Clark Center S295
Stanford University
Stanford, CA, 94305

Prof. Sharon Hammes-Schiffer
Deputy Editor of JPC B
The Journal of Physical Chemistry

Dear Prof. Hammes-Schiffer,

We are resubmitting this manuscript, whose title is now "Statistical Model Selection for Markov Models of Biomolecular Dynamics." The reviews were very helpful. Our detailed responses are included below, inline with the comments that they respond to.

Reviewer 1:

> In "Optimal parameter selection in Markov state models for biomolecular conformational dynamics", the authors elaborate on an earlier likelihood-based approach to the selection of the appropriate number of discrete states for crisp partitionings to attempt to obtain a Markov model that balances accuracy with the potential for overfitting that many-state models are capable of. Three different critera are examined: a leave-some-out cross-validation likelihood estimate, as well as two criteria that penalize overfitting by using assumptions of asymptotic normality of the likelihood function (AIC and BIC). Two different "emission probabilities" for the probability of observing a sample x_t from a given discrete state s are considered that would produce a generative model—one that can be used to assign new data to a model: a uniform probability and Gaussian probability. These are compared for consistency with each other, as well as compared with well-established metrics such as the stability of implied timescales of the model with the number of states.
>
> This manuscript will be a useful contribution to the literature on Markov state model construction, but has a number of deficiencies that hinder publication in its current form. Once these are addressed, it could be suitable for publication in JPC B.
>
> Most critically, the manuscript does not clearly explain the described work, fully justify the conclusions, or provide sufficient detail to reproduce the results. A number of points below help address these deficiencies:
>
> The title "Optimal parameter selection..." does not reflect the work contained in the paper, which focuses only on the selection of the number of states. Should the authors wish to keep the general title, more work on selecting other kinds of parameters should be included; otherwise, the title should be amended to something like "Optimal selection of the number of states... using a likelihood approach..."

We agree with the reviewer, as our contribution focuses only on picking the number of states. To address this comment, we've changed the title to reflect this fact. Furthermore, we don't wish to make the suggestion that our model selection criteria are themselves optimal. They are merely one set of statistical tools from among many. For this reason, we've changed the title to "Statistical Model Selection for Markov Models of Biomolecular Dynamics."

> "As such, the eigenvalues, and correspondingly the relaxation timescales, should remain constant as the lag time increases." The eigenvalues of which matrix should remain constant with what parameter? The eigenvalues of $T(t)$ actually change with t. Perhaps you mean the eigenvalues of some implied rate matrix K where $T(t) = \exp[Kt]$?

"choosing models whose relaxation timescales have converged" Give definition for "relaxation timescales".

The reviewer is correct. We meant that the timescales will remain constant, but this means that the eigenvalues do change. We've added a much more detailed discussion of the background on MSMs including a discussion of the propagator. This text can be found in the subsection titled *Theory: Markov State Models*

"With s(x) as the function mapping conformations into the indicator function basis set" Unclear: do you mean s(x) is the integer index of the basis function that is nonzero (has support) at x? Please be precise. The lack of precision in this manuscript substantially hampers readability and comprehension.

We were too casual with our notation, but had meant that $s(\cdot)$ maps a conformation in phase space to an integer corresponding to the index of the state in the Markov model. We've changed this notation to $\sigma(\cdot)$, and reworded the relevant section to read:

A k-state Markov model is defined on a discrete set of states, $S = \{s_i\}_{i=1}^k$, where $s_i \subseteq \Omega$ and $\bigcup_{i=1}^k s_i = \Omega$. Furthermore, $s_i \cap s_j = \emptyset$ for all $i \neq j$. In words, every point in Ω is assigned to one (and only one) state in the MSM. Let $\sigma(\mathbf{x})$ be the function that maps a point $\mathbf{x} \in \Omega$ to the index i of the state in S such that $x \in s_i$.

Eq. 1: This neglects a term for the initial conditions x_0. Eq. 1 actually refers to the conditional probability $P[x_t|x_1]$. Is this the quantity you intend to work with, rather than $P[x_t]$? If so, why?

We have changed this, such that we work directly $P\big(\{\mathbf{x}_t\}_{t=1}^N \;\big|\; \{S, \mathbf{T}\}\big) d\mathbf{x}^N$ as opposed to the conditional distribution in our work. We do not, however, fit a separate starting probability distribution. Because our goal is to build models for equilibrium dynamics, we choose $P\big(\mathbf{x}_1|\{S,T\}\big) = \pi_{\sigma(\mathbf{x}_1)}\, P(\mathbf{x}_1|\sigma(\mathbf{x}_1))$ where π the stationary distribution of $\mathbf{T}$.

"By convention, we define $P(x|s(x))$ to be zero for all x such that $s(x) \neq s_i$" Be precise! This statement is meaningless. Do you mean "we define $P(x|i) = 0$ whenever $s(x) \neq i$?"

We've reworded this paragraph to be clearer. The text can be found after Eq. (8), where we emphasize that the likelihood function introduced is valid only when

$$P(\mathbf{x}|i) = 0 \quad \forall\, \mathbf{x} \text{ s.t. } \sigma(\mathbf{x}) \neq i. \tag{1}$$

"We propose two possible emission distributions, $P(x|s(x))$" Write this as $P(x|s)$ or $P(x|i)$, to make it clear that your generative probability model MUST be a probability over configurations x given a discrete integer index s or i.

Eq. 3: Rewrite as $P(x|s)$ instead of $P(x_t|s(x_t))$, since these are supposed to be generative distributions for x given state index s.

The reviewer's suggested notation is more clear. We've rewritten the emission distribution as

$$P(\mathbf{x} \mid i) = \frac{1}{V_i} \mathbf{1}_{s_i}(\mathbf{x}). \tag{2}$$

> "A more tractable alternative is a Gaussian emission model..." Why is this even anticipated to be a good model? Won't it force your states to have gaussianlike clusters in them? And isn't a gaussian infinite in support, while you SPECIFICALLY REQUIRE the state membership functions to have finite support because they form a crisp partitioning? Why are you not truncating this model at the state boundaries, which would affect your normalization of the probability density in Eq 5. This inconsistency seems like a major flaw in this work.
>
> "We assume that the overlaps are small..." Is this enforced by some condition on sigma and mu for neighboring states? Has this assumption been checked? This is certainly, absolutely, positively not the case for even the simple model in Fig 3A.

When we first added this emission model, our goal was to provide a protocol that would be practical for practitioners to apply in high-dimensional settings. As the scaling of the volume calculation makes the volume emission model challenging except in low dimensions, we looked for an approximation whose probabilities could be evaluated analytically. One simple first guess would be to use a gaussian distribution. However, the reviewer brings up several issues with this model that make it too imprecise to be used in practice. As such, we've removed all of the Gaussian emission model from this work. Future work might be able to utilize this model as a tractable variational approximation to a hidden Markov model with Gaussian emission distributions.

> On choice of emission probabilities, an obvious question: Why not choose $P(x|i) \propto \exp[-\beta U(x)]\delta_{i,s(x)}$, where $U(x)$ is the potential energy and β the inverse temperature? The normalization constants are ALSO difficult to compute, but have the advantage that they are proportional to the stationary probability elements computed from the transition matrix T.

This is certainly one possibility, and is desirable because of its physical motivation. Our interest, however, is primarily in statistical models for molecular dynamics data that makes use of the potential function only by way of MD trajectories, and doesn't require further calculation based on the forcefield.

> "Unfortunately such approaches are intractable for problems of the size encountered in biomolecular simulations because of the need to integrate over all possible Markov models with a given number of parameters." Is this really intractable? That's a very strong claim. Perhaps it is just difficult or complex, since no evidence of intractability is presented?

In order to compute the Bayes factor for selecting the number of states, we would need to integrate over the space of all crisp state decompositions S with $|S| = k$ as well as the simplex of $k \times k$ reversible stochastic matrices. Although the second is possible analytically using the results of [**?**], we're not aware of any practical methods for the former. To avoid confusion, we've replaced the word "intractable" with "impractical."

> "N is the number of data points, assumed to be independent and identically distributed" How reasonable is this assumption for molecular simulation data? How do you even quantify the number of "iid normally distributed" data points in a molecular simulation?

This is a subtle point, and we've added some clarification to the main text, both in the subsection titled *Theory: Statistical Model Selection*, as well in the *Methods* section. Since we are parameterizing a transition matrix, we choose the number of data points to be the number of transitions used to calculate the transition probabilities. However, as we are using the sliding window approach for counting transitions, we are in fact over-counting the number of identical transitions and so the BIC score is likely under-penalizing the models. This is a weakness of the BIC for correlated timeseries applications, and is also a concern when estimating the transition matrix and its uncertainty from the observed transition counts[**?**].

> "For cross validation, we find it essential to use a prior..." Give complete equation for log likelihood. How do you initialize and sample posterior? Or maximize likelihood?

The prior refered to in the initial manuscript is in fact a psuedocount used during the estimation of $\mathbf{T}$. This is now clarified in the *Methods* section.

> "We simulated two trajectories with ... Langevin dynamics..." Which langevin integration scheme was used! How were the initial conditions selected?
>
> "The first trajectory was clustered using the k-centers clustering algorithm..." Cite. What parameters were used for clustering?

We've expanded our discussion of the simulation methods as well as the parameters used in MSM construction to address these and other issues. The new subsection, entitled *Simulation and MSM Construction* can be found in the *Methods* section.

> "The volumes did not change drastically..." What does this mean, quantitatively?
>
> "...the likelihood remained essentially constant..." What does this mean, quantitatively?

We selected M, the number number rounds of Monte Carlo volume estimation, such that the log likelihood was converged with respect to M.

> "according to the three methods" Which three methods exactly?

We've clarified this to read:

> according to the BIC, AIC, and test log-likelihood

> Fig. 2: This figure is way too small. What is "score"? Is this a log likelihood? Is lower or higher score representative of lower or higher likelihood? The numbers/units of likelihood are actually meaningful, and should be shown. What is the difference between "volume" and "Gaussian"?

All three figures have been reworked and are bigger. The y-axis label has been changed to "Selection Criterion." As discussed by Liddle and others, "the absolute value of the criterion is not of interest, only the relative value between different models."[**?**].

> Fig. 2 caption: "form the boundaries of the optimal window" Why is this region the optimal window? It looks like the optimal BIC window is 100-200. And why not just pick THE single optimal number of states? Why do you need a window that contains highly nonoptimal models?

We are not attempting to provide the "best" model given the data, as there are many statistical tests for doing this. The window illustrated in these figures represents some reasonable range of models, and is a useful visualization of the differences in the three tests. In practice, however, one would choose one of the three and then pick the best model given that test.

> "The AIC and BIC penalize the larger state models much more than the test set log-likelihood..." Why is this discrepancy so large? Doesn't this suggest AIC/BIC are unsuitable if they are not representative of the actual out-of-sample performance on the test set?

It's quite possible that the AIC and BIC are over-penalizing the likelihood. Given this, we would always suggest that one uses the test log-likelihood approach when possible. However, the "best" models according to the AIC and BIC were close to the best model from the test log-likelihood, so they may still be useful when it is impossible to leave out any data during the parameterization step. We are by no means advocating that either the AIC or BIC are perfect approximations to the out-of-sample performance.

Additionally, it's worth noting that the AIC applied to the tICA MSMs for WW domain does qualitatively agree with the test-set log likelihood.

> "causes the transition matrix to suffer" Clarify?

This original text has been removed, though the subsection entitled *Theory: Bias-Variance Dilemma* discusses the tradeoff we were referring to in the original draft.

> For Mueller potential with Gaussian likelihood: Would be useful to show log-likelihood as a function of number of states so we can see how much the model prefers 5 states.

As discussed above, we've removed the Gaussian likelihood results.

> Fig. 3: "Since k-means optimizes an objective function that is related to the emission probability" What is this objective function? As I understand it, k-means simply minimizes the sum of intracluster variances, which has nothing to do with your choice of emission probability being uniform or Gaussian or otherwise. Or do you mean you optimize the log-likelihood form of Eq 1 using a voronoi decomposition based on generators that are optimized for each number of states? Also: "The likelihood decreases as new states are added..." Which likelihoods are printed in the figure panels? Test set, uniform, or Gaussian emission probabilities?

The k-means clustering method minimizes the average mean squared distance from each data point to its cluster's mean. This is precisely the same as the $\hat{\sigma}^2$ term in the original equation. One can show that the resulting log-likelihood (evaluated on the data set) includes a term proportional to $\log \hat{\sigma}$. This is what we meant by the k-means clustering optimizing the likelihood function (and is precisely why **?**] selected the probability that they did).

This discussion, however, has been removed, since it only applies to the Gaussian likelihood.

"...AIC is comparable to the test set log-likelihood...The BIC penalizes complexity more strongly..." Why the huge difference between the AIC and BIC?

The difference just comes from the difference between the way both are defined. The BIC penalizes the number of parameters more strongly. We've added a discussion of the origins of the BIC and AIC. It's important to note that we are not attempting to advocate the use of a specific criterion, but are more interested in convincing the reader to use a statistical model selection criterion. The gold standard, of course is a cross-validation approach. Since this is not practical for most applications, we also wanted to include other selection criteria that are easier to calculate. Given our results, we would suggest the use of the AIC, but it's always important to know where these criteria come from and what assumptions are implicitly made when using one.

Fig. 4: Too small. Missing figure title, like "Comparison of likelihood models for Fip35 WW domain trajectory data". "Test set log-likelihoods are omitted in C as the dense matrix algebra is intractable for $k > 10^3$". Intractable? Really? But this is just summing log T_{ij} over N elements, isn't it?

We did not intend to use the word "intractable" in to indicate a particular computational complexity class. We've changed this word to "impractical" to avoid confusion.

"For tICA, this window is consistent with the convergence of the relaxation timescales" So....looking at convergence of the relaxation timescales is just as good as the new proposed likelihood method?

Yes it is. For well-sampled datasets built using the state-of-the-art methods, the implied timescales, which consider only the systematic error in the MSM, yield consistent recommendations with the AIC, BIC and test log-likelihood. In other regimes, where the statistical error dominates, the results are different.

"Lane et al. proposed...whereas Kellogg et al...Consistent with this spread, we find the optimal number of states to be highly dependent on the space in which the data is clustered." Were these two examples using the same metric of model quality and criteria for selecting the number of states?

This discussion is no longer in this revision since we've removed the Gaussian emission model. However, in case the reviewer was curious, Lane et al. selected the number of states by insisting that each point was within 4.5 Å of the cluster center. Kellogg et al. used a likelihood (Eq. (11) in the revision) based on a discriminative probability model.

"These results indicate why likelihood-based model selection is superior to approaches relying only on the timescale convergence." How? I don't see clear evidence of this. Which likelihood model is superior specifically? How is this superior to choosing the smallest number of states for which the timescales become flat (assuming that such a thing exists)? The only evidence presented here has judged success by examining concordance among AIC/BIC/leave-some-out cross validation and the timescale metric. In Fig 4, BIC and AIC disagree, while AIC tracks (but does not quantitatively agree with) the test set method. The agreement with the timescale convergence is shown as evidence of correctness. But it seems like SOMETHING is missing to show that this approach (or, more properly, ONE of these likelihood approaches) is superior to

> simply checking when the timescales go flat. A much more convincing test would be to show that these methods allow one to achieve smaller model error (eg in reproducing some quantitative features of the dynamics—maybe even the timescales?) for some smaller quantity of data when compared with a "gold standard" enormous dataset. The behavior of the optimal number of states as a function of the quantity of data would also be of great interest.
>
> "These results take a step toward fully automating the MSM construction process...by providing a quantitative method for selecting the most suitably complex model." Which method in particular is recommended? The Gaussian method? AIC or BIC? I'm still confused about what the actual, concrete demonstration of superiority is and which method the authors suggest should be used.

Our position in this manuscript is that in many cases the implied timescale tests, which are motivated by arguments made in the limit of infinite data, are insufficient for building MSMs on available data sets. We frequently observe that the timescales are not well-behaved for many large scale simulations. Therefore, model selection based on timescale convergence is not able to answer the question "given the data that I have, what is the best MSM I can build?" In this regime, both the statistical and systematic sources of error must be taken into account simulataneously. The likelihood-based criteria are a few examples of such methods, but they may not be the gold standard for model selection.

The statistical tools we've suggested match the timescale tests when the dataset is large and "well-behaved." However, there are many cases where the timescales do not converge at all and so the timescale tests fail to pick a reasonable MSM, which we illustrate in Fig. 4.

> Minor issues that the authors should address are below: Abstract: "in regimes in which these techniques encourage over-fitting" It is unclear precisely what "these techniques" refers to.

This is a good point, and we've rephrased to say:

> Application of techniques that consider both systematic bias and statistical error on two $100\mu s$ molecular dynamics trajectories of the Fip35 WW domain shows agreement with existing techniques based on self-consistency of the model's relaxation timescales, with more suitable results in regimes in which those timescale-based techniques encourage over-fitting.

> "Two key quantities which define the MSM are thus the state definitions, an indicator function basis over phase space..." There is a substantial quantity of literature on the use of basis functions that are not indicator functions (i.e. "fuzzy" rather than "crisp" partitionings) that should be mentioned. See, for example, the work of Marcus Weber and Susanna Roeblitz (nee Kube).

We've added a comment to the *Likelihood of a Markov State Model* section, discussing "fuzzy" clustering.

> The model has the training data set as its support rather than phase space, which makes it unable to generalize and assign probability to new data." It may also be worth noting that this is not a generative model, since the next paragraph states that a generative model is desired.

This is precisely what we are drawing attention to, and we've added to this section (*Likelihood of a Markov State Model*) to explicitly say this is a discriminative model.

> "but is intractable in high dimensions when the states are defined by general polytopes, such as those produced by a data-driven Voronoi tessellation" Be clear that *exact* volume computation is intractable in very high dimension due to the complexity of all known algorithms. Monte Carlo schemes (such as the one used here, or methods such as [http://www.cs.elte.hu/ lovasz/vol4-focs-tr.pdf]) allow the approximation of this volume. However, it's also worth noting that the Gaussian model you propose in Eq. 4 is ALSO INTRACTABLE for EXACTLY THE SAME REASON, since you do not compute the true normalizing constant for essentially identical reasons of computational tractability in high dimension. As such, phrases such as "A more tractable alternative is a Gaussian emission model.." are entirely disingenuous and misleading.

The reviewer is exactly right. For these reasons (and others) we've chosen to discuss only the volume-based emission model in this revision.

> "All MSMs were built using a lag time of 50ns" Why 50 ns?

This was the same lag time that was used in Beauchamp et al.[**?**]. This has also been specified in the methods section. In this scheme, the lag time must be specified before selecting the models, and so we picked a number that had been successfully used in the past. It's worth noting, however, that our scheme does not provide a framework for the selection of the lag time.

> "identify the simulations folding process" Something isn't quite right with this phrase. What is being referred to?

This sentence was poorly worded, but we've reworked this section, and so this should be clearer.

> "the relaxation timescales can continually increase with the number of states used" Even in the limit of an infinite quantity of data, the eigenvalue error bound theorems from Schuette et al. note that this behavior is precisely what we should expect, since the approximation error decreases as the eigenfunctions are better approximated via more states. However, the statistical error grows, though this is not quantified or plotted here.

The reviewer is exactly right, and this is precisely the point of our contribution. One advantage of the test-set likelihood approach is that it consolidates both the systematic and statistical contributions to the error into a single computable function.

Reviewer 2:

> The authors present an alternative approach for choosing the Markov time for Markov state models that will be of value for automating the construction of these models by removing the need for subjective user input. I have a few concerns, below.
>
> Line 34: Did you mean sacrificing "accuracy" instead of "complexity"?

We do mean complexity here, however the two words go hand in hand. The complexity of a model can be generally thought of as how capable the model is at handling arbitrarily complicated data. As one increases the complexity of a model, it is more able to model the training set more accurately.

> Can the authors comment on whether k-means or k-centers is preferable based on their Muller potential results?

This is an interesting question, but we feel it is beyond the scope of this work. However, the approach based on likelihood functions should be able to be used to study this question.

> The authors have access to many other data sets from their own lab and the Shaw group. Would it be possible to show results for other systems (maybe one larger system and one system with more disorder)? I think this would help to establish the generality of their results.

This is a very good idea and something we're interested in doing. However, due to the complexity of the volume calculation it is a difficult task. We hope that our work serves as a stepping stone to new methods that optimize a likelihood supported in phase space, but admit it may not be the one discussed here.

Reviewer 3:

Many of the comments made by Reviewer 1 and Reviewer 3 were closely related, and our responses reflect the consolidated concerns of both reviewers.

> Optimal Parameter Selection in Markov State Models for Biomolecular Conformational Dynamics by McGibbon, Schwantes and Pande. I enjoyed reading this paper. It is well suited for Bill Swope's birthday issue. It addresses an important problem in chemical physics, namely choosing amongst the various possible parameter settings when constructing a Markov State Model. However, the following general points should be addressed in a revision:
>
> I find the title Optimal Parameter Selection in Markov State Models... inappropriate: It suggests that the "Parameter selection in MSMs" is a well-defined problem, and that the solution approach here would be optimal, i.e. not further improvable. Neither is true. I suggest to pick a title that is specific to the content of the paper, e.g. "Likelihood-approach for selecting the number of states / choosing between discretizations in MSMs"

As discussed above, we've changed the title to "Statistical Model Selection for Markov Models of Biomolecular Dynamics."

> The paper starts with a discussion of approaches that choose the lag time based on implied timescales, and then proposes a way to choose the number of discrete states as an alternative. I can see the idea behind it, but to general reader it will probably be unclear what the connection between the two is. It should be spelled out that there is a change of viewpoint, where instead of fixing the discretization and varying the lagtime, one fixes the lagtime and varies the discretization. And the hope why both approaches are meaningful is that the implied timescales should converge to their true values for either (1) increasing lag time [see Djurdjevac, Sarich, Schtte, MMS 2011 and Prinz, Chodera, Noe, PRX 2014 (http://arxiv.org/abs/1207.0225)], or (2) better approximation of the dominant eigenfunctions via more discrete states [see (19) and Sarich, Noe, Schütte MMS 8, p1154 (2010)].

We've added significantly more exposition and background on the theory behind MSMs, and in particular the analysis of the error as a function of the number of discrete states.

> My main concern is that the paper suggests that the best result is obtained by maximizing the present likelihood. But stationary and dynamical expectation values such as the relaxation timescales are well defined by the underlying dynamics, and the ultimate objective of any estimation procedure must be to correctly approximate them. Using the present likelihood could systematically give wrong estimates for various reasons.
>
> An inappropriate statistical model for the local output function (which is really hard to make) would lead to biased estimates. Then there is uncertainty in choosing the Bayesian criterion - AIC, BIC or something else? If the lag time chosen is too short, it will not lead to correct estimates for feasible numbers of states.
>
> The differences in Fig 4 (see Details below) indeed suggests that the present criterion fails to provide the correct estimate in at least some cases. If a new estimator is derived, it needs to be shown that it converges to the correct result in some limit.

These are valuable points, which we fully agree with. Our focus is on best fitting the available data, and ideally validating models on new data which was left out during training. The ultimate goal is to fit the underlying dynamics, but our only "window" into the underlying data is by way of MD trajectories. Clearly convergence proofs, that MSM methods converge to the exact transfer operator in the limit of infinite sampling and infinite model complexity (i.e. $k \to \infty$) are essential, but these results do not provide direct guidance to practitioners on balancing competing sources of error.

> I suggest to build a fine discrete model, e.g. by direct gridding of the Müller potential, and constructing a simple MCMC dynamics on neighboring gridpoints, and then considering different clusterings on this data set. There, the exact timescales are known, and an embedding in a geometric space is given, such that the method can be applied such that the present method can be applied as is. It then needs to be shown that the present likelihood approach selects MSMs with timescales that coincide with the true values. In addition, it would be useful to compare the results here with implied timescale plots. Do the results agree in the cases where ITS do converge?

In the finite-data regime, our methods will not select MSMs with timescales that correspond to the "true values," precisely because there is a competing statistical error. Proofs of the consistency and asymptotic optimality of the AIC and BIC for linear regression do exist, which guarentee that as $n \to \infty$, the correct model will be selected with proability one[**? ?**]. Our intuition is that similar results may apply for MSMs in the joint infinite-data / infinite number of states limit, but we have not shown this. Given that our approach is focused on practical tools in the limited data regime, we prefer to focus there.

> Please check the timescales in Fig 2(A,B) and 4(A,B) - they look identical, but shouldn't. Details / specific points:

These timescales were correctly displayed. The timescales in A and B are calculated from the same set of models, but the evaluation of those models just provides a different "score." The results with the Gaussian emission distribution have been removed.

> "ever-more states". Ever-more states or ever-larger lag time? (when did number of states come in?)
>
> "Often the choice is made to use the model whose timescales better match an experimental observable:" I hope that's not true and I wouldn't generally accuse the community of practicing this. When timescales don't converge, this is a numerical or statistical problem with the data or the estimators used. Ignoring this fact and deliberately choosing a model that coincidentally delivers the numbers one hopes to get would be simply cheating.

The reviewer is definitely correct here. We are not accusing anyone of "cheating" but more drawing attention to the fact that with the absence of an optimizable objective function, a variety of human factors may come into play. We've nonetheless removed this comment from the paper.

> Gaussians could be truncated and normalized - it seems that renormalization would require to solve an even harder problem than for uniform output probabilities, because youd need to integrate the Gaussian density over a high-dimensional polytope.
>
> "We assume that the overlaps are small" But are they small in your application? Ensuring this puts requirements on the sizes of states and sigma. How is it checked that this is then a good model?
>
> Are you sure that the timescales shown in Fig 2A(right) and 2B(right) are from different data? They look identical to me.

As discussed above, we've removed the Gaussian emission distribution.

> I'm surprised that there are three slow relaxation processes for the Müller potential before the second gap. The potential looks like there should be only two (due to three metastable states). Were MSMs build with reversible transition matrix estimation, and are all eigenvalues thus real-valued?

Yes, they are all real-valued.

> Müller potential: lag time was chosen to be 30 steps. How is this choice justified? Since the timescales in Fig. 2 are not independent of N, this must mean that at least for some state decompositions the ITS are not converged in the lag time. In these cases, it's not even appropriate to describe the state-to-state dynamics by a Markov chain at all. Is the statistical model still meaningful in this case?
>
> WW domain, lag time was chosen to be 50 ns - same comment as for the Mueller potential, above.

Our goal is to build the best possible Markov model given the data and a prespectified lag time. From the perspective of model selection based on implied timescale convergence with respect to lag time, it's usually required that the modeller have some physical intuition about the correct number of states. Here, we work from the opposite direction, as likelihood based approach gives no guidance on the selection of τ.

> Müller potential, Gaussian likelihood: How were the Gaussian sigma-parameters chosen? The k-means states in Fig 3 have very different sizes and their centers are sometimes very close, so the assumption that output distributions shall not overlap seems to be hard to realize.
>
> WW domain, Gaussian likelihood - same comment as for the Müller potential, above.

As discussed above, we've removed the Gaussian emission distribution.

> Are you sure that the timescales shown in Fig. 4A(right) and 4B(right) are from different data? They look identical to me.

They were supposed to be the same timescales, but different likelihoods. Now that we have removed the Gaussian emission distribution, this point of confusion should not be present.

> Doesn't Fig 4(c) suggest that there's a problem here? For the likelihood-based choice made, the slowest timescale is estimated between 1500 and 3000 ns with RMSD clustering. In contrast, the tICA-based estimates suggest a timescale of around 5000 ns. So if the tICA results is indeed correct, the method should be going for larger state number in the RMSD case.

This is a question of perspective. Our view is that even if the models based on RMSD or PCA underestimate the timescales, it nonetheless introduces an unacceptable amount of statistical error in the finite-data context to add states without limit.

> In some timescale vs number of state plots the timescales descreases with increasing number of states (e.g. Fig 2A). That's surprising. Can the authors explain that? Has a prior been used that may be responsible for this behavior?

We have no strong theortical justification for this observation, but we have seen slight decreases in the timescales with respect to k on occation, regardless of the method by which $\mathbf{T}$ was estimated (i.e. with and without psuedocounts).

Sincerely,

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

References

[1] Andrew R Liddle. How many cosmological parameters? *Monthly Notices of the Royal Astronomical Society*, 351(3):L49–L53, 2004.

[2] Ryuei Nishii. Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics*, 12(2):758–765, 06 1984.

[3] Dan Pelleg, Andrew W Moore, et al. X-means: Extending k-means with efficient estimation of the number of clusters. In *ICML*, pages 727–734, 2000.

[4] Jan-Hendrik Prinz, Hao Wu, Marco Sarich, Bettina Keller, Martin Senne, Martin Held, John D. Chodera, Christof Schütte, and Frank Noé. Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.*, 134(17):174105, 2011.

[5] Jun Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 7(2):221–242, 1997.