Supplemental Information for

On the Removal of Initial State Bias from Simulation Data

Marco Bacci, Amedeo Caflisch, and Andreas Vitalis

Inventory

I. Supplemental Figures S1-S10.

I. Supplemental Figures

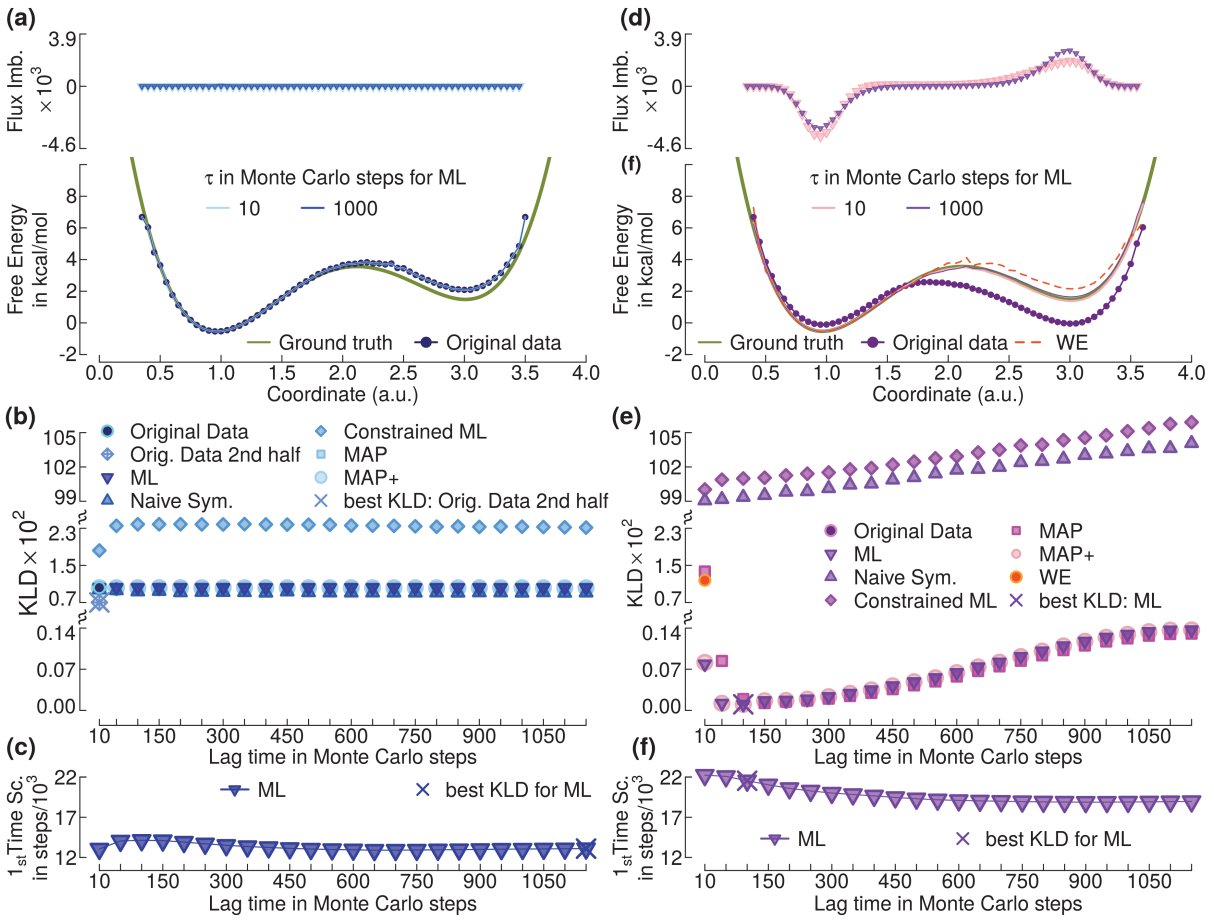

Figure S1: This is the same as Fig. 3 in the main text only that the discretization is done using regular space binning at a coarser resolution (0.05 a.u. bin width). Note that, because of the coarser resolution, the deterioration of the performance of the MAP estimator at short lag time is less pronounced. Conversely, the ML and MAP+ estimator used in conjunction with the shortest lag time exhibit small errors, which are likely due to issues with Markovianity (compare Fig. S2).

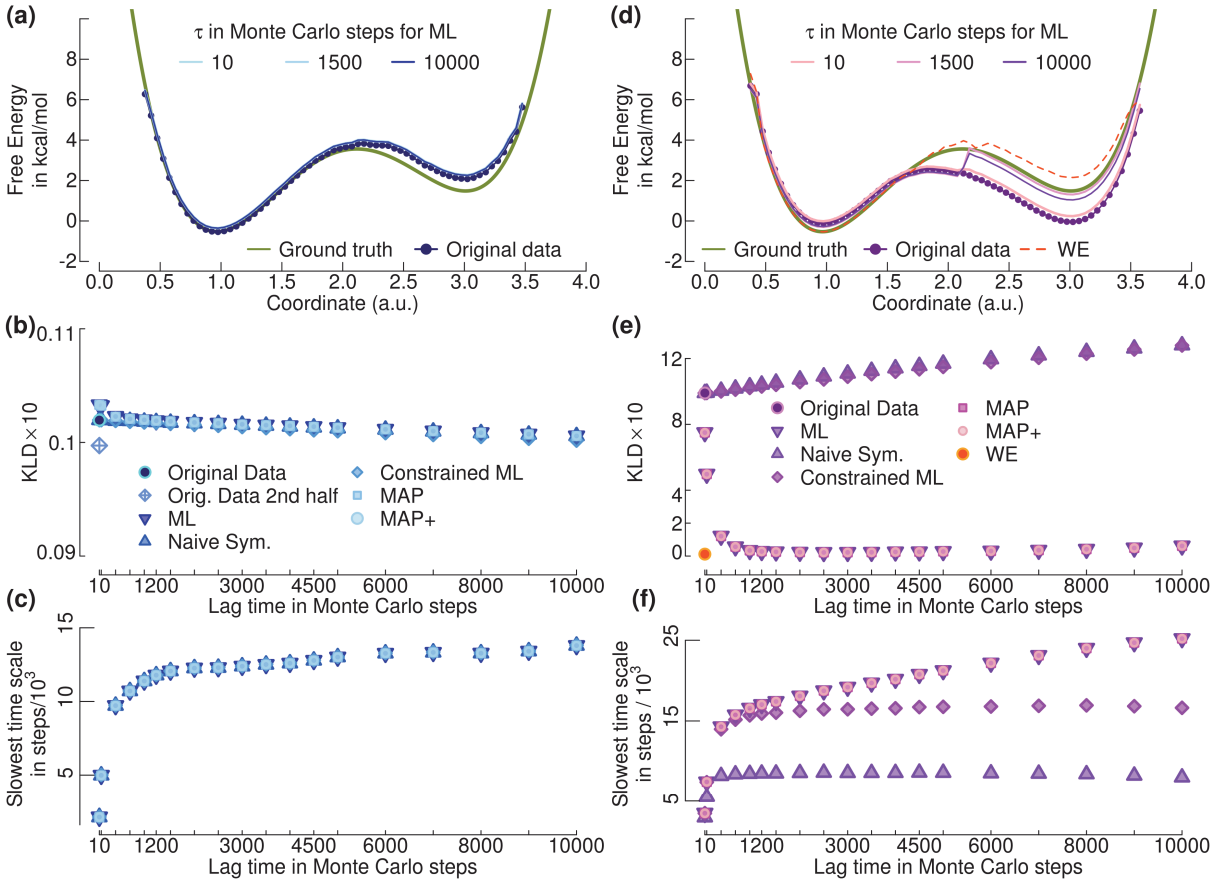

Figure S2: This is similar to Fig. 3 in the main text and to Fig. S1. Here, the discretization is done using only two states (smaller and larger than a coordinate value of 2.12, respectively). This is the maximally coarse MSM one can construct for this system. **(a)-(c)** Data for 16 trajectories of 10^5 steps each with no reseedings (CS). The ground truth along with the potential inferred from the observed distribution is displayed in (a). Potentials from ML reweighting are shown as well but the lines superpose with the observed data. Panel (b) shows KL divergences from the ground truth for all attempted reweighting strategies and lag times (note the y-axis range). Finally, (c) shows the value of $-\tau / \ln \lambda_2$ where λ_2 is the 2nd largest eigenvalue of the ML estimate of $\mathbf{T}(\tau)$ and τ is the lag time. Here, this corresponds to the time to cross the barrier between the two states. Because the system is at equilibrium (but poorly sampled), different inference strategies do not affect the results significantly. **(d)-(f)**. The same as (a)-(c) for a PIGS data set of identical extent. In addition to the analogous data shown in (a)-(c), the WE result, which is independent of any discretization or lag time, is included in panels (d) and (e). In panel (d), the "step" in the reweighted distribution is because the underlying MSM is so coarse. Despite this, the relative weight of the two states is captured accurately at appropriate lag times. Note the y-axis scale in (e), also in relation to (b). In panel (e), ML, MAP, and MAP+ all overlap and exhibit errors at short lag times, which are most likely due to Markovianity violations. Panel (f) shows the plateauing of the slow mode with lag time, which is especially robust for the detailed balance-imposing models. This should be juxtaposed with the failure in applying these models to the reweighting problem as seen in panel (e).

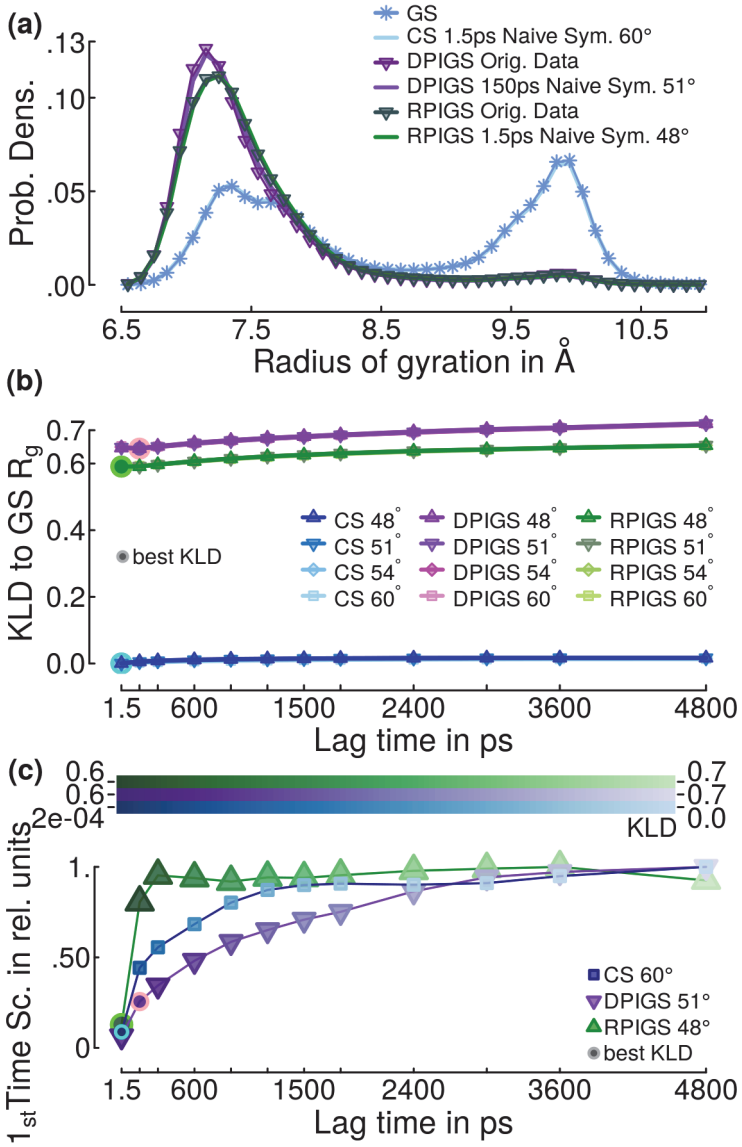

Figure S3: This is the same as Fig. 5 in the main text but for a naïve imposition of detailed balance on **T**, *i.e.*, letting $c_{ij}^{*} = c_{ji}^{*} = \max\left(c_{ij}, c_{ji}\right)$. The predicted steady states are essentially the same as the raw sampling weights, (a) and (b), but the slowest time scale is well-behaved, (c). Note that all MSM predictions overlap with their respective original data in (a) and that all resolutions for a given data set overlap in (b).

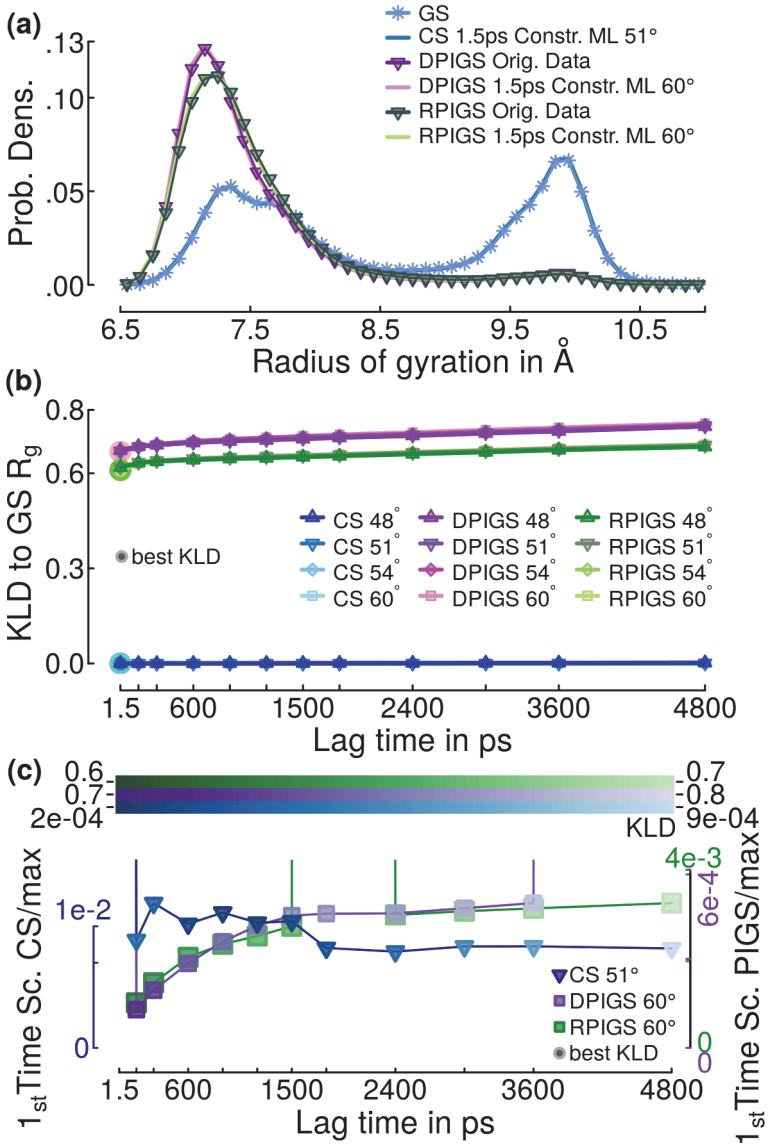

Figure S4: This is the same as Fig. 5 in the main text but for a ML inference of **T** under a detailed balance constraint (see eq. (2) in the main text). The predicted steady states are essentially the same as the raw sampling weights, (a) and (b), and the slowest time scale is prone to dramatic outliers occurring without an identifiable pattern, (c). Note the individual y-axis scales in (c): the plot is truncated to facilitate visualization since the maximum relaxation time scale for outliers exceeded the base level by factors of 100 or larger. Note that all MSM predictions overlap with their respective original data in (a) and that all resolutions for a given data set overlap in (b).

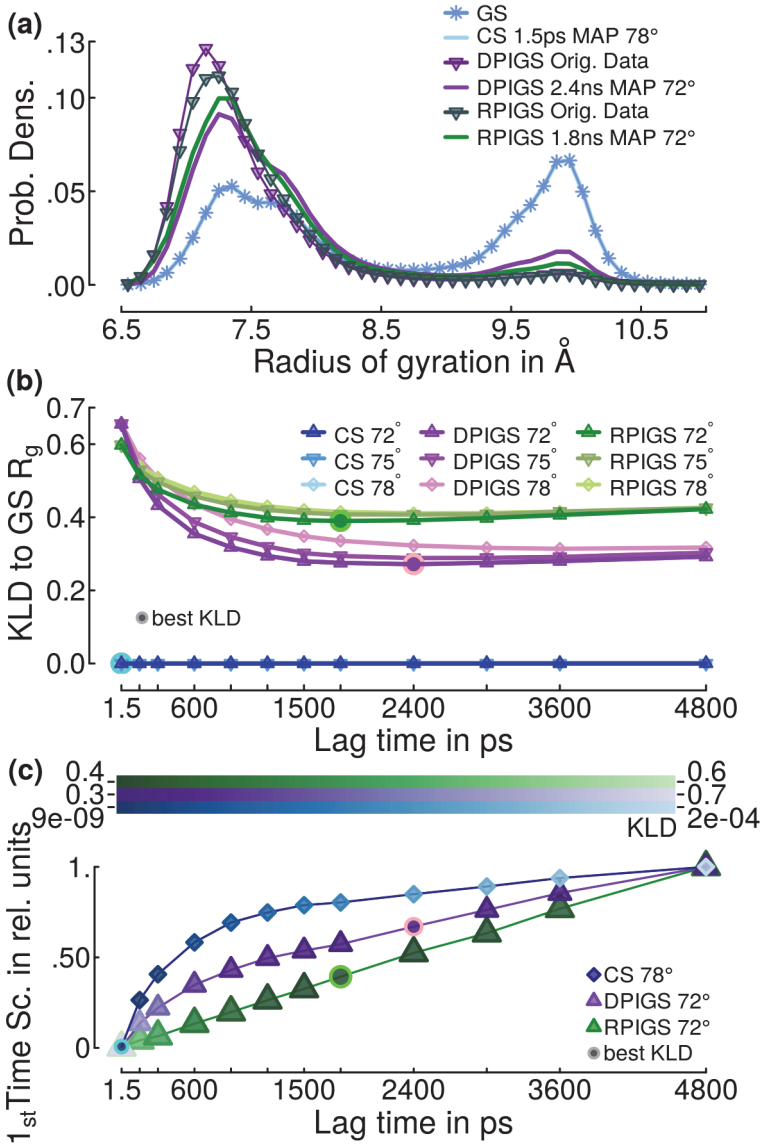

Figure S5: This is the same as Fig. 5 in the main text but for the MAP estimator (symmetric Dirichlet prior per row with $\alpha = 1 + 1/N$ where N is the number of states, see eqs. (3) and (5) in the main text). Because the prior creates a maximally dense transition matrix, we could not calculate properties for comparable clustering resolutions (numerical bottleneck with memory and compute time). Note that the CS and GS data overlap in (a) and that all CS data overlap in (b).

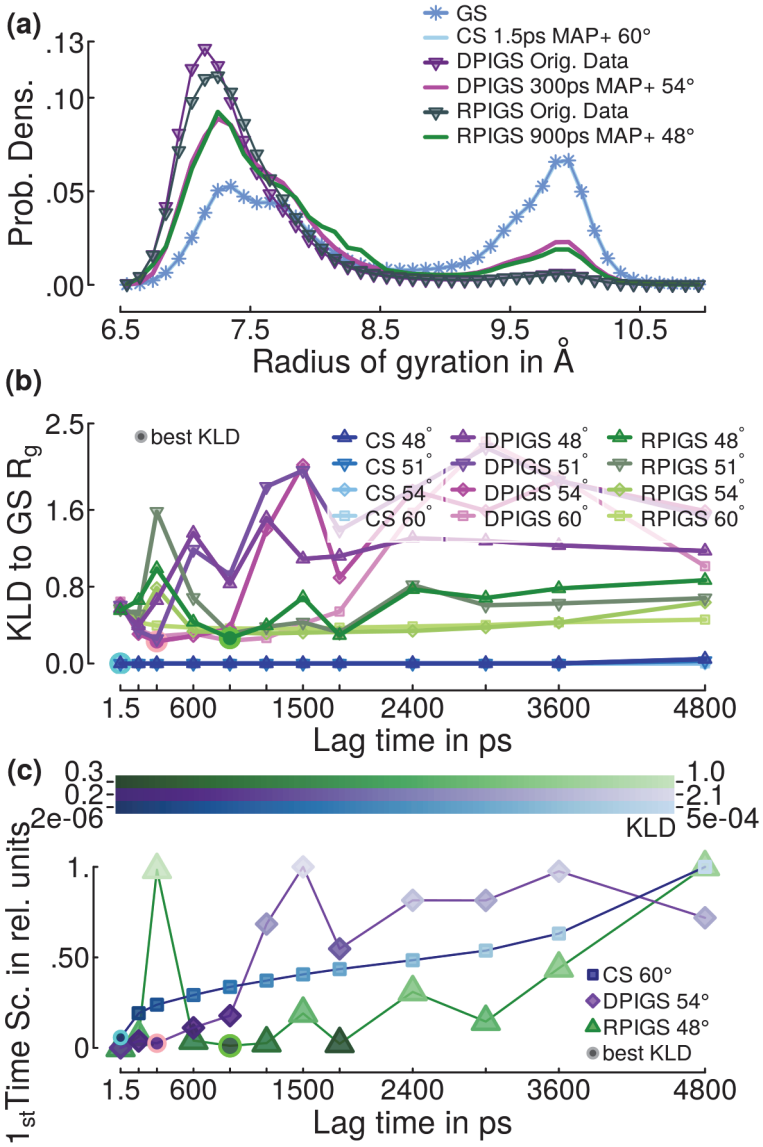

Figure S6: This is the same as Fig. 5 in the main text but for the MAP+ estimator (nonsymmetric Dirichlet prior, see eq. (11) in the main text). The peak performance results are similar to those in Fig. 5, but the trends with clustering resolution and resolution are more (rather than less) erratic than those for the ML estimator. This seems to be a particular problem for the DPIGS data set. The relevance of prior information increases with increasing lag time and finer resolution. Evidently, this does not give rise to a consistent improvement across conditions, which could be linked to the fact that the shape of the prior is data-derived. Note that the CS and GS data overlap in (a) and that all CS data overlap in (b).

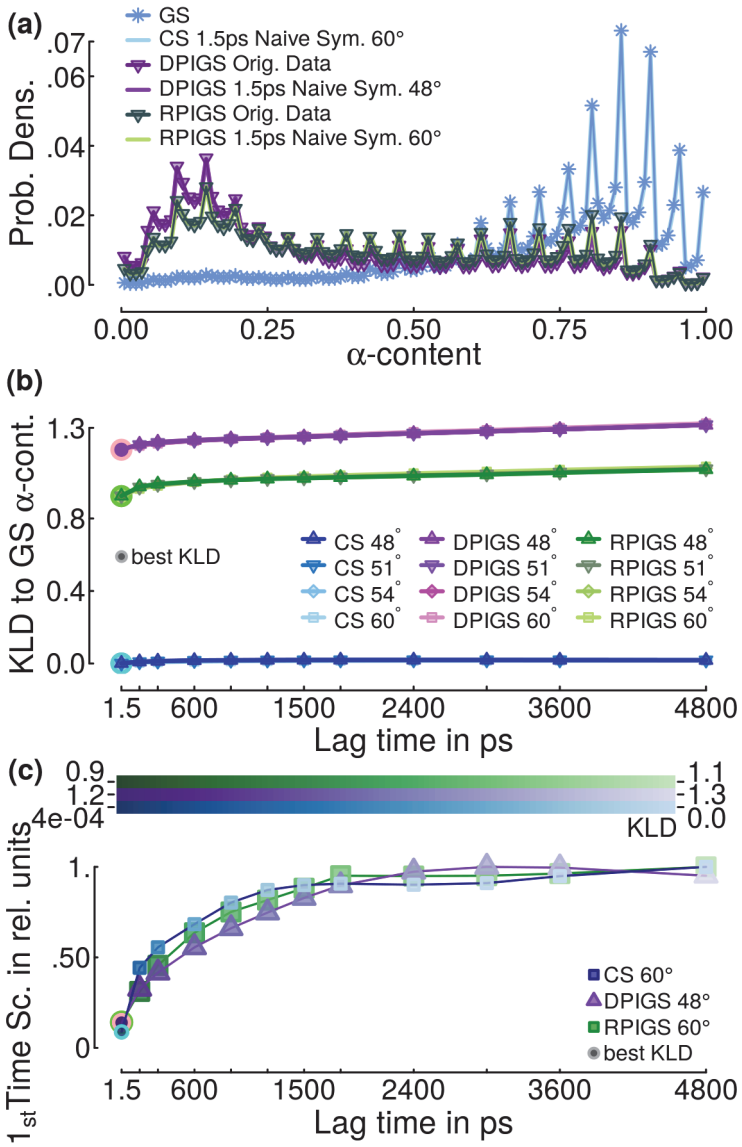

Figure S7: This is the same as Fig. 6 in the main text but for a naïve imposition of detailed balance on **T**, *i.e.*, letting $c_{ij}^{*} = c_{ji}^{*} = \max\left(c_{ij}, c_{ji}\right)$. Note that all MSM predictions overlap with their respective original data in (a) and that all resolutions for a given data set overlap in (b).

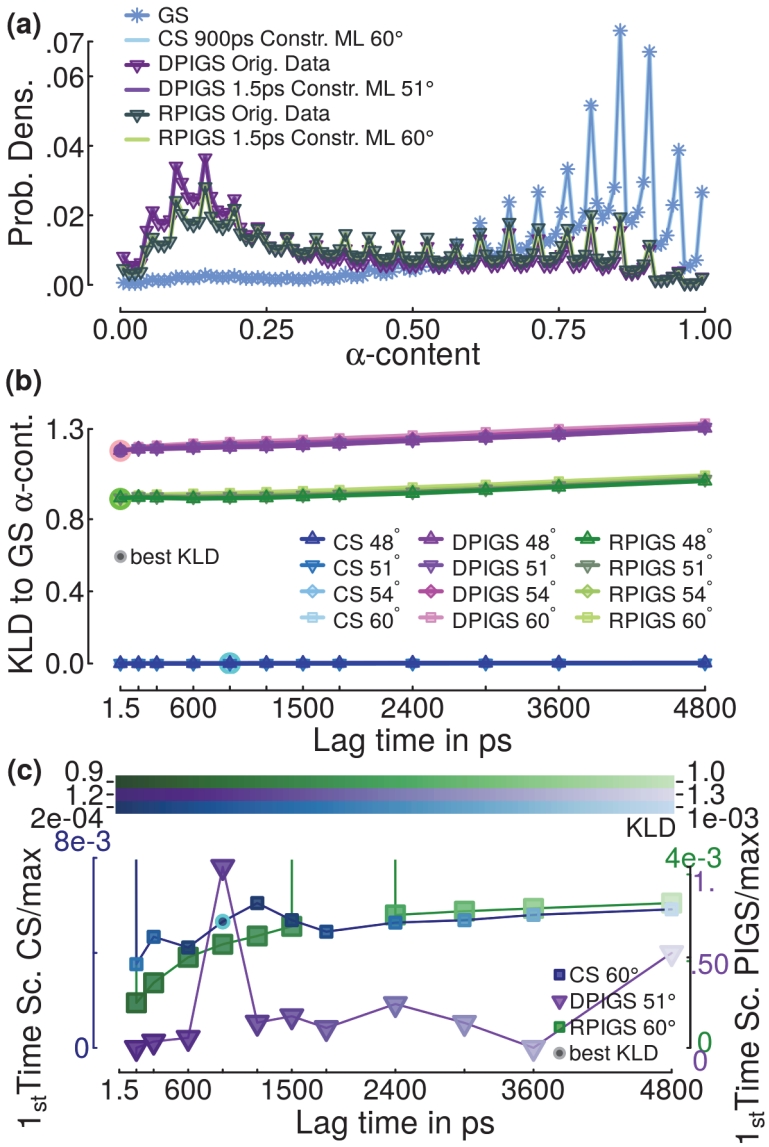

Figure S8: This is the same as Fig. 6 in the main text but for a ML inference of **T** under a detailed balance constraint (see eq. (2) in the main text). Note that all MSM predictions overlap with their respective original data in (a) and that all resolutions for a given data set overlap in (b).

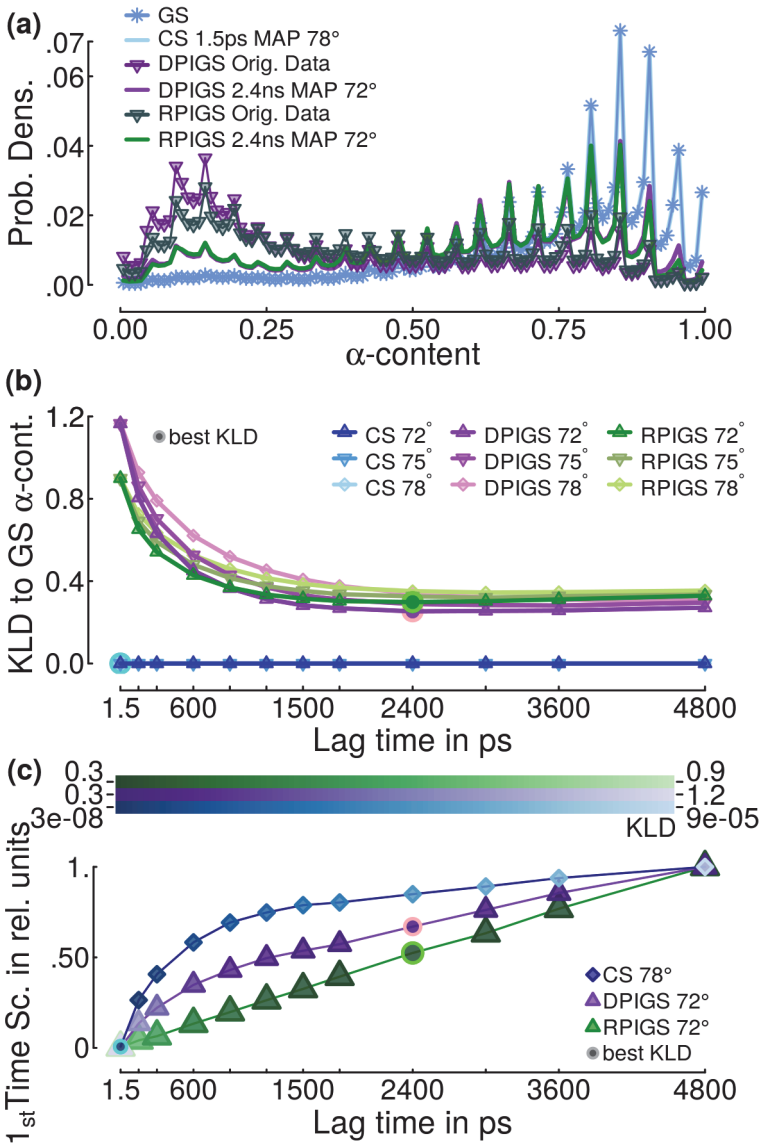

Figure S9: This is the same as Fig. 6 in the main text but for the MAP estimator (symmetric Dirichlet prior per row with $\alpha = 1 + 1/N$ where N is the number of states, see eqs. (3) and (5) in the main text). Note that the CS and GS data overlap in (a) and that all CS data overlap in (b).

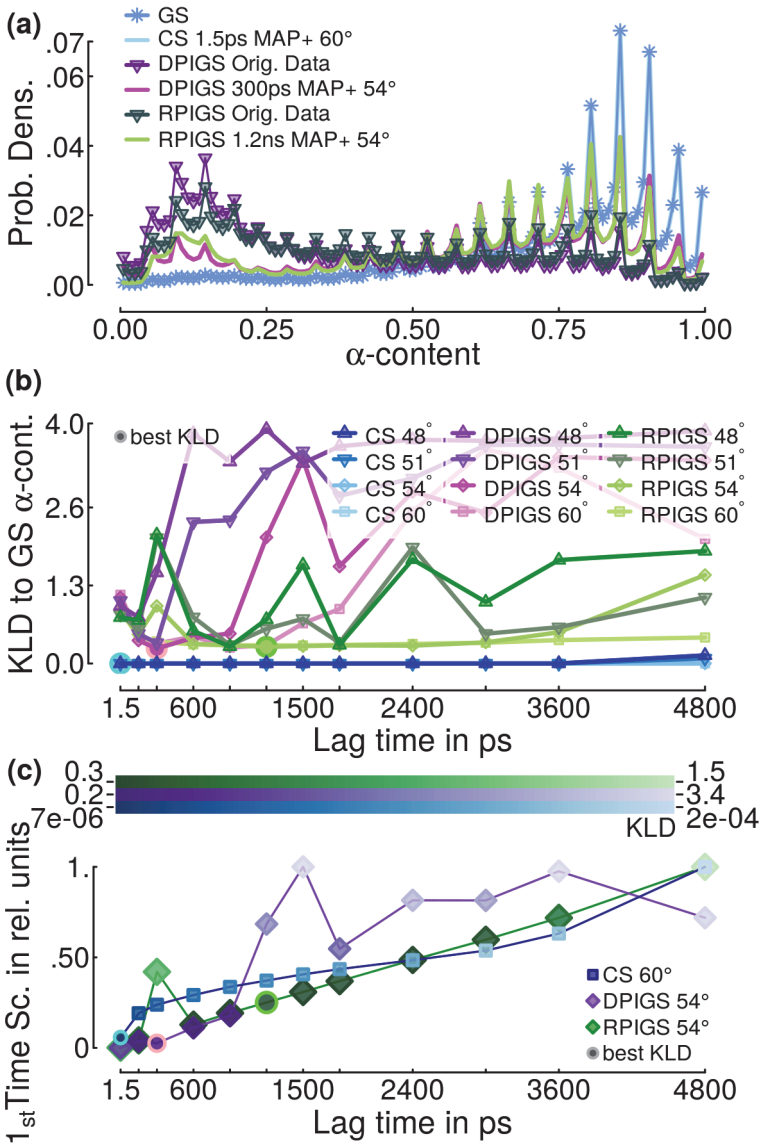

Figure S10: This is the same as Fig. 6 in the main text but for the MAP+ estimator (nonsymmetric Dirichlet prior, see eq. (9) in the main text). Note that the CS and GS data overlap in (a) and that all CS data overlap in (b).