

# Accelerated Sampling of Mutants: A Hierarchical Bayesian Markov State Model Strategy

Robert McGibbon and Vijay Pande

Department of Chemistry, Stanford University, Stanford, CA.



STANFORD  
UNIVERSITY

## Introduction

- ▶ Mutational analysis is the bread and butter of experimental protein biophysics.
- ▶ In simulations, mutational analysis is a major challenge, because the costs scale proportional to the number of mutants.
- ▶ The extensive commonalities between mutants is what makes the analysis *useful*.
- ▶ This suggests that large-scale simulations of a wild-type protein can be an **informative prior** on new simulations of a mutant.
- ▶ Our Ansatz: mutation perturbs the rates of interconversion between a unknown subset of the conformational states of a protein.

## Model

- ▶ We use a common discretization of the state space between the mutant and wild type.
- ▶ Let  $\vec{p}_i^M$  be the outbound transition probabilities from state  $i$  in mutant,  $\vec{c}_i^{WT}$  be the observed outbound transition counts from state  $i$  in the wild type, and  $q_i \in (0, 1)$ , the transfer coefficients, be the degree of information transfer between wildtype and mutant dynamics from state  $i$ .
- ▶ **Informative prior** on  $\vec{p}_i^{MT}$ :  

$$\vec{p}_i^{MT} \sim \text{Dirichlet}(q_i \cdot \vec{c}_i^{WT} + 1/2)$$
- ▶ When  $q_i = 0$ , we have Jeffreys prior, and when  $q_i > 0$ , counts are **inherited** from the wildtype into the mutant.
- ▶  $q_i$  must also be learned from the data. For convenience, the prior on  $q_i$  is Beta, with shared hyperparameters, which constrains  $q_i \in (0, 1)$ :

$$q_i \sim \text{Beta}(\alpha, \beta)$$

- ▶ Given observed transitions in  $MT$ , posterior distribution on  $p_i$ :

$$P(\vec{p}_i^{MT} | \vec{c}_i^{MT}) \propto$$

$$\int_0^1 dq_i \text{Dir}(q_i \cdot \vec{c}_i^{WT} + \vec{c}_i^{MT} + 1/2) \cdot P_{\alpha, \beta}(q_i)$$

## Methods

- ▶ Our sampling principle: choose actions that maximize the model's **expected information gain (EIG)**.
- ▶ The actions considered are “from which state shall I start further sampling?”
- ▶ Conditional on  $q_i$ , the EIG of observing a “count”,  $e$ , is given by the expected Kullbeck–Leibler divergence from the current posterior,  $P(\vec{p}_i^{MT} | q_i)$ , to the updated posterior,  $P(\vec{p}_i^{MT} | q_i, e)$ .

$$D_{KL}(P || Q) = \int_{-\infty}^{\infty} dx \ln \left( \frac{p(x)}{q(x)} \right) p(x)$$

- ▶ For Dirichlet distributions,

$$D_{KL}(\lambda^q || \lambda^p) = \log \frac{\Gamma(\lambda^{qt})}{\Gamma(\lambda^{pt})} + \sum_{s=1}^m \log \frac{\Gamma(\lambda_s^p)}{\Gamma(\lambda_s^q)} + \sum_{s=1}^m [\lambda_s^q - \lambda_s^p] [\Psi(\lambda_s^q) - \Psi(\lambda^{qt})]$$

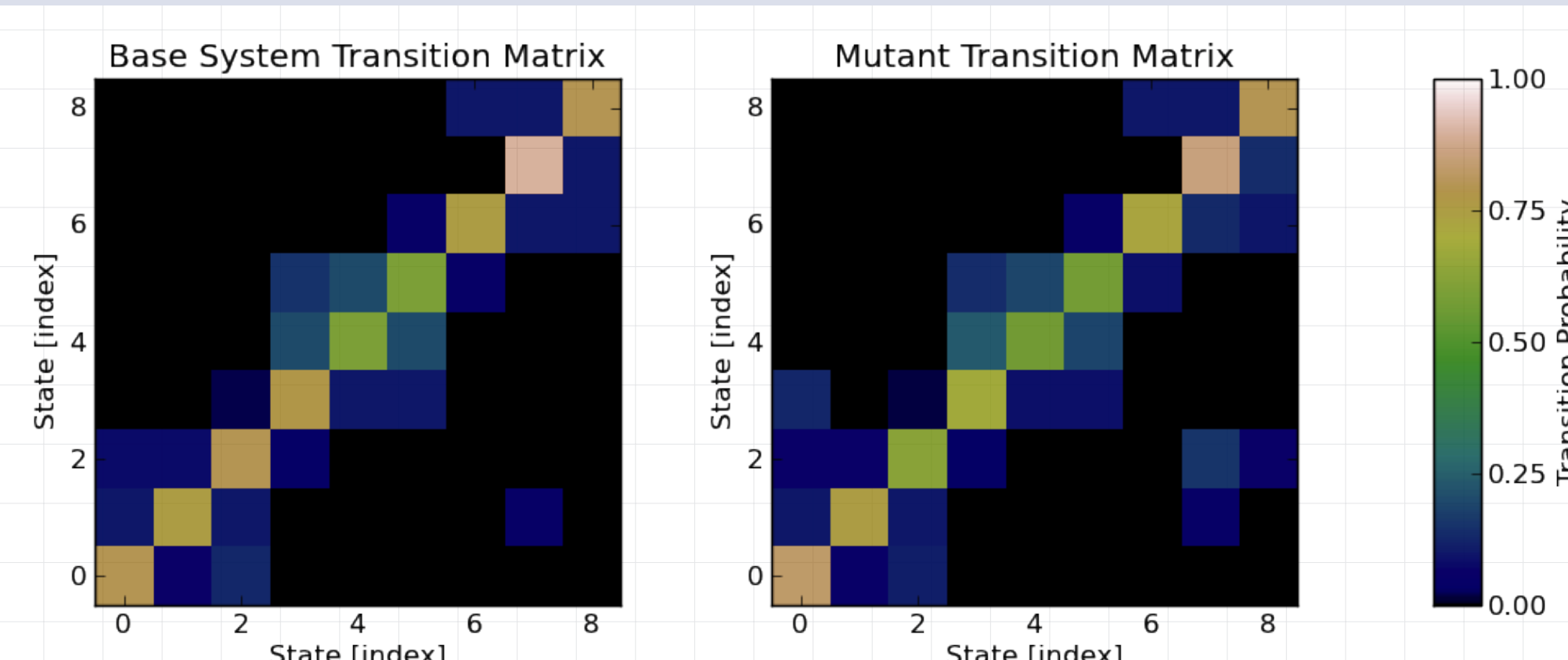
- ▶ If  $e$  is a single count, distributed according to the current Dirichlet-Multinomial posterior, than this simplifies to

$$E[D_{KL} | q_i] = \Psi(\lambda^t) - \log(\lambda^t) + \frac{1}{\lambda^t} \sum_l [\lambda_l (\log(\lambda_l) - \Psi(\lambda_l))]$$

where  $\lambda = q_i \cdot \vec{c}_i^{WT} + \vec{c}_i^{MT} + 1/2$  and  $\Psi$  is the digamma function.

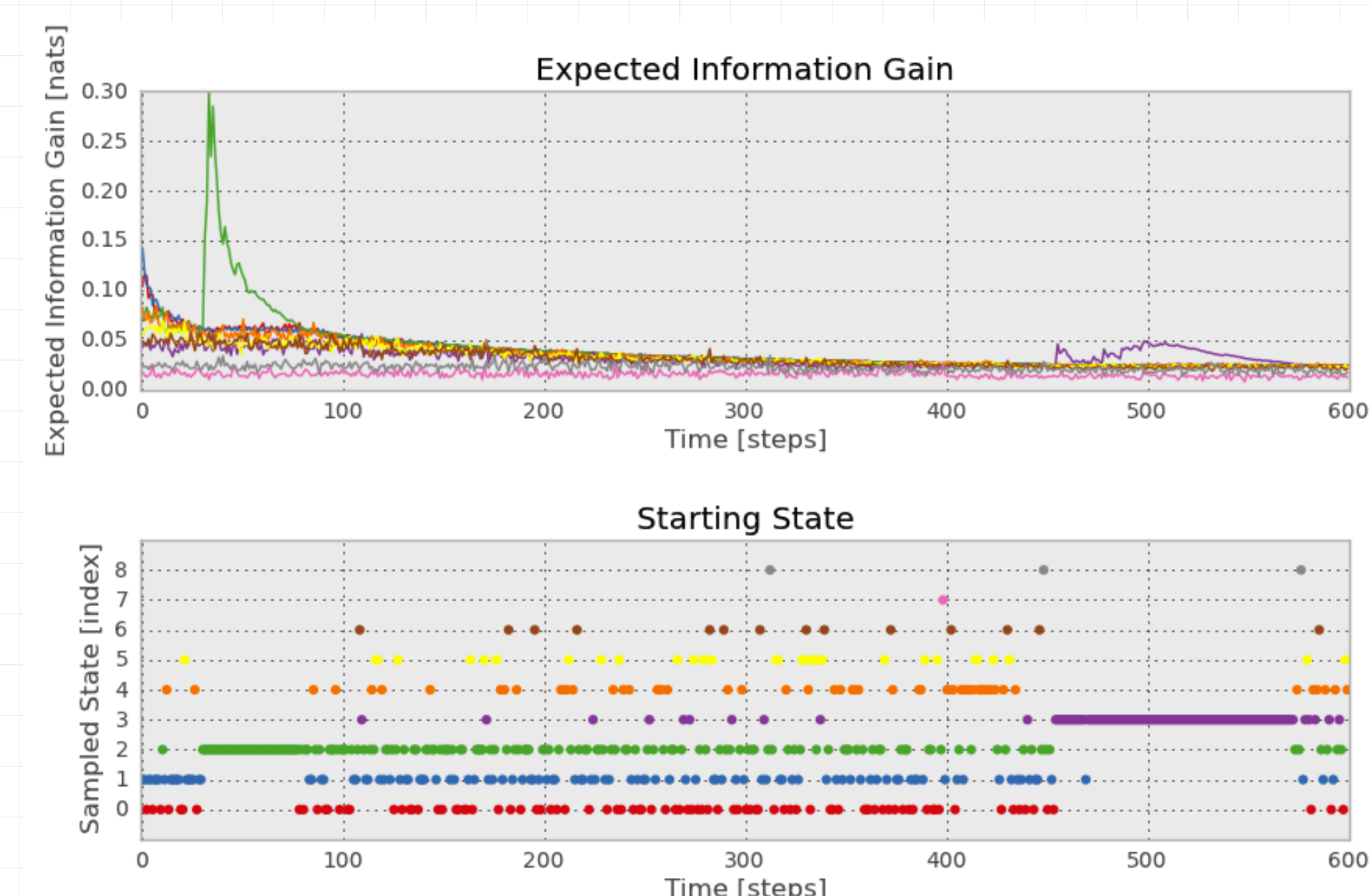
- ▶ Still requires Markov chain Monte Carlo over  $q_i$ .

## Example

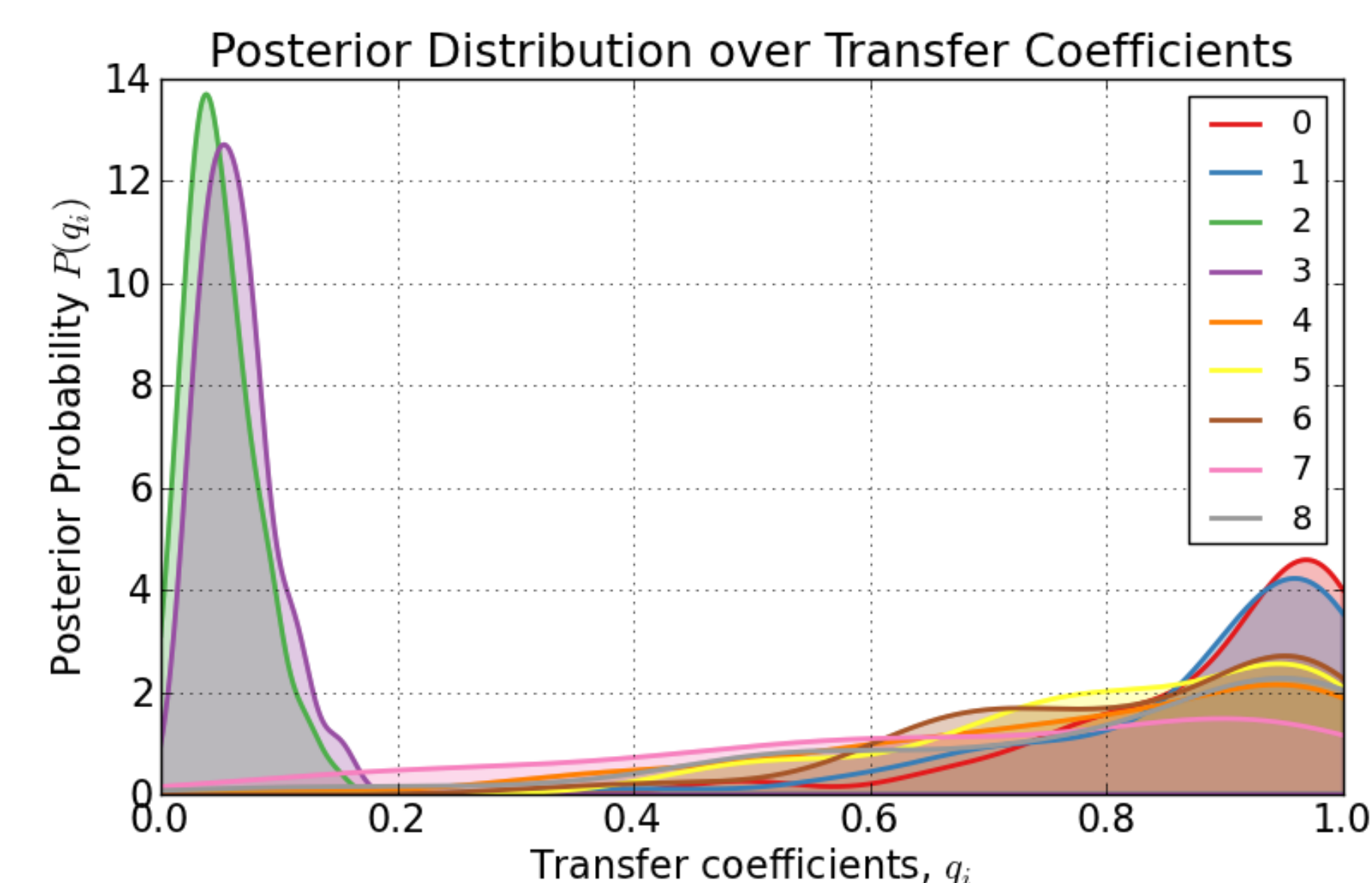


**Figure:** Transition probability matrix for an example system with nine states. Here, the mutant has new connections  $3 \rightarrow 0, 2 \rightarrow 7, 2 \rightarrow 8$

## Example



**Figure:** When the model discovers unexpected behavior anomalies, like the transitions from 2 and 3, it focuses its sampling there.



**Figure:** The posterior distributions over  $q$  show that the model has correctly “discovered” that states 2, 3 are dissimilar in the mutant. Some states (e.g. 8) that are undersampled still feature very wide posteriors.

## Limitations

- ▶ Conformational states are assumed to be the same for wildtype and mutant.
- ▶ No enforcement of detailed balance.
- ▶ Uncertainty due to low wildtype counts in low-population unfolded microstates can swamp model, focus sampling there.

## References

- ▶ Smith, C. “Cloning and mutagenesis: tinkering with the order of things.” *Nat. Methods* 4 455 (2007)
- ▶ Patil, A.; Huard, D.; Fonnesbeck, C. J. “PyMC: Bayesian stochastic modelling in Python.” *J. Stat. Softw.* 35 1 (2010)
- ▶ Cohn, D; Atlas, L.; Ladner, R. “Improving generalization with active learning.” *Machine Learning* 15 201 (1994)