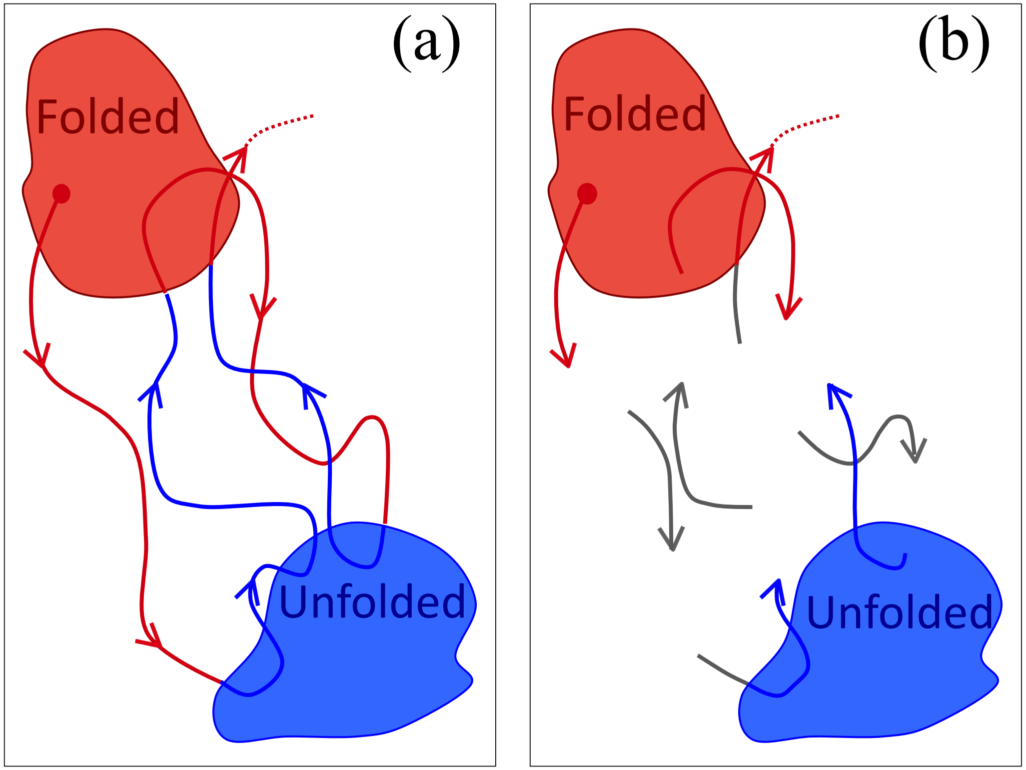
Non-Markovian Analyses (Folding)



Reduced data set:

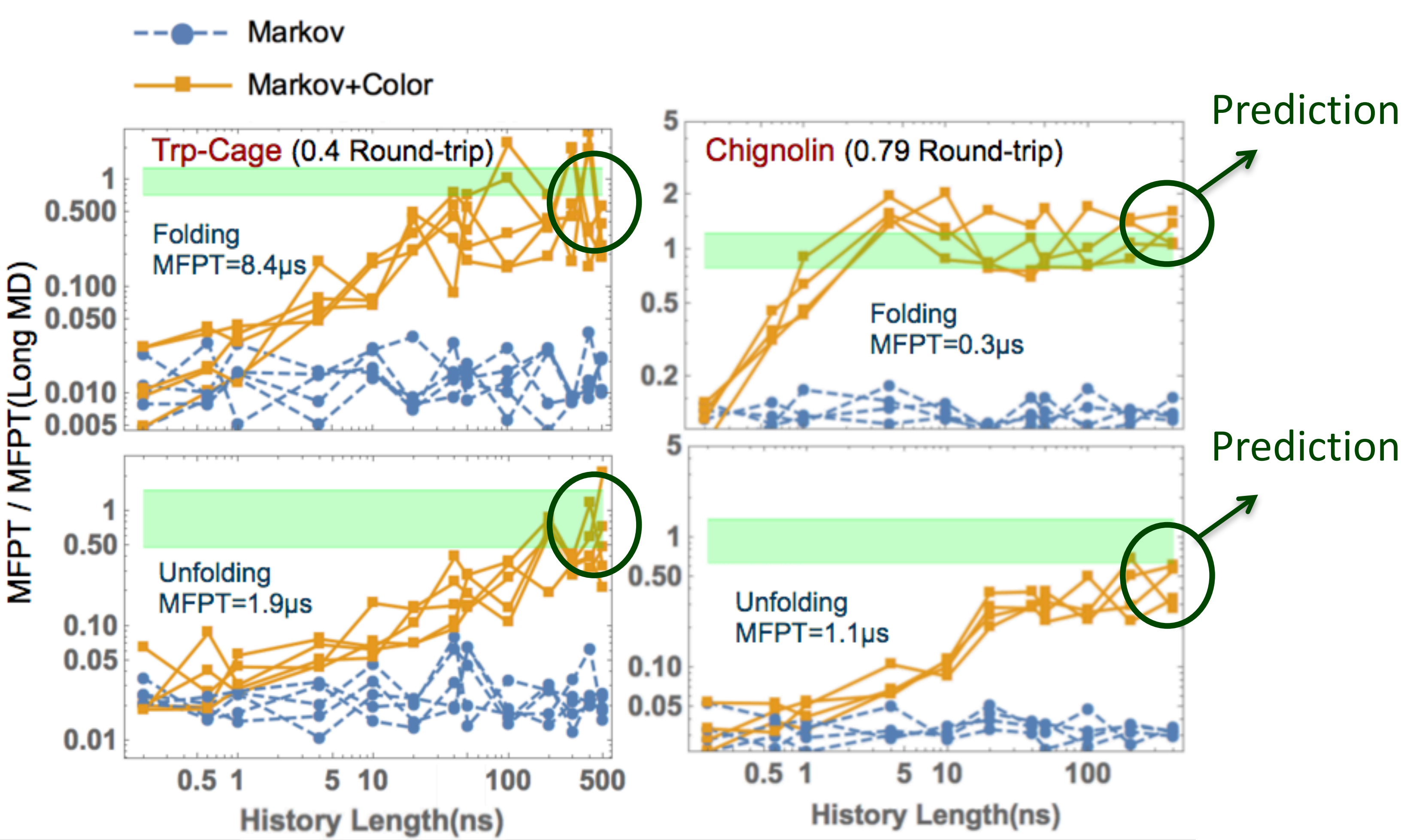
Non-Markovian Analysis

Reducing the amount of Data (< 5%)



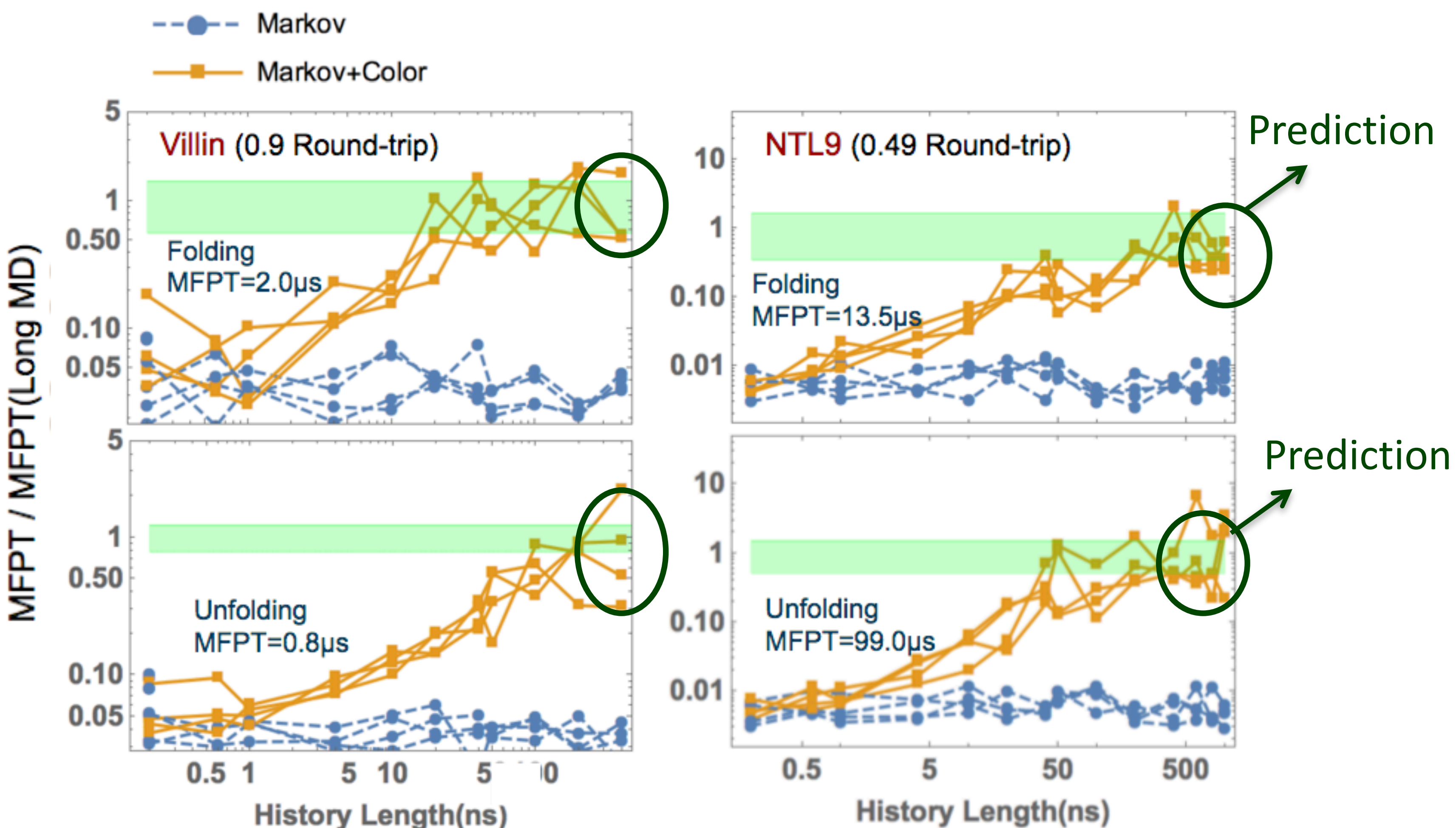
Non-Markovian Analyses

Reduced data, MSMBuilder States



Non-Markovian Analyses

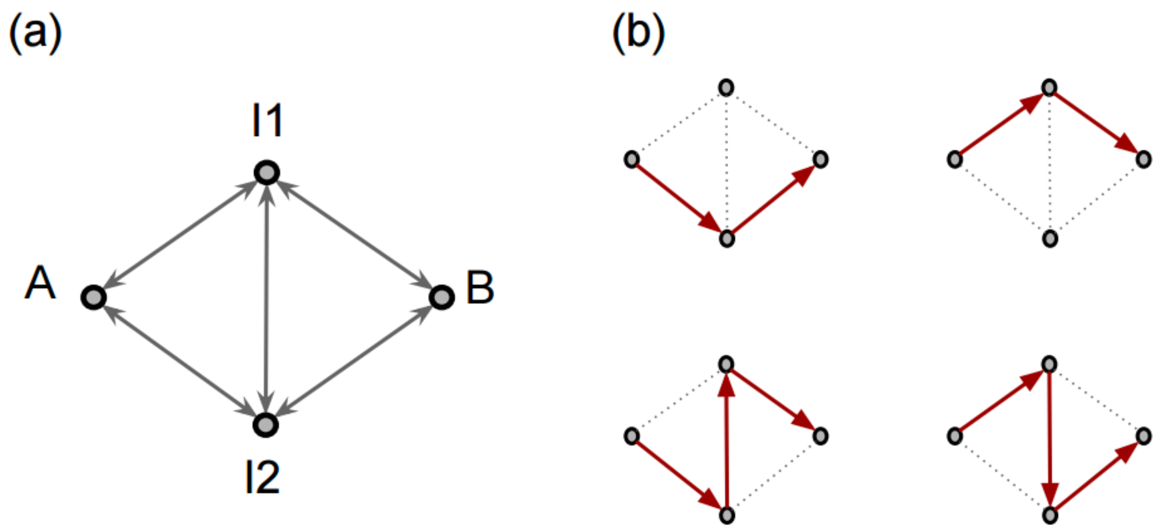
Reduced data, MSMBuilder States



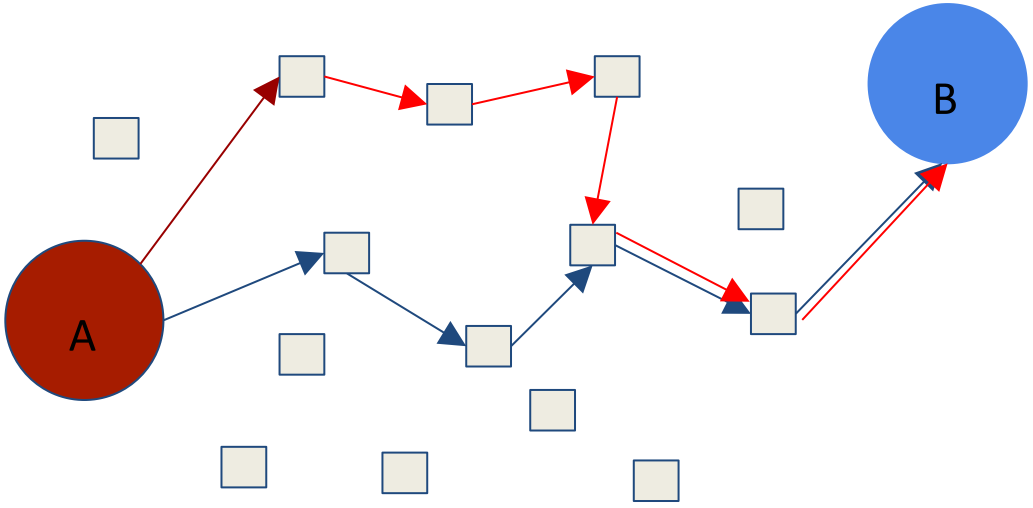
Mechanism

Markov vs Non-Markovian

Mechanism

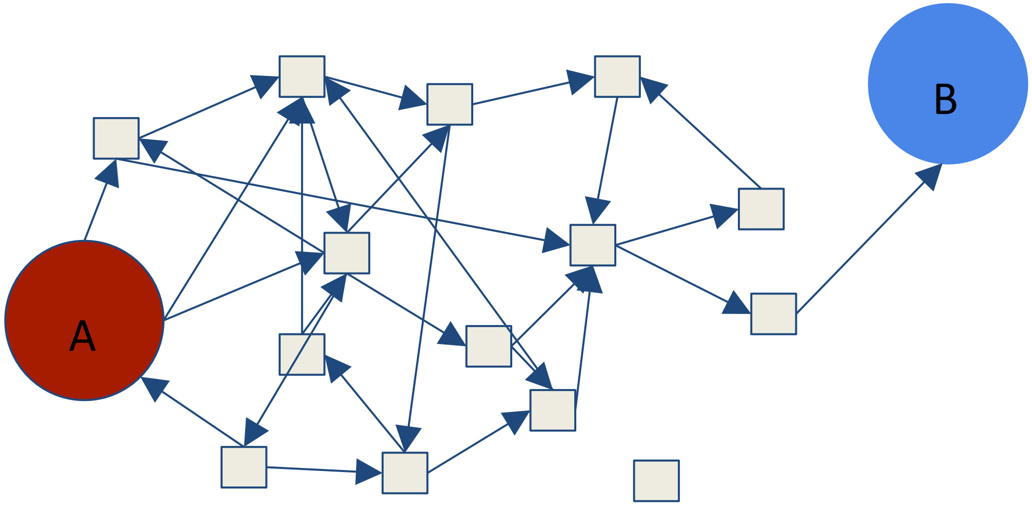


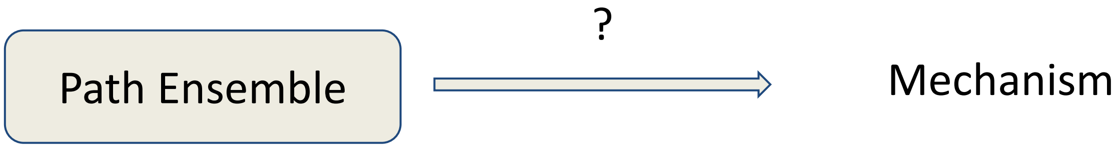
Mechanism



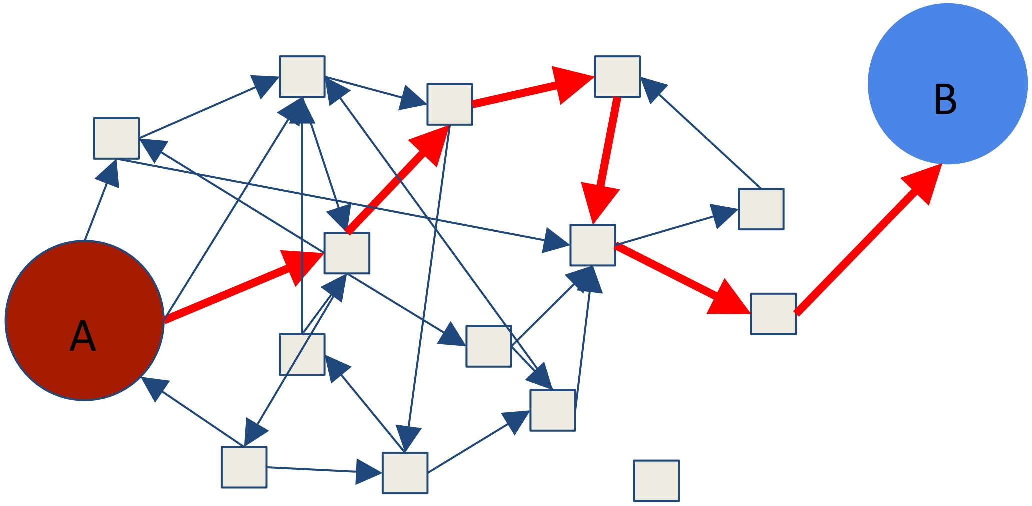
Mechanism

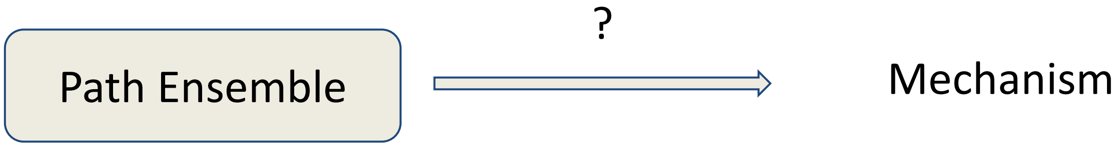
In practice we have...





Mechanism: Fundamental Sequence





Mechanism: Fundamental Sequence

Defining the ”backbone” of the path or *fundamental sequence* will allow us to divide the path ensemble in classes using an equivalence relation. Two paths that share the same fundamental sequence belong to the same class.

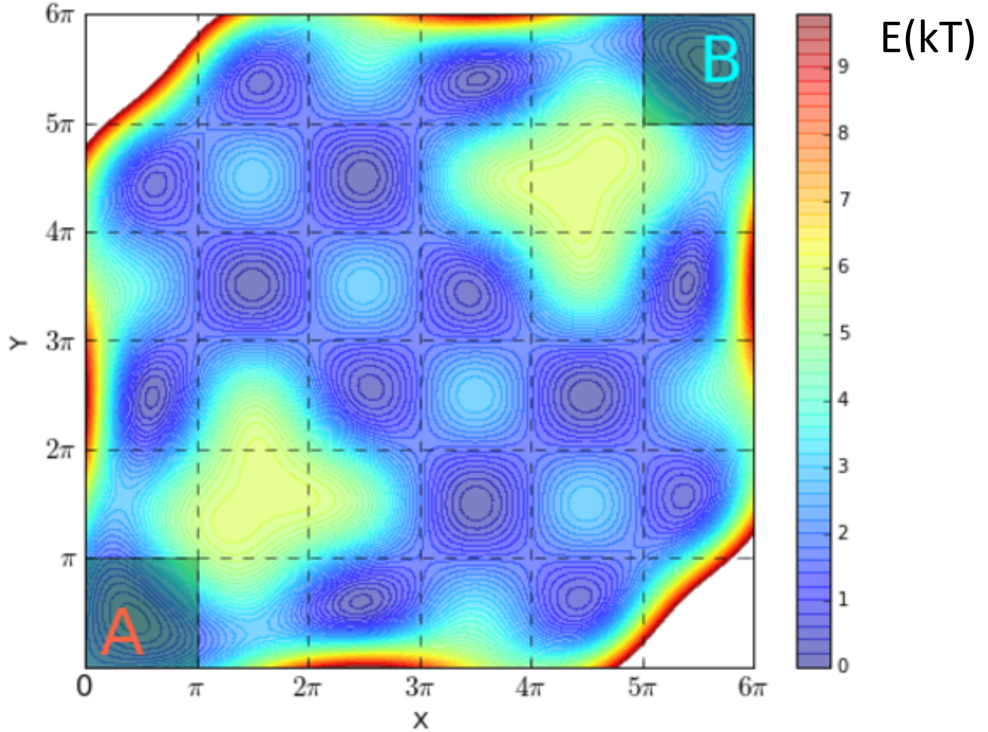
Def.

The fundamental sequence of a path is the most likely sequence that is consistent with the connectivity of the path. The likelihood is maximized in both directions.

$$\mathrm{FS}^* = \arg\min_{\mathbf{q}\in\Gamma(G)} \left\{ \sum_{i=1}^{|\mathbf{q}|-1} -\log(k_{q_i,q_{i+1}} k_{q_{i+1},q_i}) \right\}$$

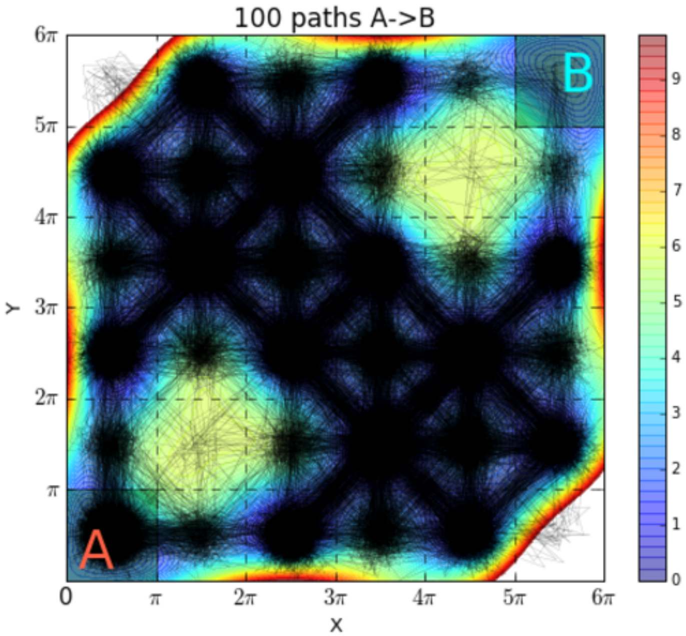
Mechanism: Fundamental Sequence

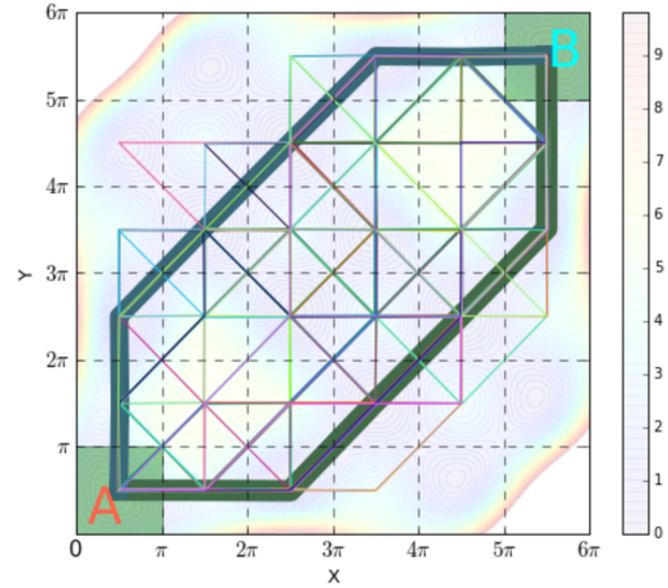
Example: 2D toy model



Mechanism: Fundamental Sequence

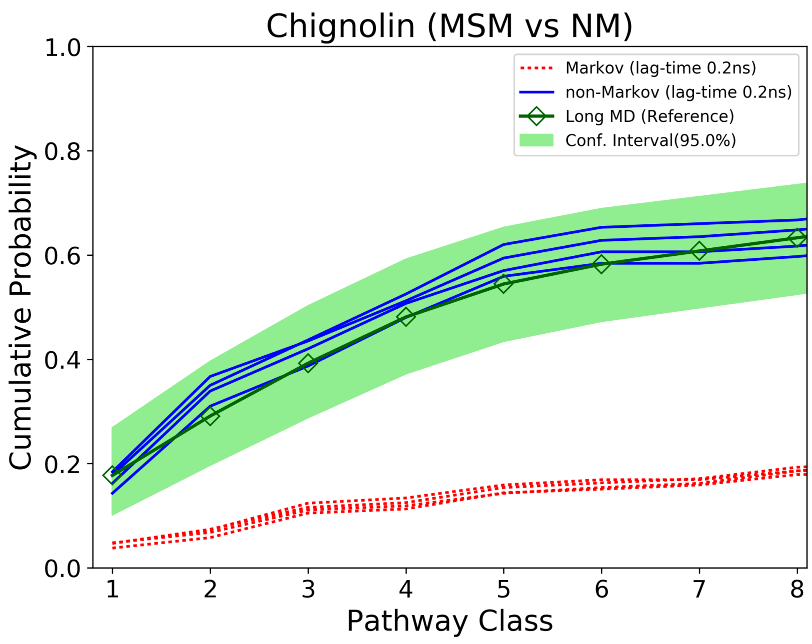
Example: 2D toy model

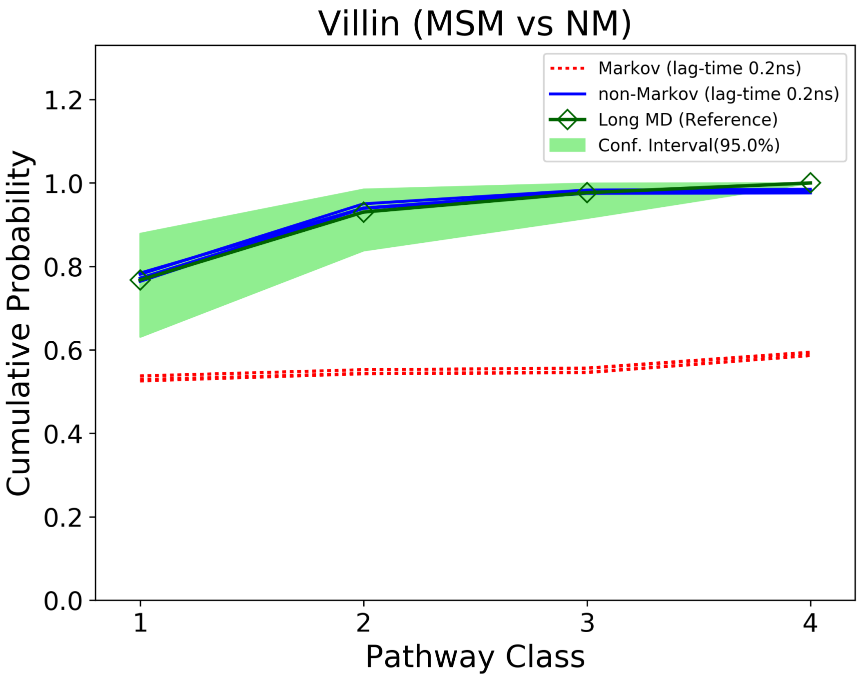




Mechanism: MSM vs NM

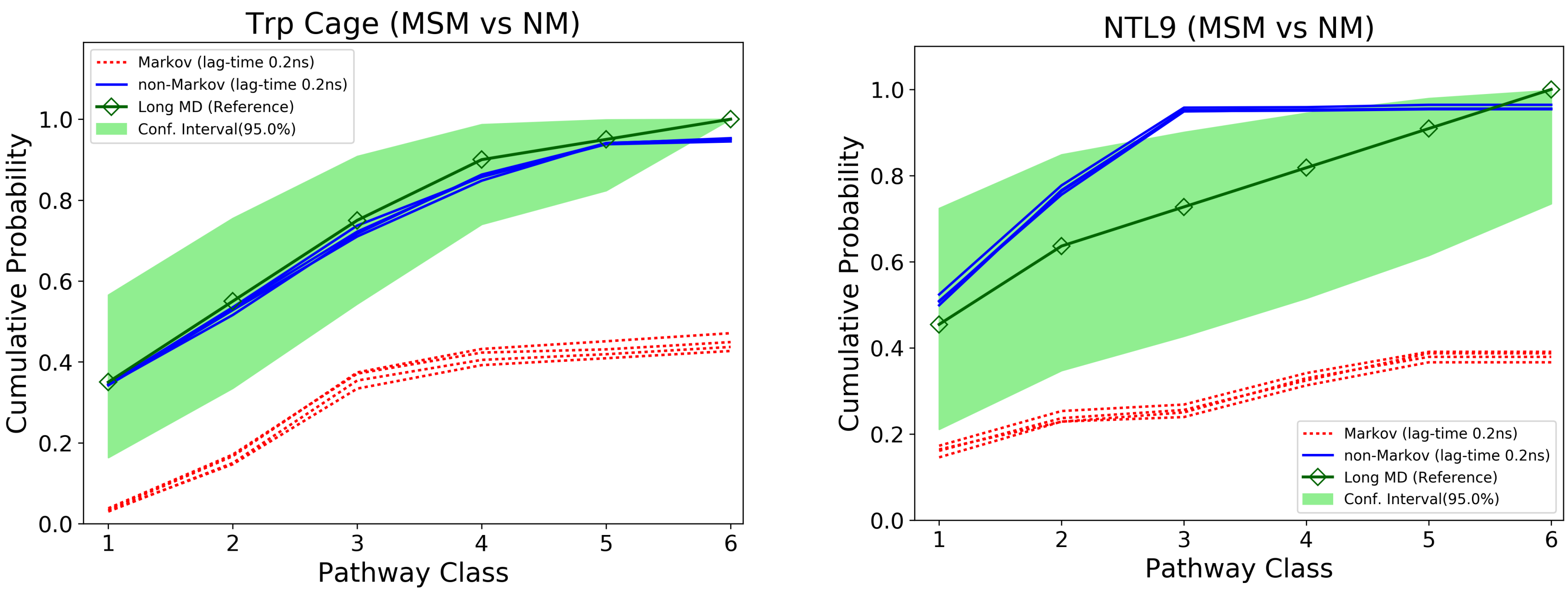
Classification based on the FS





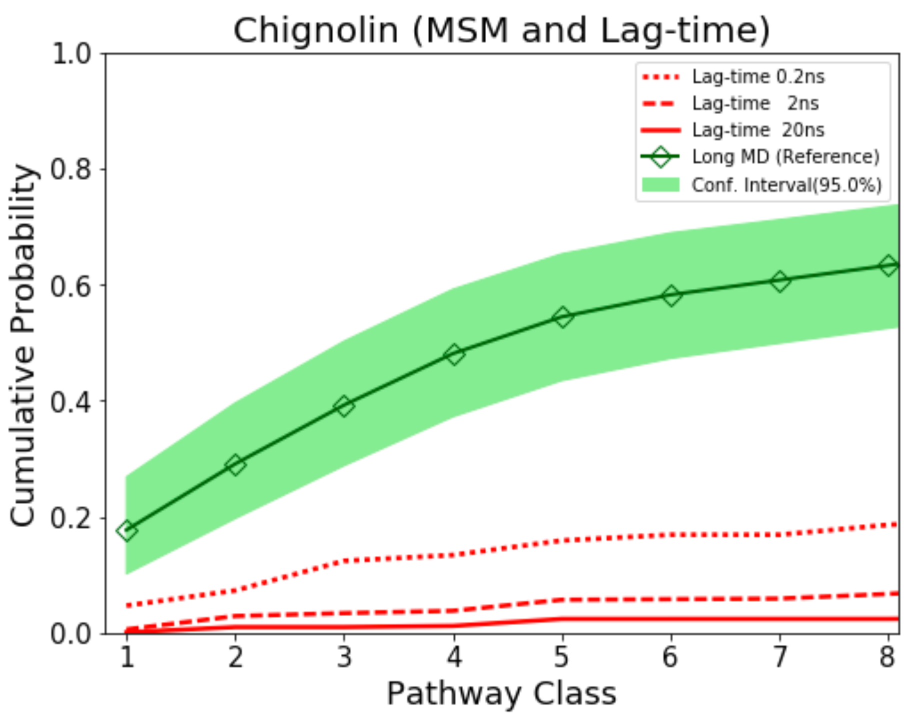
Mechanism: MSM vs NM

Classification based on the FS



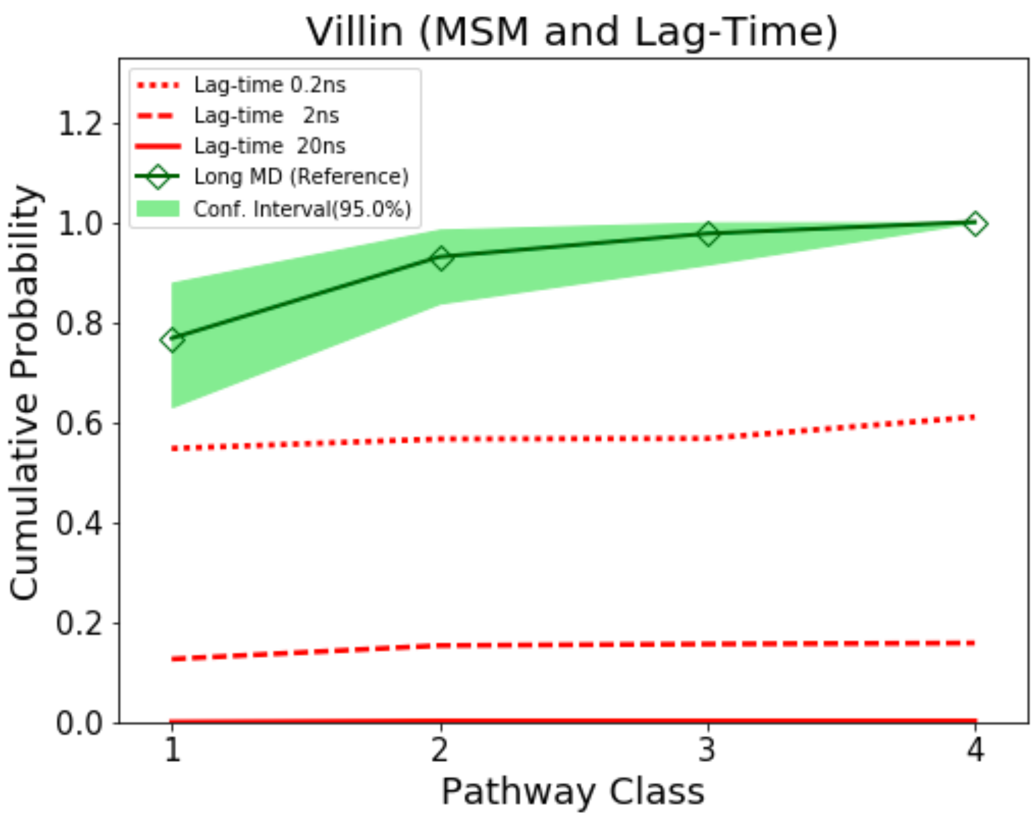
MSM vs Lag-time

Classification based on the FS



MSM vs Lag-time

Classification based on the FS



Conclusions

- The inclusion of color information in the analysis allows us to obtain unbiased MFPTs even when the partition of the space in bins is not optimal.
- In a non-Markovian regime, even with a relatively small amount of history (available in most of the MD simulations), we can improve dramatically the estimation of the MFPTs with respect to regular Markov Models.
- We can drastically reduce the amount of data and still obtain reasonable results.
- If the history is taken in to account, there is no need of lag-time “optimization”.
- The NM approach drastically outperforms MSM in the description of the mechanism/path ensemble.

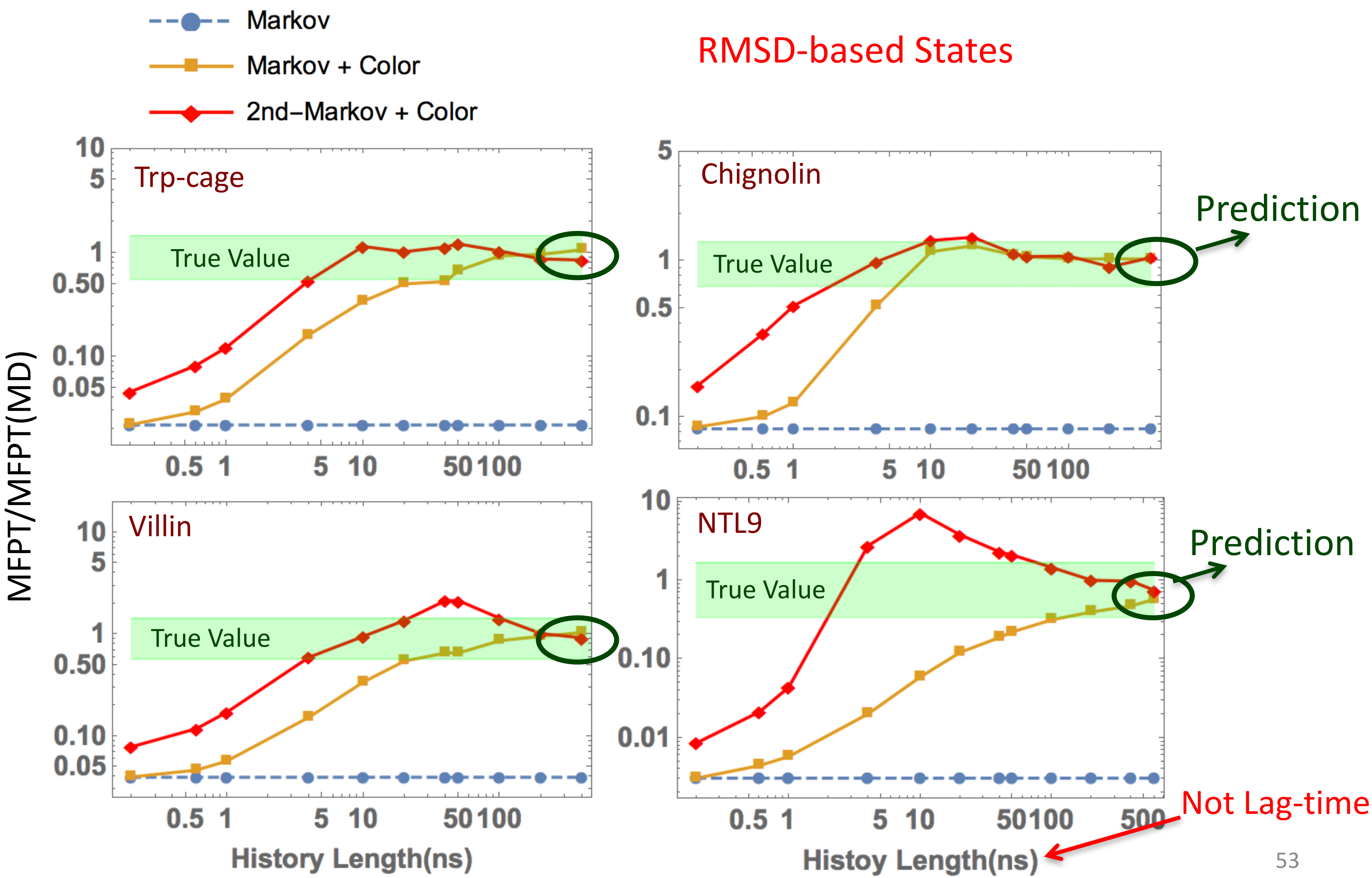
Acknowledgment

- Joshua Adelman
- Daniel Zuckerman
- Justin Spiriti
- Rory Donovan
- Ramu Anandakrishnan
- Ariane Nunes
- Ian Welland

- Shaw group
- NIH
- NSF

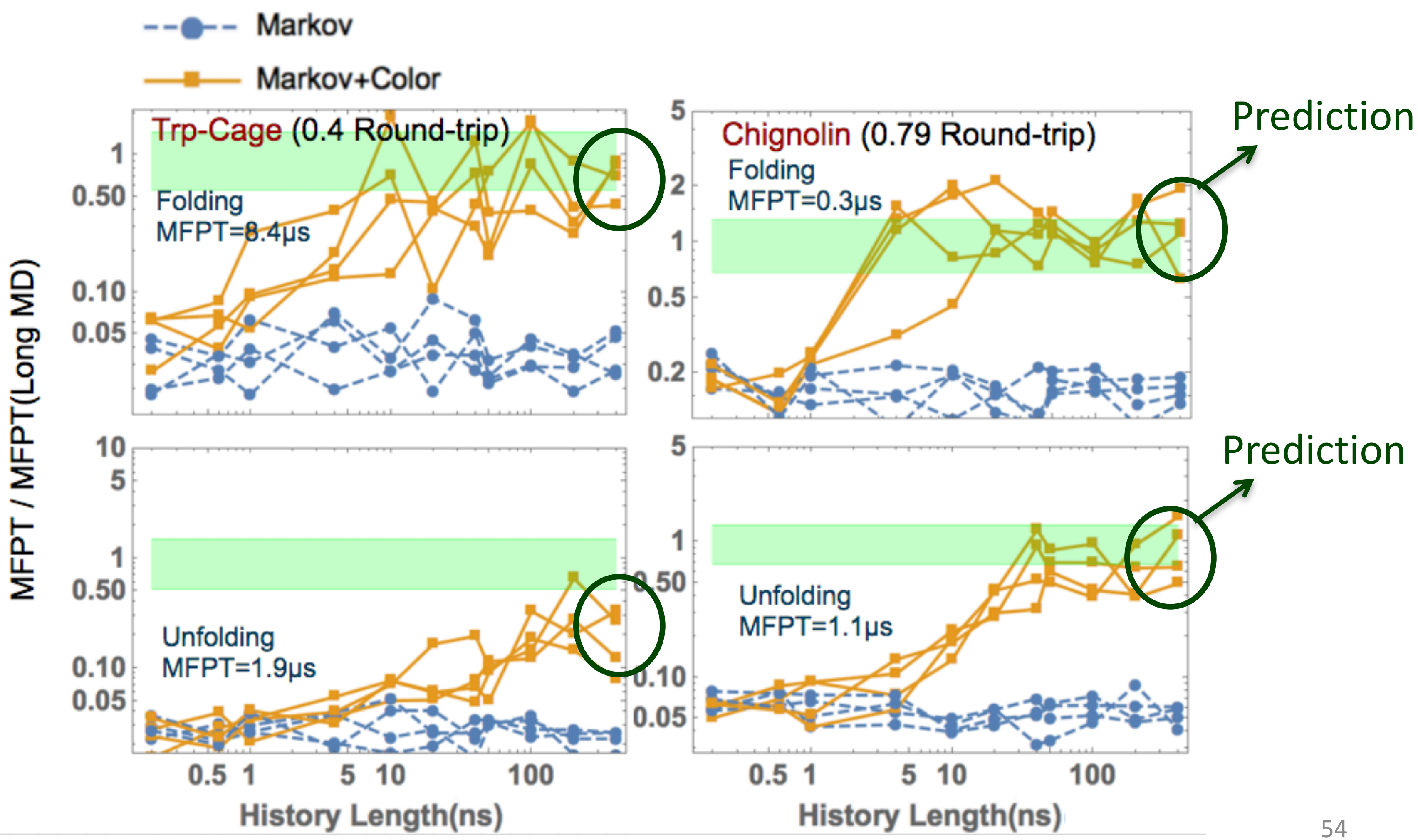
Supporting Information

Non-Markovian Analyses(Folding)

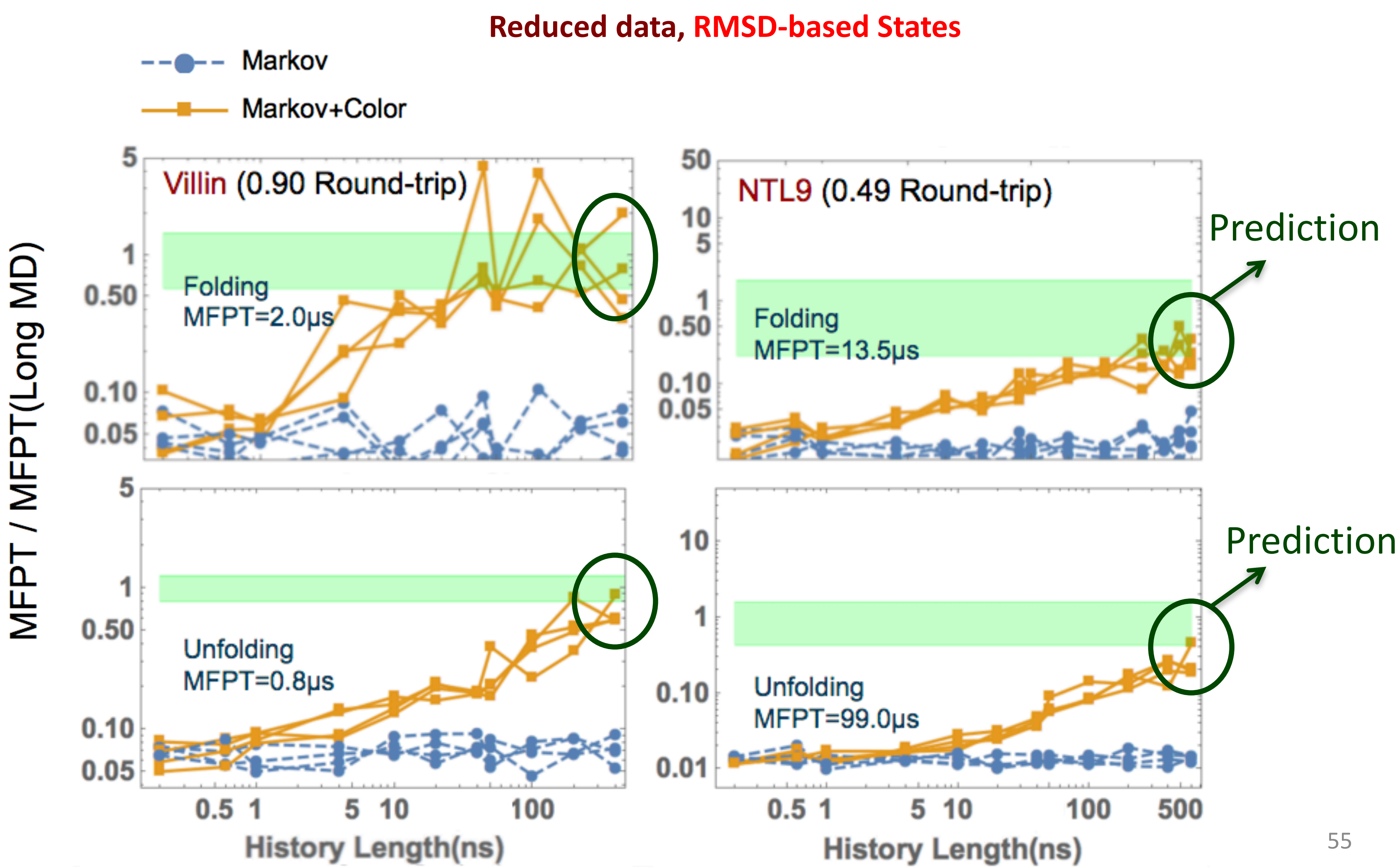


Non-Markovian Analyses

Reduced data, RMSD-based States

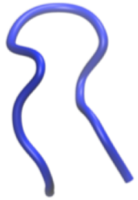


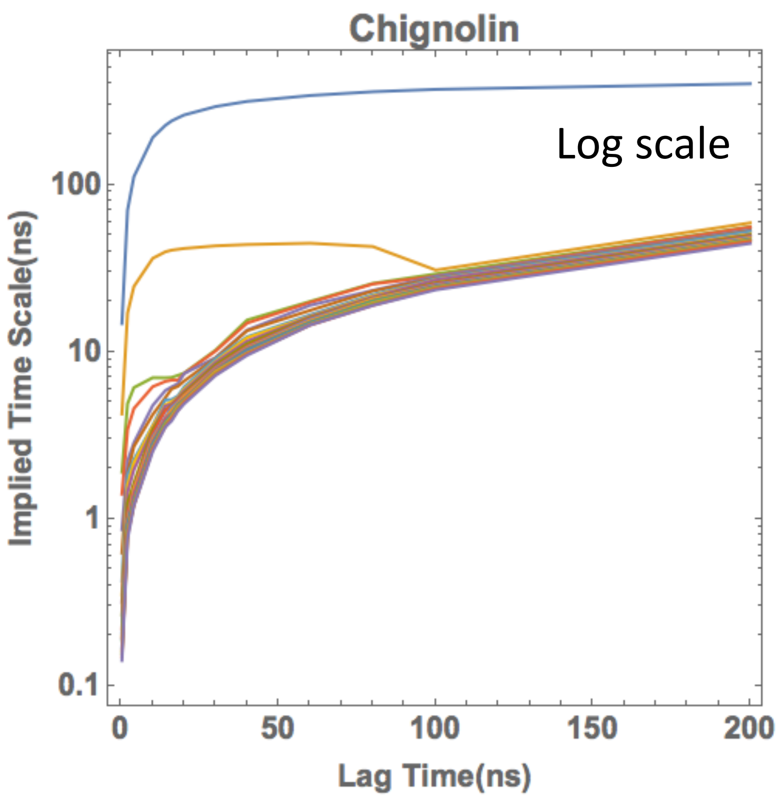
Non-Markovian Analyses

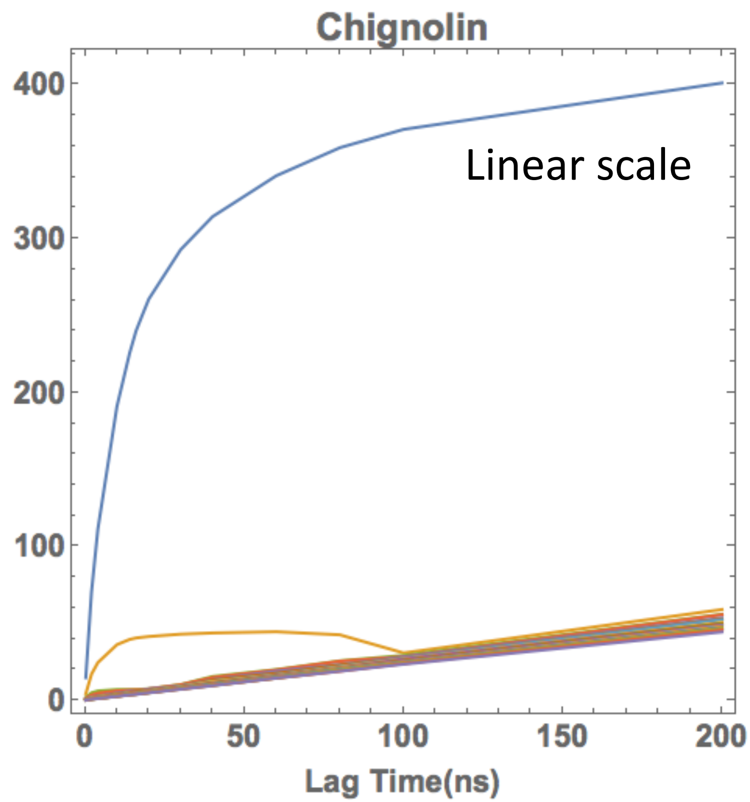


MSM: Implied time scales

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





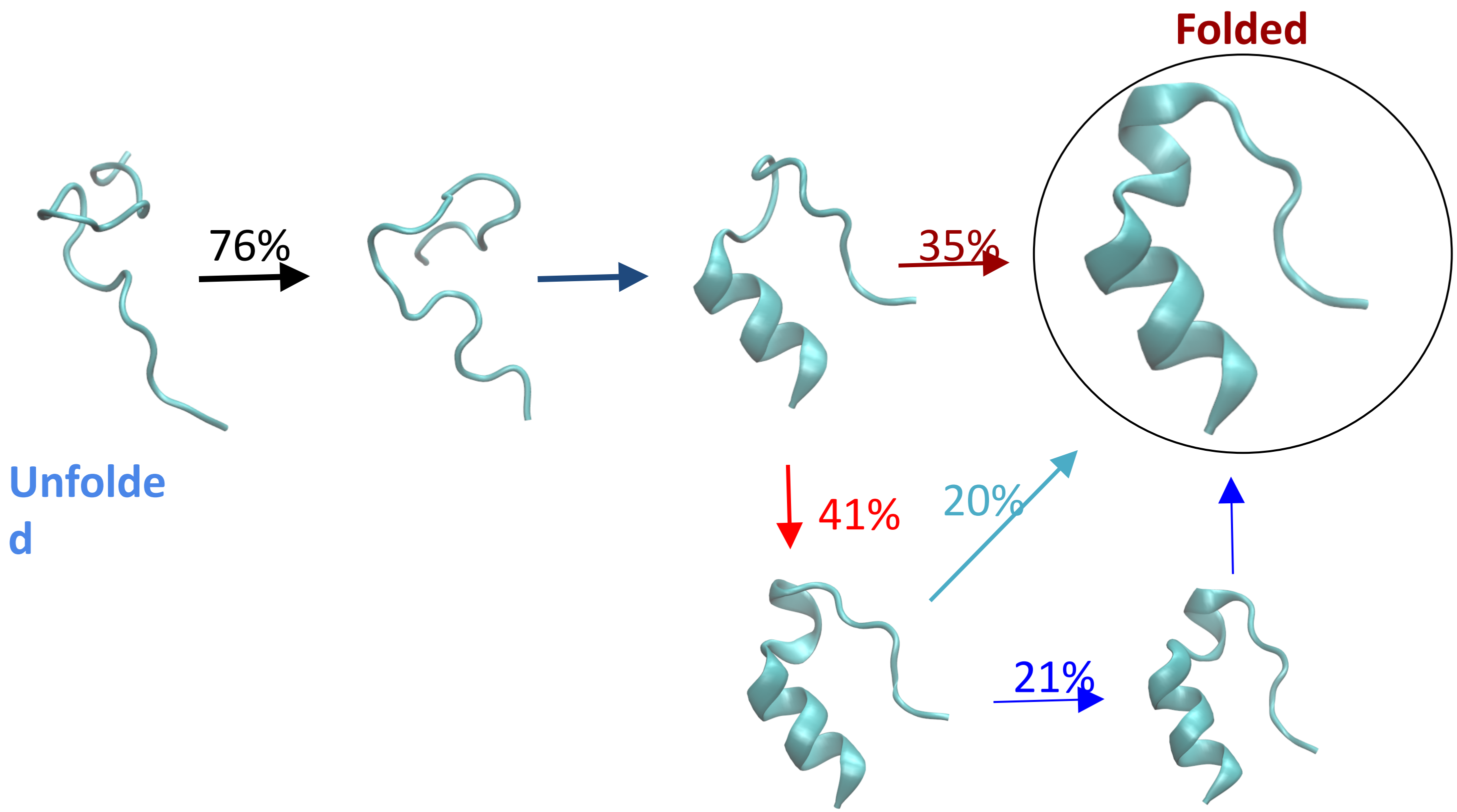


Beyond Markov: Color

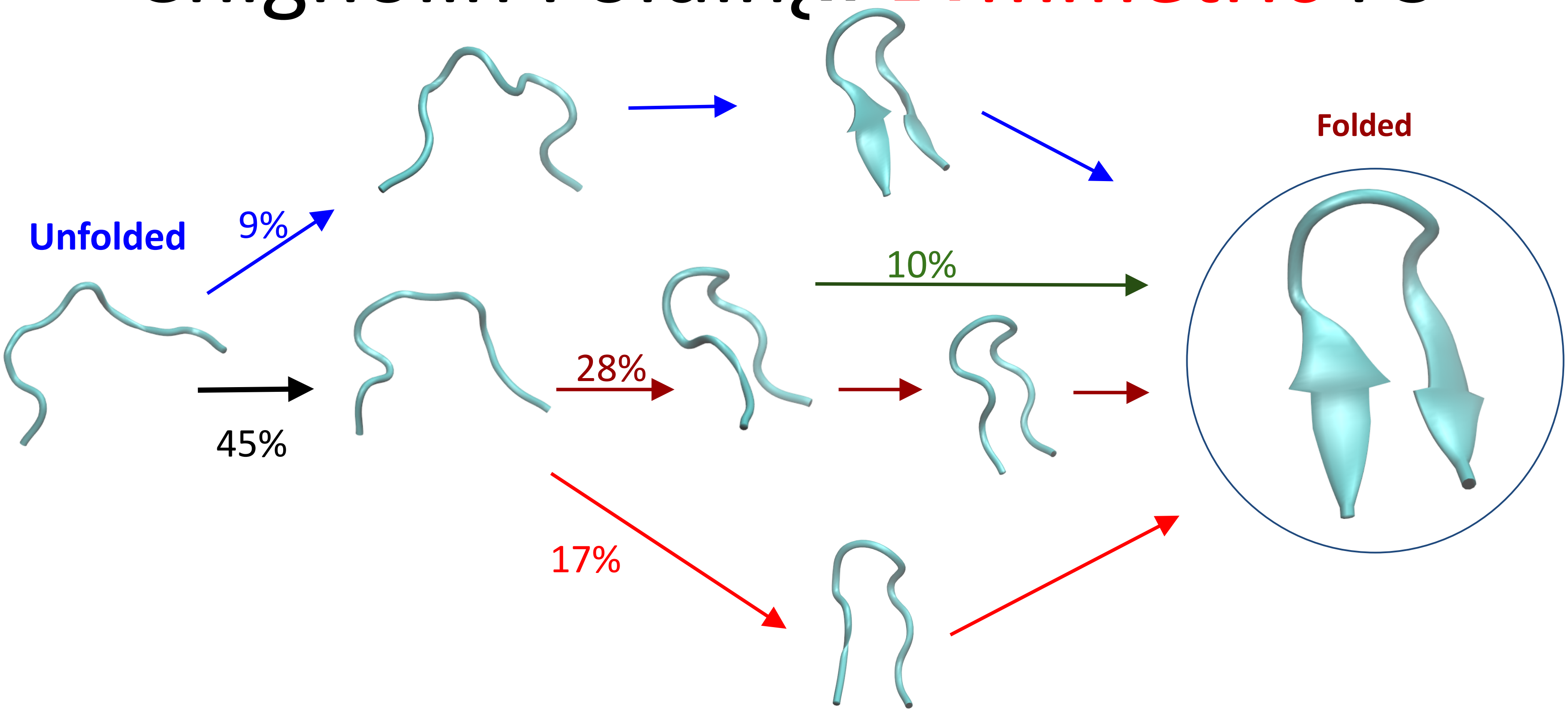
$$\mathcal{K}_{ml} = P\{X_{t+\tau} = \left\lceil \frac{l}{2} \right\rceil, L_{t+\tau} = \nu | X_t = \left\lceil \frac{m}{2} \right\rceil,\ L_t = \mu\},\ \ \mu, \nu = \begin{cases} \alpha, \alpha & \text{if } m,\ l \text{ are odd} \\ \alpha, \beta & \text{if only } m \text{ is odd} \\ \beta, \alpha & \text{if only } m \text{ is even} \\ \beta, \beta & \text{if } m,\ l \text{ are even} \end{cases}$$

Suarez et al., J. Chem. Theory Comput., 2014, 10 (7), pp 2658–2667
Vanden-Eijnden et al., J. Chem. Phys., 2009, 131(4), pp 44120

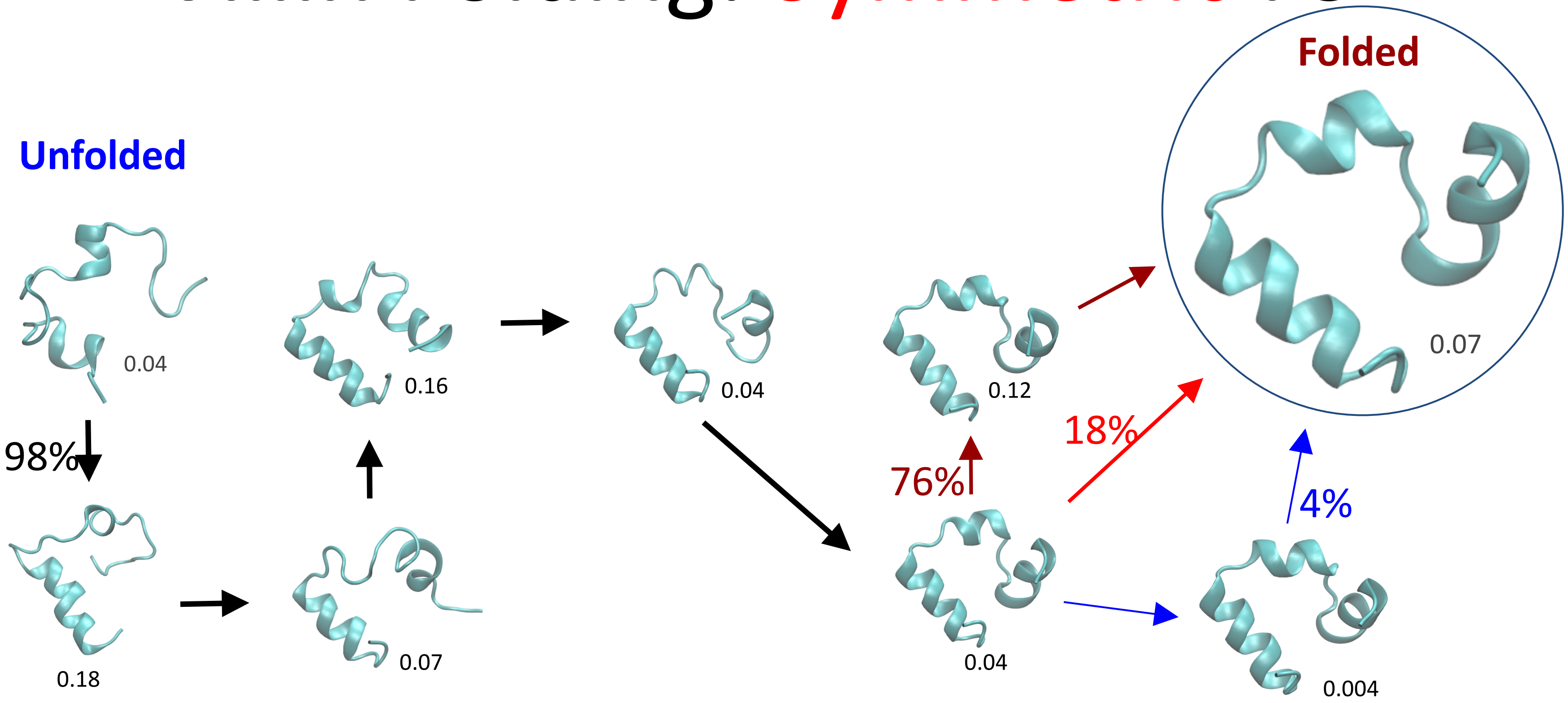
Trp-cage Folding: Symmetric FS



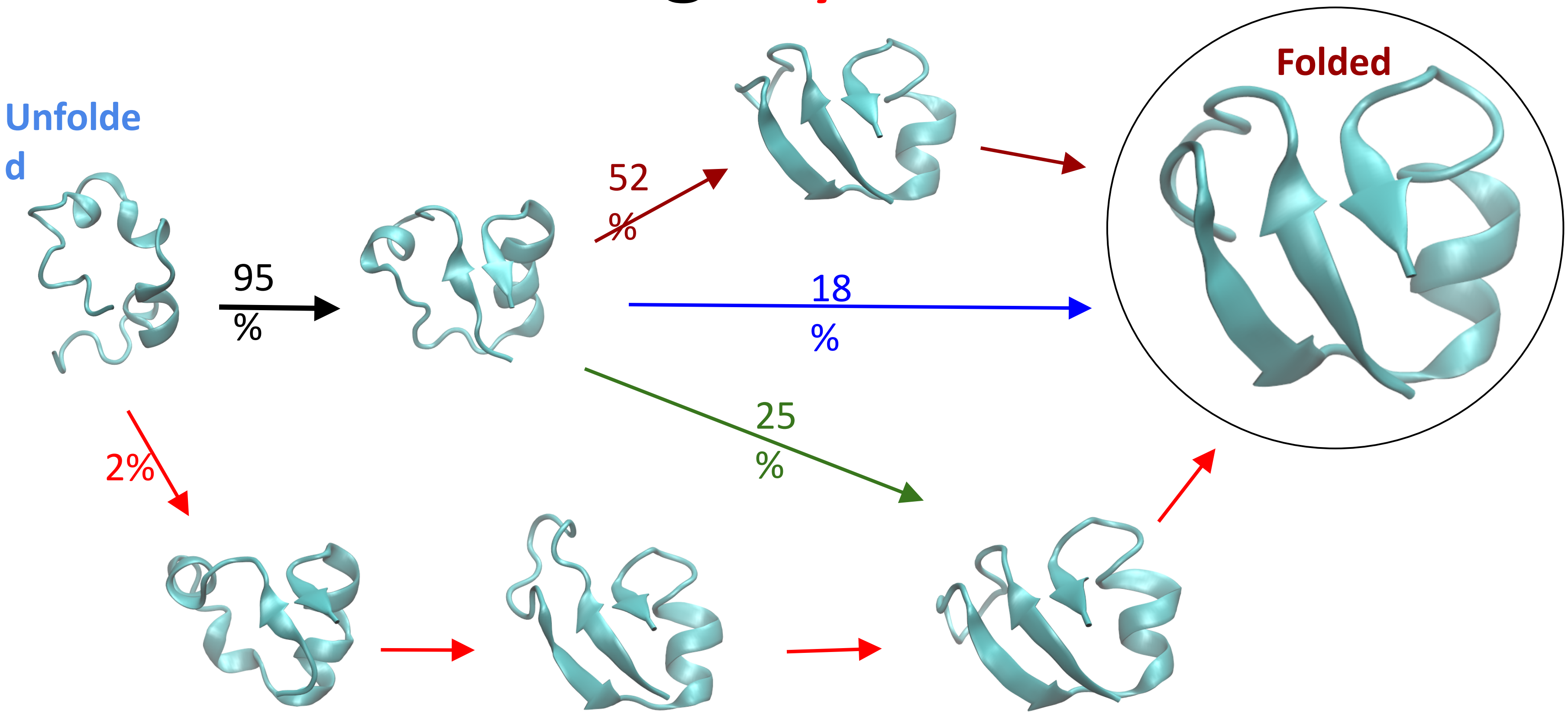
Chignolin Folding: Symmetric FS



Villin Folding: Symmetric FS



NTL9 Folding: Symmetric FS

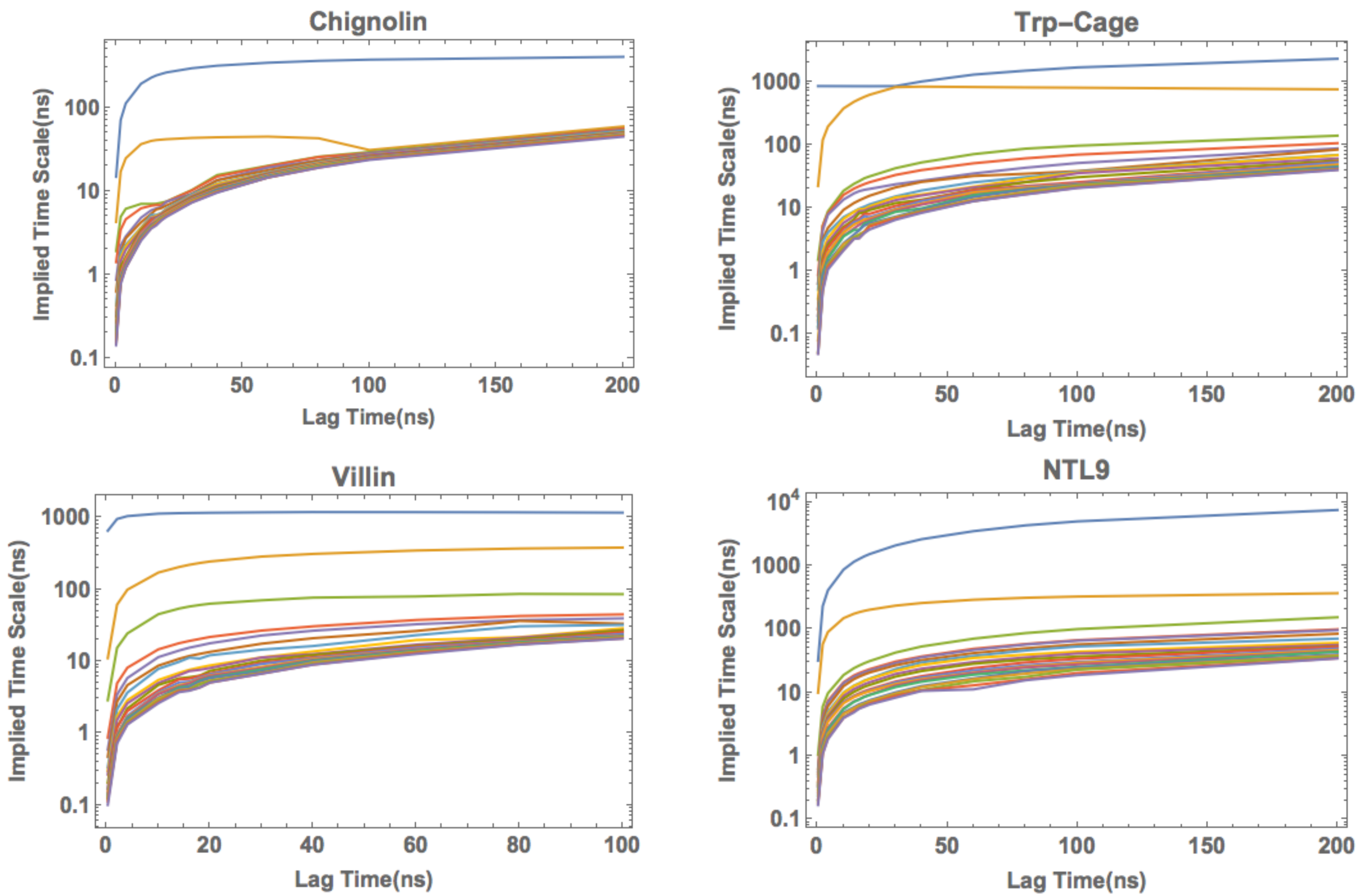


Protein models

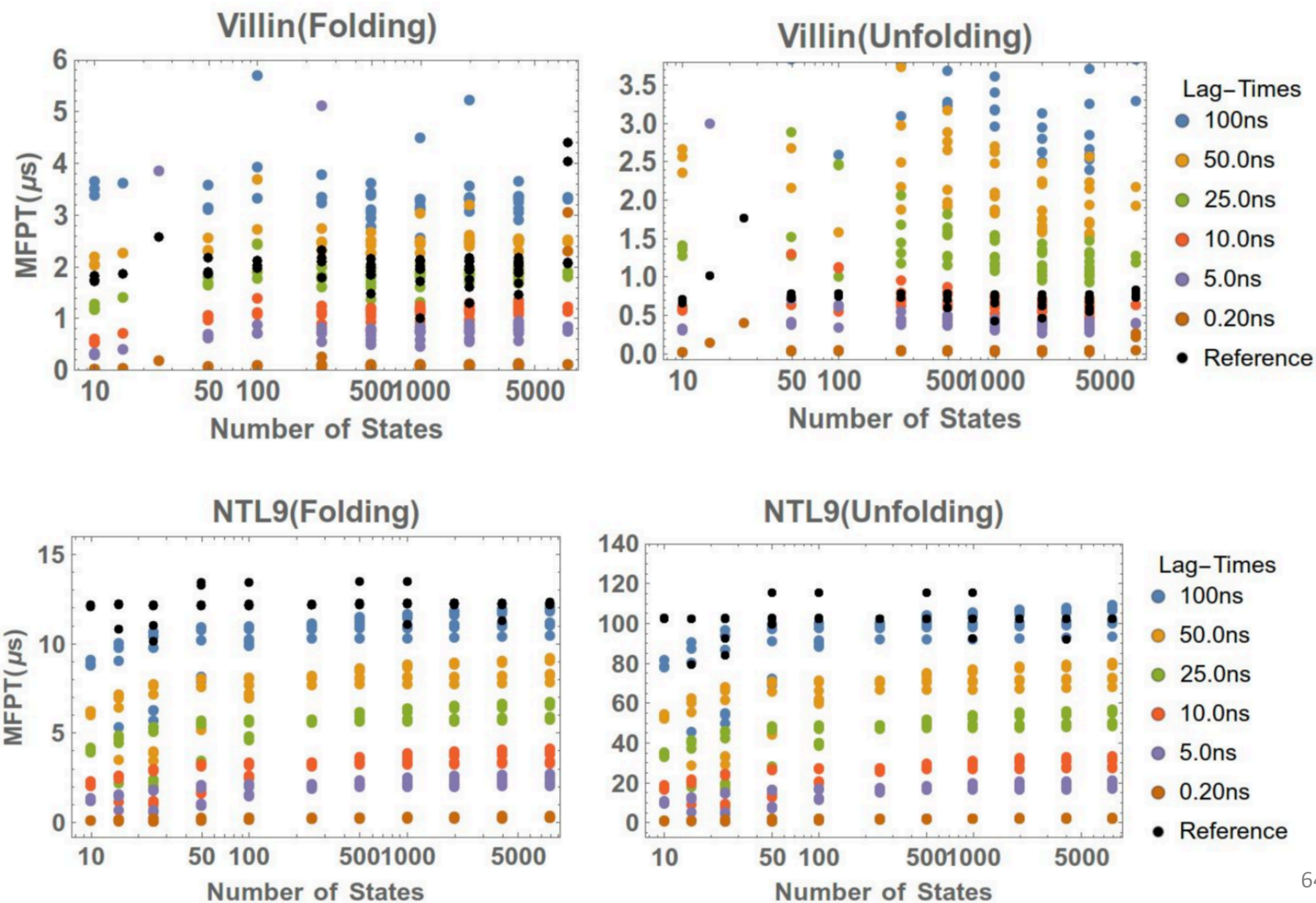
Table 1: Protein models used for Markovian and non-Markovian analyses. For each system, the table shows the number of residues, the total simulation time used in the analysis and the state definitions based on heavy-atom RMSD with respect to the folded structure whose protein data bank code (PDB ID) is given in the last column.

Protein	Num. Residues	Time(μs)	RMSD (Folded)	RMSD (Unfolded)	Reference Structure (PDB ID)
Chignolin	10	106	< 1.10Å	> 7.00Å	5AWL
Trp-cage	20	208	< 1.75Å	> 10.0Å	2JOF
NTL9	39	1100	< 1.50Å	> 10.0Å	2HBA
Villin	35	125	< 1.50Å	> 11.0Å	2F4K

Implied time scales

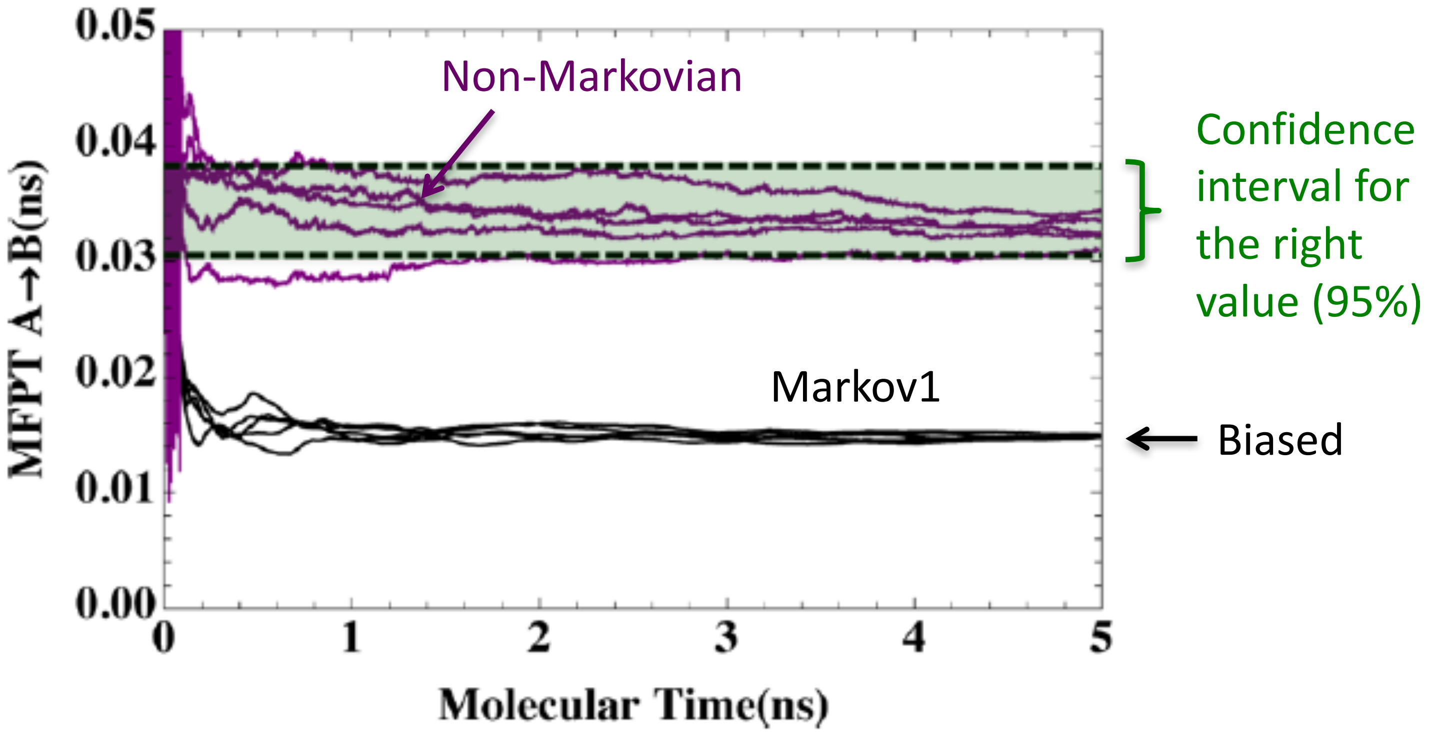


Markov State Models



Example: Methane/Methane

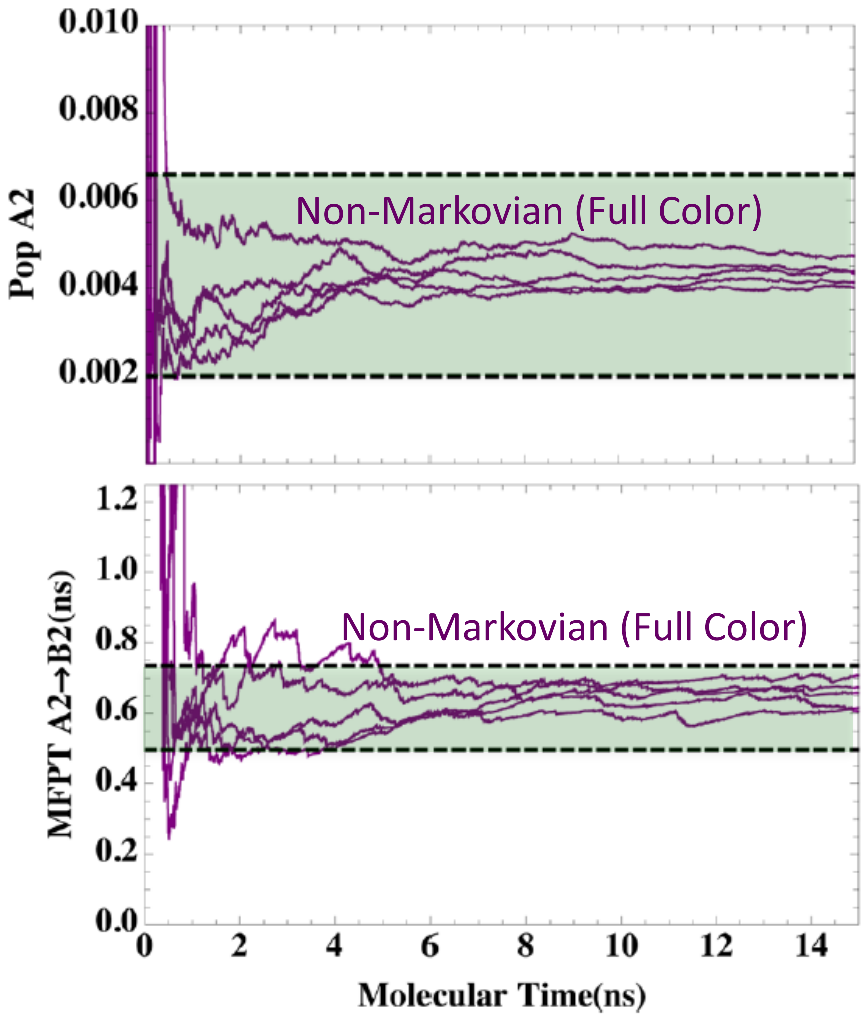
Dissociation process, 5 independent WE simulations.



J. Chem. Theory Comput., 2014, 10 (7), pp 2658–2667

Example: Ala4

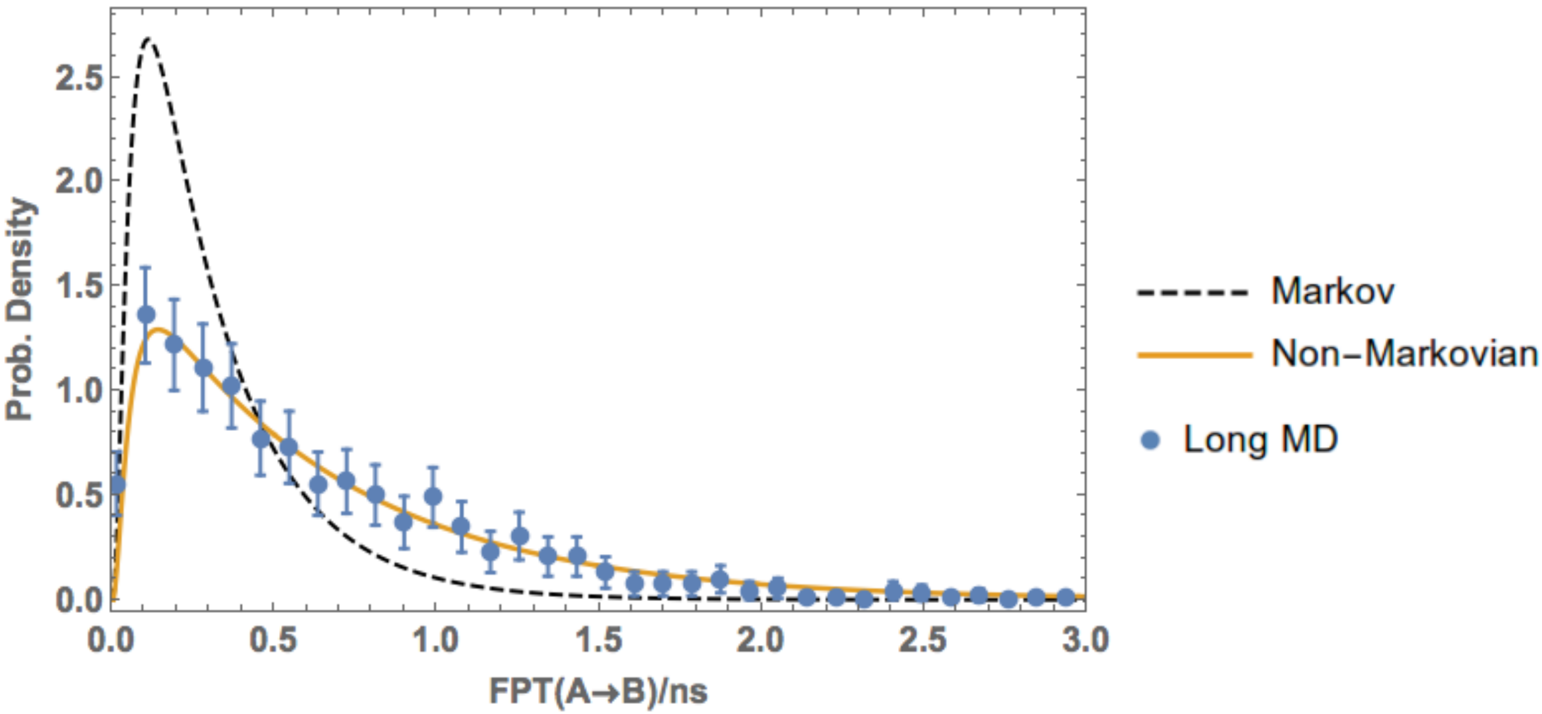
5 independent WE simulations.



J. Chem. Theory Comput., 2014, 10 (7), pp 2658–2667

Ala4

First passage time distribution



Protein Science 2016, 25, pp 67-78