# Optimal Parameter Selection in Markov State Models for Biomolecular Conformational Dynamics

Robert T. McGibbon, Christian R. Schwantes, and Vijay S. Pande

Markov state models (MSMs) are a powerful tool for the analysis of molecular dynamics simulations, but have been hampered by the need for manual selection of the number of states. We report a new method for the optimal selection of the number of states in an MSM based on the Bayesian information criterion. We demonstrate the approach on three systems of increasing complexity...

## I. INTRODUCTION

Protein dynamics are an essential aspect of biomolecular function. These dynamics span a wide range of length scales, timescales and complexity, including folding and aggregation, conformational change between functional native substates, ligand binding, and allostery [1–4]. Whereas classical experimental probes have often been interpreted in two-state frameworks, ensemble measurements with increasingly high temporal resolution as well as sensitive single molecule probes have uncovered a vast array of complex kinetics [5, 6]. But atomic-resolution structural characterization of these dynamics is often a Herculean challenge, particularly in dynamical settings – as molecular probes like Förster resonance energy transfer, small-angle x-ray scattering, and nuclear magnetic resonance techniques measure complex projections of the intrinsic structure, generally reporting simultaneously on many degrees of freedom[7, 8].

Computer simulations can complement experiments by providing atomic-resolution insight into the structural dynamics. With advances at the algorithmic, hardware, and software levels, modern molecular simulation paradigms, incorporating specialized or accelerated hardware, often in combination with highly parallel distributed computing frameworks, are capable of generating extensive simulation data sets[9–12]. In fact, the minimally-biased kinetic analysis of such simulations is often a central bottleneck and presents a major challenge to the field. The analysis paradigms often entail the construction of lower resolution models parametrized from the high resolution simulation data set which capture the essential features in an interpretable framework[13, 14]. For example, by projecting the data down onto one or two degrees of freedom we create a simpler model for the system, such as one characterized by diffusion along a single reaction coordinate[15].

Markov state models (MSMs) are one approach for analyzing MD data sets and driving further MD simulations that are able to smoothly move between high and low-resolution models[16–19]. Such detailed models maintain quantitative agreement with the underlying simulation data, while low-resolution models capture the salient features of the potential energy landscape, sacrificing some degree of model complexity. In an MSM, the dynamics are modeled as a memory-less jump process between a discrete set of conformational states. The two key quantities which define the MSM are thus the state definitions, an indicator function basis over phase space, and the pairwise transition probabilities or transition rates, which parameterize the kinetics.

A significant challenge in the automated construction of Markov state models is the choice of the number of states[20]. Although classical Hamiltonian dynamics form a continuous-time Markov chain in $\mathbb{R}^{6N}$, the Markov property does not hold after the projecting the dynamics onto a basis of discrete indicator functions. In particular, when states contain within them free energy barriers of substantial magnitude, the validity of the Markov assumption begins to suffer considerably. While this source of modeling error can be addressed by increasing the number of microstates, the reduction in one error comes at the expense of the increase in another. This second source of error is statistical in origin. As the number of states in the model grows, so does the number of parameters required to completely specify the kinetic model between all pairs of states. Because the amount of data is constant, each additional parameter leads to a decrease in the amount of data available per model parameter, which makes the approach susceptible to over-fitting.

Here, we seek to build models that are *suitably* complex, given the data, yielding complex descriptions of the system only to the extent that their additional parameters are implied by the observed dynamics. To that end, we introduce a new procedure for scoring the likelihood of an MSM, which, together with cross validation and the Bayesian information criterion (BIC), enables the optimal selection of the state space, which we express both in terms of the number of states and the clustering algorithm employed to group sampled conformations into states. This approach complements validation procedures performed primarily based on human intuition, such as Chapman-Kolmogorov tests, and enables the treatment of model selection as an optimization problem amenable to automated methods.

## II. LIKELIHOOD OF A MARKOV STATE MODEL

With the kinetic model expressed as a set of pairwise state to state transition probabilities at a given lag time, the likelihood of an ensemble of trajectories after projection into the indicator function basis is given simply by the product of the transition matrix elements along the

observed trajectories. However, as we vary the number of states, it is not permissible to simply compare these likelihoods as part of an optimization of the state definitions. In doing so, the optimal model would always be the trivial one state model, whose computed likelihood is unity regardless of the data.

The appropriate likelihood is instead a path action in phase space, on which the discrete states are merely an indicator function basis. With $s(X)$ as the function mapping conformations into the indicator function basis set, $s : \mathbb{R}^{3N} \to \{1, 2, \ldots, K\}$, the likelihood can be written as

$$P[x_{0\ldots T-1}]dx^T = \prod_{i=0}^{T-1} T(s(x_i) \to s(x_{i+1})) \cdot \prod_{i=0}^{T} p_{s(x_i)}(x_i) \tag{1}$$

With a discrete, non-overlapping state space, the likelihood of a sampled trajectory can be decomposed into a product of terms of two types: the state to state transition probabilities, $T(s_i \to s_j)$, and so-called emission distributions of each state, the conditional probability of observing a conformation at a given location in phase space given that the conformation, $x_t$ is within a certain state, $s(x_t)$.

For example, consider two Markov state models sharing the same transition matrix, $T$. In one model, the state emission distributions are highly peaked at specific locations in phase space, whereas in the other model the emission distributions are uniform over the volume of the states. If an observed trajectory does go through the first models' regions of high likelihood, it is appropriately termed a more likely model given the data.

However, models' long timescale behavior – the rates and fluxes between metastable basins – are independent of the choice of the emission distributions. The emission distributions characterize only the fine details of the equilibrium distribution within states, a quantity that the MSM approach does not seek to model; implicit in the decision to group conformations together into states is the idea that we decline to model the differences between conformations belonging to the same state. If such differences exist and are sufficiently large to warrant attention, than the states are too large.

Therefore, the most appropriate emission distribution for discrete state MSMs is that of the uniform distribution over the phase-space volume of the state. That is, the likelihood of observing a conformation in phase space given that the conformation is assigned to state $i$ is 0 if the conformation is outside of the bounding volume of the state and constant if the conformation is within the volume. This constant is set so that the distribution integrates to 1, and is thus the reciprocal volume of the microstate.

$$P[x_{0\ldots T-1}]dx^T = \prod_{i=0}^{T-1} T(s(x_i) \to s(x_{i+1})) \cdot \prod_{i=0}^{T} \frac{1}{V_{s(x_i)}} \tag{2}$$

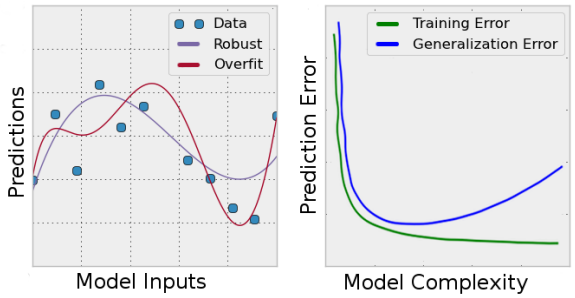


FIG. 1. An example of overfitting. Left: in the presence of noise, a more complex model is able to fit the observed training data better, with lower residues and a higher empirical likelihood, but fails to distinguish the signal from the noise. Right: more complex model classes exhibit lower training errors, but generalize poorly to unobserved data.

## III. CROSS VALIDATION AND THE BAYES INFORMATION CRITERION

Likelihood maximization is insufficient for model selection when the number of parameters varies between proposed models, as more complex models generally exhibit higher empirical likelihoods, often at the cost of larger generalization errors due to overfitting[21, 22]. Statistical learning theory provides a number of alternative approaches for this problem. Conceptually, the most straightforward is a full Bayesian treatment in which all unknown model parameters are represented by probability distributions. The evidence for a model is computed by formally integrating over the model parameters and the evidence ratio, or Bayes factor[23], then provides a rigorous basis of model selection that appropriately punishes overly complex models as they become poorly constrained in parameter space. Unfortunately such approaches are intractable for problems of this size because of the need to integrate over all possible Markov models of a given size.

Instead, we explore both cross validation and Schwarz's Bayesian information criterion (BIC)[24] for choosing the number of states in a Markov state model. For cross validation, we parameterize the model, building both the state space and transition matrix on a subset of the data, but evaluating the likelihood on the left-out portion in an attempt to directly measure the generalization error. The BIC on the other hand involves augmenting the likelihood with a penalty on the number of parameters, and is an asymptotic approximation Bayesian evidence.

$$\mathrm{BIC} \equiv -2 \cdot \ln L + k \cdot \ln N \tag{3}$$

where $L$ is the likelihood, $k$ is the number of free parameters, and $N$ is the number of data points, assumed to be independent and identically distributed.

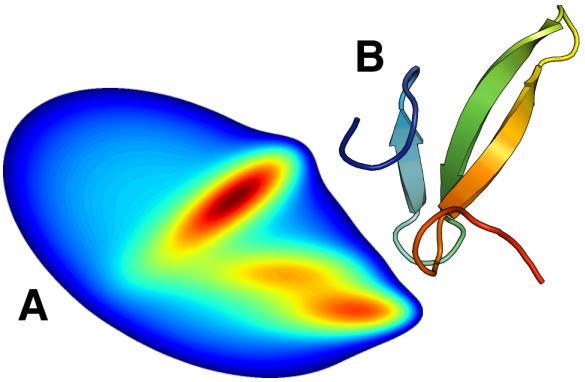


FIG. 2. Systems studied in this work. (A) Langevin dynamics on the two dimensional Müller potential. (b) 200 $\mu s$ of dynamics of the Fip35 WW domain[25], courtesy of D.E. Shaw research [26].

## IV. COMPUTATIONAL METHODS

The uniform distribution emission model presents a computational challenge: its use requires the calculation of the (hyper)volume of the MSM's states, which, when defined by clustering are high-dimensional Voronoi cells. While trivial in two or three dimensions, this computational geometry task becomes challenging in high-dimensional settings. The computation of such volumes has occupied significant attention in recent years in the computational geometry literature, especially via randomized algorithms[27–29]. We opt to approximate the volumes using naive Monte Carlo rejection sampling, which we find tractable for large systems only when the molecular dynamics dataset is first projected into a suitable small vector space of up to perhaps ten dimensions.

A further challenge is the procedure by which to model the volume of states which are at the "edge" of the MSM – whose Voronoi cells extend to infinity in some direction. Is the volume of these states unbounded? It is appropriate to assert that the volume of these edge states is bounded in some way by the extent of our dataset. For example, the volume of a state might be defined as the volume of the intersection of its Voronoi cell and the convex hull of the whole dataset, which would encode the assumption that the likelihood of observing conformations outside the convex hull of the sampled data is vanishing.

We use a slightly modified version of this definition that adopts the same spirit. Instead of taking the outer bounding envelope to be the convex hull of the data, we take it to be the set of all trial points such that the nearest sampled configuration to the trial point is closer than a certain cutoff, $R$. For further efficiency, it is feasible to use only a random subsample of the dataset for this nearest neighbor computation.

## V. RESULTS AND DISCUSSION

### A. Müller Potential

We simulated $1\times10^7$ steps of Langevin dynamics on the Müller potential, clustered the data using the $k$-centers clustering algorithm with a euclidean distance metric, and computed the state volumes with $1 \times 10^5$ rounds of Monte carlo rejection sampling, using a padding distance of $R = 0.2$ around the state centers from the 500 state model to define the dataset's envelope.

The volumes did not change drastically upon doing ten times more Monte Carlo rejection sampling, and furthermore, the likelihood remained almost exactly constant, which gives us some confidence that the volumes do not need to be perfect in order to have a good estimate of the likelihood. I can make a figure, and we can put it in the SI.

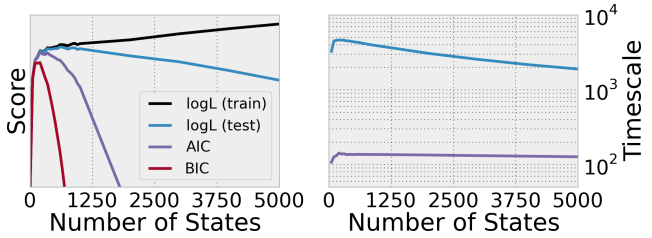


FIG. 3. In models built using between 50 and 1,000 states with the $k$-centers algorithm, the log likelihood function increased quickly and plateaued at approximately 300 states. There are many methods that can be used to test the transferability of an MSM. We've plotted the AIC, BIC, and cross-validation scores for the Müller potential simulations. These results indicate that, at least for this dataset, the BIC and AIC may penalize the number of states too heavily. See SI section: "AIC vs BIC" for a discussion of the various methods in detail.

As shown in Fig. 3, models built with too few states achieve a drastically reduced likelihood, but above a threshold region the likelihood increases relatively slowly. Here, the AIC and BIC's penalty on the number of parameters, which scales with the square of the number of states, begins to dominate. The optimal models according to the three methods are between 200 and 1000 states for this system. This observation is consistent with the convergence of the timescales of the MSMs, but allows for a direct optimization of the state space, as opposed to a heuristic way of comparing models.

Interestingly, the AIC and BIC penalize the larger state models much more than the cross-validation approach. This could be due to the simplicity of this system. Since the Müller potential is only two-dimensional, it is difficult to actually produce a drastically over-fit model. We suspect that the cross-validation score on a real protein system would be closer to the BIC and AIC. However, we note, that cross-validation requires estimating a model without some subset of the data, in which

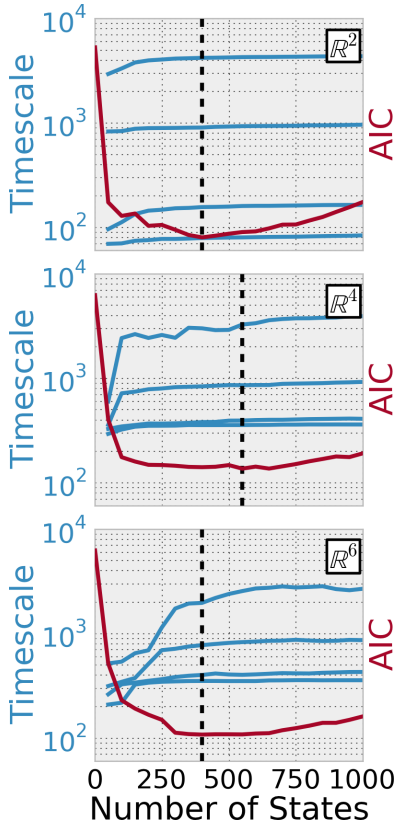


FIG. 4. Implied timescales and AIC scores for a series of Markov state models for the Fip35 WW domain. The panels differ by the dimensionality of the tICA projection. All models were built using a Markovian lag time of 50 ns. Qualitatively, the optimal model according to the AIC, corresponds roughly to where the timescales begin to flatten out.

case, it is not feasible for large systems that are not oversampled.

### B. Fip35 WW Domain

To test the procedure on a larger protein system, we reanalyzed two ultra-long 100 $\mu s$ molecular dynamics trajectories of the Fip35 WW domain[25], provided courtesy of D.E. Shaw Research [26]. In order to reduce the dimensionality of the problem, especially critical for the computation of the state volumes, we first preprocess the trajectories using time-structure based independent components analysis[30], retaining between two and six uncorrelated linear combinations of residue-residue distances.

How many states are required for a Markov state model? Lane et. al., when analyzing this same dataset, proposed a 26,000 state Markov state model after clustering the data by RMSD, whereas Kellogg et. al. using an approach similar to that described herein, arrive at a 1,000 state model after clustering on contact maps[31, 32]. That number is the micro state model from KCenters and CM's, but they also did KMeans and used 100 states.

According to this likelihood approach, the most likely models built on tICA projections are in the hundreds of states range. This does not mean, however, that the previous models are "wrong." The value of a model can only be judged based on the questions one is asking. If the question is, what are the slowest dynamics in the data, then an over-fit model may actually give a better estimate than a "perfectly-fit" model. However, if the question requires an inference of out-of-sample information, then we would trust the well-fit model over an over-fit one.

For protein simulations, the slowest timescale (typically corresponding to folding) has been observed to be insensitive to differences in MSM construction methodology [20**?** ]. We hypothesize that this behavior is due to the large structural changes associated with a folding event – even a two state model, given the same dataset, could estimate the slowest timescale by simply counting the number of transitions. However, teasing out the slow timescales that are associated with subtler structural changes is a different challenge.

Kellogg stuff.

### C. Limitations and Future Work

The likelihood described above allows for the comparison of MSMs with varying number of states, however it requires the definition of a vector space. As such, models built on different representations cannot be compared. This is because the likelihood requires the calculation of relative volumes, but the volumes will be different depending on what space you are working in. In addition, the lag time must be specified and cannot be selected using the likelihood. This is an improvement over the current state of constructing an MSM where the number of states is selected and then the lag time is tuned according to the eigenspectrum of the model. Selecting a lag time *a priori* is a much more intuitive parameter, that one typically knows based on the system.

We can also extend these results to models that do not have discrete states. Since all we require is an emission probability into a vector space, it is easy to use this likelihood for evaluating models that no longer have discrete states.

## VI. CONCLUSIONS

Markov State Models are a promising analysis technique for automatically analyzing simulations generated from Markovian dynamics. There are a number of steps in the construction process, however, that require human intervention, which limits the use of MSMs to experts as well as introduces significant uncertainty into the model building process. These results take a large step toward fully automating the construction process by providing

a quantitative way of selecting the number of states to use, as well as a way of selecting the clustering method. This will allow for non-experts to begin to use MSMs as well as allow adaptive sampling methods to be applied without human intervention.


[1] C. M. Dobson, Nature **426**, 884 (2003).
[2] S. Kim, B. Born, M. Havenith, and M. Gruebele, Angew. Chem. Int. Ed. **47**, 6486 (2008).
[3] R. H. Austin, K. W. Beeson, L. Eisenstein, H. Frauenfelder, and I. C. Gunsalus, Biochemistry **14**, 5355 (1975).
[4] I. Bahar, C. Chennubhotla, and D. Tobi, Curr. Opin. Struct. Biol. **17**, 633 (2007).
[5] G. Cosa, Y. Zeng, H.-W. Liu, C. F. Landes, D. E. Makarov, K. Musier-Forsyth, and P. F. Barbara, J. Phys. Chem. B **110**, 2419 (2006).
[6] X. Zhang, V. Q. Lam, Y. Mou, T. Kimura, J. Chung, S. Chandrasekar, J. R. Winkler, S. L. Mayo, and S.-o. Shan, Proc. Natl. Acad. Sci. U.S.A. **108**, 6450 (2011).
[7] H. D. Mertens and D. I. Svergun, J. Struct. Biol. **172**, 128 (2010).
[8] S.-R. Tzeng and C. G. Kalodimos, Curr. Opin. Struct. Biol. **21**, 62 (2011).
[9] P. Eastman, M. S. Friedrichs, J. D. Chodera, R. J. Radmer, C. M. Bruns, J. P. Ku, K. A. Beauchamp, T. J. Lane, L.-P. Wang, D. Shukla, T. Tye, M. Houston, T. Stich, C. Klein, M. R. Shirts, and V. S. Pande, J. Chem. Theory Comput. **9**, 461 (2013).
[10] M. Shirts and V. S. Pande, Science **290** (2000).
[11] D. Shaw, R. Dror, J. Salmon, J. Grossman, K. Mackenzie, J. Bank, C. Young, M. Deneroff, B. Batson, K. Bowers, E. Chow, M. Eastwood, D. Ierardi, J. Klepeis, J. Kuskin, R. Larson, K. Lindorff-Larsen, P. Maragakis, M. Moraes, S. Piana, Y. Shan, and B. Towles, in *High Performance Computing Networking, Storage and Analysis, Proceedings of the Conference on* (2009) pp. 1–11.
[12] B. Hess, J. Chem. Theory Comput. **4**, 116 (2008).
[13] P. L. Freddolino, C. B. Harrison, Y. Liu, and K. Schulten, Nat. Phys. **6** (2010).
[14] T. J. Lane, D. Shukla, K. A. Beauchamp, and V. S. Pande, Curr. Opin. Struct. Biol. **23**, 58 (2013).
[15] R. B. Best and G. Hummer, Proc. Natl. Acad. Sci. U.S.A. **107**, 1088 (2010).
[16] V. S. Pande, K. Beauchamp, and G. R. Bowman, Methods **52**, 99 (2010).
[17] K. A. Beauchamp, R. McGibbon, Y.-S. Lin, and V. S. Pande, Proc. Natl. Acad. Sci. U.S.A. **109**, 17807 (2012).
[18] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schutte, and F. Noe, J. Chem. Phys. **134**, 174105 (2011).
[19] G. R. Bowman, L. Meng, and X. Huang, J. Chem. Phys. **139**, 121905 (2013).
[20] R. T. McGibbon and V. S. Pande, J. Chem. Theory Comput. **9**, 2900 (2013).
[21] A. R. Liddle, Mon. Not. R. Astron. Soc. Lett. **377**, L74 (2007).
[22] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics (Springer New York Inc., New York, NY, USA, 2001).
[23] A. E. Gelfand and D. K. Dey, J. R. Statistic. Soc. B **56**, pp. 501 (1994).
[24] G. Schwarz, Ann. Stat. **6**, 461 (1978).
[25] F. Liu, D. Du, A. A. Fuller, J. E. Davoren, P. Wipf, J. W. Kelly, and M. Gruebele, Proc. Natl. Acad. Sci. U.S.A. **105**, 2369 (2008).
[26] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan, and W. Wriggers, Science **330**, 341 (2010).
[27] R. Kannan, L. Lovász, and M. Simonovits, Random Structures & Algorithms **11**, 1 (1997).
[28] M. Simonovits, Math. Program. **97**, 337 (2003).
[29] L. Lovász and S. Vempala, J. Comput. System Sci. **72**, 392 (2006).
[30] C. R. Schwantes and V. S. Pande, J. Chem. Theory Comput. **9**, 2000 (2013).
[31] T. J. Lane, G. R. Bowman, K. Beauchamp, V. A. Voelz, and V. S. Pande, J. Am. Chem. Soc. **133**, 18413 (2011).
[32] E. H. Kellogg, O. F. Lange, and D. Baker, J. Phys. Chem. B **116**, 11405 (2012).