the state histories $\mathbf{S}$ as an auxiliary variable, sampling from the augmented posterior

$$P(\boldsymbol{\Theta}, \mathbf{S}|\mathbf{O}) \propto P(\boldsymbol{\Theta}) \times \prod_{n=1}^{N} \rho_{s_0^{(n)}} \varphi(o_0^{(n)}|\mathbf{e}_{s_0^{(n)}}) \prod_{t=1}^{T^{(n)}} T_{s_{t-1}^{(n)} s_t^{(n)}} \varphi(o_t^{(n)}|\mathbf{e}_{s_t^{(n)}}) \quad (10)$$

If we presume the prior is separable

$$P(\boldsymbol{\Theta}) \equiv P(\mathbf{T})P(\mathbf{E}) \quad (11)$$

we can sample from the augmented posterior (Eq. 10) using the framework of *Gibbs sampling* [15], in which the augmented model parameters are updated by sampling from the conditional distributions,

$$P(\mathbf{S}|\mathbf{T}, \mathbf{E}, \mathbf{O}) \propto \prod_{n=1}^{N} \rho_{s_0^{(n)}} \varphi(o_0^{(n)}|\mathbf{e}_{s_0^{(n)}}) \prod_{t=1}^{T^{(n)}} T_{s_{t-1}^{(n)} s_t^{(n)}} \varphi(o_t^{(n)}|\mathbf{e}_{s_t^{(n)}})$$

$$P(\mathbf{T}|\mathbf{E}, \mathbf{S}, \mathbf{O}) \propto P(\mathbf{T}) \prod_{n=1}^{N} \prod_{t=1}^{T^{(n)}} T_{s_{t-1}^{(n)} s_t^{(n)}}$$

$$P(\mathbf{E}|\mathbf{S}, \mathbf{T}, \mathbf{O}) \propto P(\mathbf{E}) \prod_{n=1}^{N} \prod_{t=0}^{T^{(n)}} \varphi(o_t^{(n)}|\mathbf{e}_{s_t^{(n)}}) \quad (12)$$

When only the model parameters $\boldsymbol{\Theta} \equiv \{\mathbf{T}, \mathbf{E}\}$ or the hidden state histories $\mathbf{S}$ are of interest, we can simply marginalize out the uninteresting variables by sampling from the augmented joint posterior for $\{\mathbf{T}, \mathbf{E}, \mathbf{S}\}$ and examining only the variables of interest. In addition, the structure of the Gibbs sampling scheme above allows individual components (such as the observable distribution model $\varphi(o|\mathbf{e})$ or transition probability matrix $\mathbf{T}$) to be modified without affecting the structure of the remainder of the calculation.

[JDC: Talk about how insight can be extracted from parameters, such as lifetimes, rates, branching probabilities, etc.]

The Bayesian treatment equips us with both a model of the parameters given data (Eq. 9) and a model of the data given the parameters (Eq. 4), allowing an experimenter to use prior knowledge or preliminary experimental data to model the outcome of new experiments, and how the collection of additional experimental data can be expected to reduce model uncertainties. For example, suppose we have conducted an experiment $\mathcal{E}_1$ which yielded data $\mathbf{O}_1$. Using the information from this experiment, we can model the probability that a yet-to-be-performed experiment $\mathbf{E}_2$ will yield data $\mathbf{O}_2$,

$$P(\mathbf{O}_2|\mathcal{E}_2, \{\mathbf{O}_1, \mathcal{E}_1\}) = \int d\boldsymbol{\Theta} P(\mathbf{O}_2|\mathcal{E}_2, \boldsymbol{\Theta}) P(\boldsymbol{\Theta}|\{\mathcal{E}_1, \mathbf{O}_1\})$$

As a simple illustration of the utility in experimental design, we assume that a prior observation has been made to produce observed dataset $\mathbf{O}_1$, and that the distribution $P_2(\mathbf{O}_2|\boldsymbol{\Theta})$ describes the probability of observing some data $\mathbf{O}_2$ (from a potentially different observable) given the model parameters $\boldsymbol{\Theta}$. Based on the information gathered from the first observation $\mathbf{O}_1$, the *expected* information content of the second experiment to collect $\mathbf{O}_2$ can be written as

$$E[I(\mathbf{O}_2|\mathbf{O}_1)] = H[P_1(\boldsymbol{\Theta}_1|\mathbf{O}_1)] - \int d\mathbf{O}_2 \, H[P_2(\boldsymbol{\Theta}_2|\mathbf{O}_2, \mathbf{O}_1)] \int d\boldsymbol{\Theta}_1 \, P_2(\mathbf{O}_2|\boldsymbol{\Theta}_1) P_1(\boldsymbol{\Theta}_1|\mathbf{O}_1) \quad (13)$$

where $H[P(\boldsymbol{\Theta})] \equiv -\int d\boldsymbol{\Theta} \, P(\boldsymbol{\Theta}) \ln P(\boldsymbol{\Theta})$ denotes the Shannon entropy or uncertainty of a distribution $P(\boldsymbol{\Theta})$. While direct computation of Eq. 13 can be challenging, approaches have ben developed to compute useful approximations for use in Bayesian experimental design [19]. [JDC: We need to come up with a more useful way to decide between new experiments, or quantify th information content of existing experimental data.]

III. ALGORITHMS

Below, we outline the algorithms we use for generating an initial model subject to prior constraints, computing a maximum-likelihood hidden Markov model (MLHMM), and sampling from the Bayesian posterior (BHMM).

[JDC: We should discuss how the experimenter can determine an appropriate Markov time Δt for a given dataset.]

A. Generating an initial model

To initialize either computation of the MLHMM or sampling from the posterior for the BHMM, an initial model that respects any constraints imposed in the model prior $P(\boldsymbol{\Theta})$. Here, we employ a Gaussian observable distribution model for $\varphi(o|\mathbf{e})$ (Eq. 3) and enforce that the transition matrix $\mathbf{T}$ satisfy detailed balance[2].

1. *Observable parameter estimation*

We first initialize the observed distributions of each state by fitting a Gaussian mixture model with M states to the pooled observed data O, ignoring temporal information:

$$p(\mathbf{O}|\pi, \mathbf{E}) = \prod_{n=1}^{N} \prod_{t=0}^{T^{(n)}} \sum_{m=1}^{M} \pi_m \varphi(o_t^{(n)}|\mu_m, \sigma_m^2) \quad (14)$$

where the state observable emission probability vector $\mathbf{E} \equiv \{\mathbf{e}_1, \ldots, \mathbf{e}_M\}$ and $\mathbf{e}_m \equiv \{\mu_m, \sigma_m^2\}$ with μ_m denoting the observable mean and σ_m^2 the variance for state m for the Gaussian observable model. The vector $\boldsymbol{\pi}$ is composed of equilibrium state populations $\{\pi_1, \ldots, \pi_M\}$ with $\pi_m > 0$ and $\sum_{m=1}^{M} \pi_m = 1$.

A first approximation to $\boldsymbol{\pi}$ and $\mathbf{E}$ is computed by pooling and sorting the observed $o_t^{(n)}$, and defining M indicator functions $h_m(o)$ that separate the data into M contiguous regions of the observed range of o of roughly equal population. Let $N_m \equiv \sum_{n=1}^{N} \sum_{t=1}^{T^{(n)}} h_m(o_t^{(n)})$ denote the total number of observations falling in region m, and $N_{\mathrm{tot}} = \sum_{m=1}^{M} N_m$. The

[2] Physical systems that are not driven by an external force or energy reservoir should satisfy detailed balance [20], and its use has been shown to provide a large reduction in transition matrix uncertainty in data-poor conditions [21].