

# Statistical Inference of Transport Mechanisms and Long Time Scale Behavior from Time Series of Solute Trajectories in Nanostructured Membranes.

Benjamin J. Coscia, Christopher P. Calderon      Michael R. Shirts

April 7, 2020

## 1 Introduction

There is a need for highly selective membranes in order to perform efficient separations of components of complex aqueous streams.

- Organic micropollutants
- Desalination and boric acid removal from seawater.
- While many researchers focus on membrane permeability, we may be able to reduce costs of commercial nanofiltration and reverse osmosis with higher selectivity. [1]

Lyotropic liquid crystals (LLC) are a class of amphiphilic molecules whose ordered phases can be cross-linked into mechanically strong membranes capable of highly selective separations.

- The shape of the LLC monomers and water content dictates the ordered phase that they form. There are two phases of particular interest for membrane applications.
- $H_{II}$  phase lyotropic liquid crystals are characterized by hexagonally packed, straight, pores while the  $Q_I$  phase consists of a tortuous network of 3D interconnected pores.
- In both cases the pores are uniform in size with radii on the order of 1 nm giving them a very strict molecular size cut-off.
- Additionally, they have the potential to disrupt conventional membrane separation techniques by being selective based not only on size and charge, but on chemical functionality as well.
- Their pores are lined with LLC monomer functional groups which can potentially be designed to interact with solutes in a chemically-specific manner

There are limits to what we can learn from experiment about LLC membrane design.

- Experimental observables like permeability and selectivity allow us to speculate about the molecular origins of separation processes.
- This drives an empirical design approach which can potentially neglect key interactions which influence selectivity.
- LLC membranes have been shown to exhibit selectivities which cannot be fully explained by relatively simple macroscopic models.

Molecular Dynamics (MD) simulations can give us mechanistic insights with atomistic resolution so that we can intelligently design new membranes for solute-specific separations.

- In our previous work, we built a detailed atomistic model which we used to understand the nanoscopic structure of an LLC Membrane. [2]
- We also used the model in order to gain a qualitative understanding of trapping mechanisms which lead to subdiffusive transport behavior. [3]

Unfortunately, the timescales that we can simulate with MD are insufficient to be able to make well-converged predictions of macroscopic transport properties traditionally used to characterize membranes in the lab.

- However, if we use descriptive stochastic models that can capture solute dynamics, then we could project long timescale behavior in addition to gaining a deeper understanding of solute behavior on short timescales.

In our previous work, we designed two different approaches which used solute time series in order to parameterize stochastic models that could be used to project transport on much longer timescales.

- In our first approach we modeled solute trajectories as subordinated fractional Brownian and Lévy motion, called the anomalous diffusion (AD) model.
- We generated solute trajectories by generating a series of anti-correlated hops separated by random periods of entrapment drawn from a power law distribution.
- Our second approach treated solute motion as a Markov state model with state-dependent dynamics, called the Markov state-dependent dynamical model (MSDDM).
- We parameterized the state transition probabilities between each of eight discrete states as well as the solute dynamics within each of these states. We generated stochastic trajectory realizations by drawing a state sequence based on the transition probability matrix and incorporating the state dynamics while solutes were trapped in each state.

Although both models had reasonable success at predicting solute mean squared displacements (MSDs) on MD simulation timescales, they had shortcomings.

- The MSDDM failed to reproduce the hopping and trapping behavior that characterizes solute center-of-mass trajectories in our MD simulations.
- The AD model did not suffer this qualitative shortcoming, but the persistent curvature of the predicted MSD curves suggested that the model might underestimate MSDs on long timescales.
- The formulation of both models required careful examination and characterization of the interactions and dynamics exhibit by MD trajectories which required considerable human effort.

In this work, we apply the infinite hidden Markov Model (IHMM), a modeling approach that is agnostic to the source of time series data, in order to automatically detect and infer the parameters of an unknown number of latent autoregressive (AR) modes present in solute center-of-mass time series.

- In addition to AR parameters for each state, the IHMM estimates the state transition probability matrix.
- The model helps simultaneously uncover underlying transport mechanisms which give rise to dynamical behavior and project that behavior on longer timescales so that we can estimate macroscopic transport observables.

We use the parameters of the states identified by the IHMM in order to infer dominant solute-membrane interactions and transport mechanisms.

- We compare the inferred mechanisms to those which we manually identified in our previous work.
- Some kind of conclusion here. Did we find more or less states. Any new states/ subdivisions of states?

We can also use the IHMM to generate stochastic trajectory realizations that share the same dynamical characteristics as solute trajectories observed in our MD simulations.

- The trajectories are qualitatively similar, showing expected hopping and trapping behavior.
- They are quantitatively similar in that they reproduce the MSDs measured in MD.

Finally, we use the stochastic trajectory realizations in order to compute the macroscopic flux of each solute and selectivity of the LLC membranes studied towards each solute.

- We relate these macroscopic properties to our nanoscopic model by simulating mean first passage time (MFPT).
- Some kind of conclusion. This membrane is selective towards solutes with this functionality.
- Does the conclusion agree with our previous work? Any length dependence? (I think not)

## 2 Methods

We ran all MD simulations and energy minimizations using GROMACS 2018. We performed all post-simulation trajectory analysis using python scripts which are available online at <https://github.com/shirtsgroup/LLC.Membranes>.

### 2.1 Molecular Dynamics Simulations

We studied transport of solutes in the  $H_{II}$  phase using an atomistic molecular model of four pores in a monoclinic unit cell with 10 % water by weight.

- Approximately one third of the water molecules occupy the tail region with the rest near the pore center.
- We chose to study the 10 wt % water system because solutes move significantly faster than in the 5 wt % system studied previously.
- Appropriate stochastic modeling requires that solutes sample the accessible mechanisms with representative probability.

We chose to study a subset of 4 of the fastest moving solutes from our previous work: methanol, acetic acid, urea and ethylene glycol.

- In addition to exploring membrane structural space the most, these solutes have a relatively diverse set of chemical functionality.
- For each solute we created a separate system and to each system we added 6 solutes per pore for a total of 24 solutes.
- This number of solutes per pore provides a balance of a low degree of interaction between solutes and sufficient amount of data from which to generate statistics on the time scales which we simulate.
- Further details on the setup and equilibration of these systems can be found in our previous work.[3]

We extended the 1  $\mu$ s simulations of our previous work to 5  $\mu$ s in order to collect ample data.

- We simulated the system with a time step of 2 fs at a pressure of 1 bar and 300 K controlled by the Parinello-Rahman barostat and the v-rescale thermostat respectively.
- We recorded frames every 0.5 ns

## 2.2 The Infinite State Hidden Markov Model

Hidden Markov models (HMMs) are a useful and widely used technique for modeling sequences of observations where the probability of the next observation in a sequence depends, at least in part, on a previous unobserved, latent or hidden, state. [4]

- In the context of our simulations, the observations correspond to the center of mass coordinates of the solutes versus time, and the states correspond to the dynamical behavior which give rise to those types of observations.
- The probability of transitioning to a state based on the current state is mathematically defined in terms of an  $n \times n$  transition probability matrix,  $T$ , where  $n$  is the number of states.
- Unfortunately, standard HMMs require  $n$  to be known *a priori*.
- One can partially overcome this by testing a range of numbers of hidden states and determining which is the best representation of the data.

The infinite-state HMM overcomes this drawback by placing a hierarchical Dirichlet process (HDP) prior on the transition probabilities.

- Using some base probability distribution,  $H$ , a Dirichlet process (DP) generates discrete distributions,  $G_0$ , over a countably infinite number of probability measures:

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k} \quad \theta_k \sim H, \beta \sim GEM(\gamma) \quad (1)$$

where the  $\theta_k$  are values drawn from the base distribution and the weights  $\beta_k$  come from a stick-breaking process parameterized by the concentration parameter  $\gamma$  (equivalently referred to as  $GEM(\gamma)$ ).

- Also expressed  $G_0 \sim DP(\gamma, H)$
- The concentration parameter,  $\gamma$ , expresses one's confidence in the base distribution  $H$ .
- We use a uniform base distribution.
- When  $\gamma \rightarrow 0$ , the first weight of  $G_0$ ,  $\beta_1$ , approaches unity and for  $\gamma \rightarrow \infty$ , the weights become uniform and  $G_0$  closely resembles  $H$ .
- Each row,  $G_j$ , of the transition matrix is produced by drawing from a DP specified using the  $\beta$  vector as a discrete base distribution and a separate concentration parameter,  $\alpha$ .

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k} \quad \pi_j \sim DP(\alpha, \beta) \quad (2)$$

- This hierarchical specification ensures that the transition probabilities in each row share the same support points  $\{\theta_1, \dots, \theta_k\}$ .
- Once the model has converged only a finite number of states will have significant sampling.

We describe the dynamics of each state visited by solutes using a first order vector autoregressive (VAR(1)) model.

- In general, a VAR( $r$ ) process is characterized by a vector of observations in a time series that are linearly dependent on  $r$  previous values of the time series vector:

$$\mathbf{y}_t = \mathbf{c} + \sum_{i=1}^r A_i \mathbf{y}_{t-i} + \mathbf{e}_t \quad \mathbf{e}_t \sim \mathcal{N}(0, \Sigma) \quad (3)$$

Previous observations are weighted by coefficient matrices,  $A_i$ . The VAR( $r$ ) process is further characterized by a shift in the mean of each dimension by the vector  $\mathbf{c}$  and a white noise term  $\mathbf{e}_t$ . [5]

- We assumed  $\mathbf{e}_t$  to be multivariate Gaussian noise, with mean zero and covariance,  $\Sigma$ .
- We limited our analysis to an autoregressive order of  $r = 1$ .
- This means that we only parameterize  $A_1$ . To simplify notation, we will just call it  $A$ .
- We used a matrix-normal inverse-Wishart prior on parameters  $A$  and  $\Sigma$  and a Gaussian prior on  $\mathbf{c}$  in order to infer their values. [6]

Using the IHMM framework, we estimated the most likely number and sequence of hidden states in each solute center-of-mass trajectory while simultaneously inferring VAR(1) parameters for each state and the overall state transition probability matrix,  $T$ .

- We created a python implementation of this process which we heavily adapted from the MATLAB code of Fox et al. [7]
- Parameter estimation is iterative. Therefore, we looked for convergence of each entry in the  $\mathbf{c}$  vector as well as the  $A$  and  $\Sigma$  matrices of Equation 3.
- We detected equilibration of the parameters using the module `pymbar.timeseries.detect_equilibration` on the time series of parameter estimates.
- Plots illustrating parameter convergence are available in the SI.
- We refer the interested reader to much more extensive descriptions of the inference and sampling procedures used to estimate the VAR(1) parameters and the state sequence. [4, 8, 9, 6, 10]

We carefully applied the IHMM algorithm in a way which takes advantage of the system’s cylindrical symmetry.

- There are a number of viable ways in which one could choose to analyze these time series.
- We could use all 3 dimensions straight from the output trajectory.
- We could use just the  $z$  dimensional coordinate since we are primarily interested in through-plane membrane transport.
- We chose to use a multi-step procedure that we believe adequately distinguishes every type of distinct dynamical behavior exhibited by the solutes.
- Ultimately, we work in cylindrical coordinates since each pore is a cylinder where we expect solutes to exhibit radially symmetric dynamics.

We start by applying the IHMM algorithm to the 3D solute center-of-mass coordinate trajectories transformed relative to the closest pore center.

- We tracked the solute’s motion along the pore axis with the center-of-mass  $z$  coordinate.
- Using the nearest pore center as the origin, we represented the radial distance of each solute’s center-of-mass from the pore center in 2 dimensions,  $x$  and  $y$  (see Figure 1).
- By working in cartesian coordinates, we avoid mathematical complexity introduced by cylindrical coordinates while estimating the state sequence.

We applied the IHMM to each of the 24 solute trajectories independently.

- Although the IHMM is capable of identifying an infinite number of states, a Dirichlet Process tends to exhibit a “rich get richer” effect, favoring a fewer number of states.
- By applying the algorithm to each trajectory independently, we reduce the possibility of lumping together multiple similar states which we would prefer to stay separated before clustering.

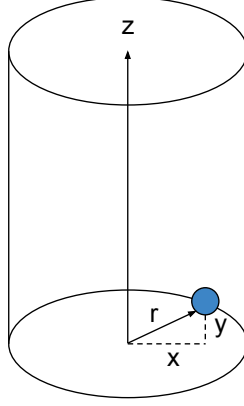


Figure 1: Using the nearest pore center as the origin, we represented the solute’s location along the pore axis in terms of the  $z$  coordinates and their radial distance from the pore centers in 2 dimensions,  $x$  and  $y$ .

The states identified by the IHMM are heavily influenced by the Gaussian prior placed on  $\mathbf{c}$  in Equation 3.

- The entries of  $A$  and  $\Sigma$  do not vary over a wide range, so the final parameters were relatively insensitive to the priors.
- In order to maximally automate the IHMM procedure, we attempted to parameterize the prior on  $\mathbf{c}$  in an intelligent way.
- The prior parameters should be chosen such that the mean level of each state lies within a region of reasonable probability of the prior (see Figure 2).
- In each dimension, we defined the prior mean to lie halfway between the maximum and minimum of each trajectory dimension.
- To parameterize the prior’s variance in each dimension, we defined the maximum and minimum to be 2 standard deviations from the prior mean.
- Although this approach has worked quite well for the data in this work, it is important to check the results to determine whether further adjustments to the prior might be needed.
- In the Supporting Information, we show the result of a parameterization where the prior parameters of  $\mathbf{c}$  were poorly chosen.

We ran 2000 iterations of the IHMM procedure in order to arrive at converged state sequences and parameters for each state.

- Most parameters converged quickly (within 50-100 iterations) while others took up to 1000 iterations.
- We defined the finalized parameters of each state as the average of the parameters from each iteration recorded after the equilibration time point.
- We found the equilibration time point of each component of the  $\mathbf{c}$  vector as well as  $A$  and  $\Sigma$  matrices, and then used the longest equilibration time of all dimensions as the equilibration time point.

We reparameterized the time series, preserving the state sequence, in terms of the radial and axial coordinates ( $r, z$ ) because  $x$  and  $y$  aren’t meaningful alone due to the system’s radial symmetry.

- We converted the  $x$  and  $y$  center-of-mass coordinates to  $r$ .
- We fixed the state sequence from the 3D parameterization and applied the inference component of the IHMM procedure to the cylindrical trajectories in order to estimate the VAR(1) parameters in terms of  $r$  and  $z$ .

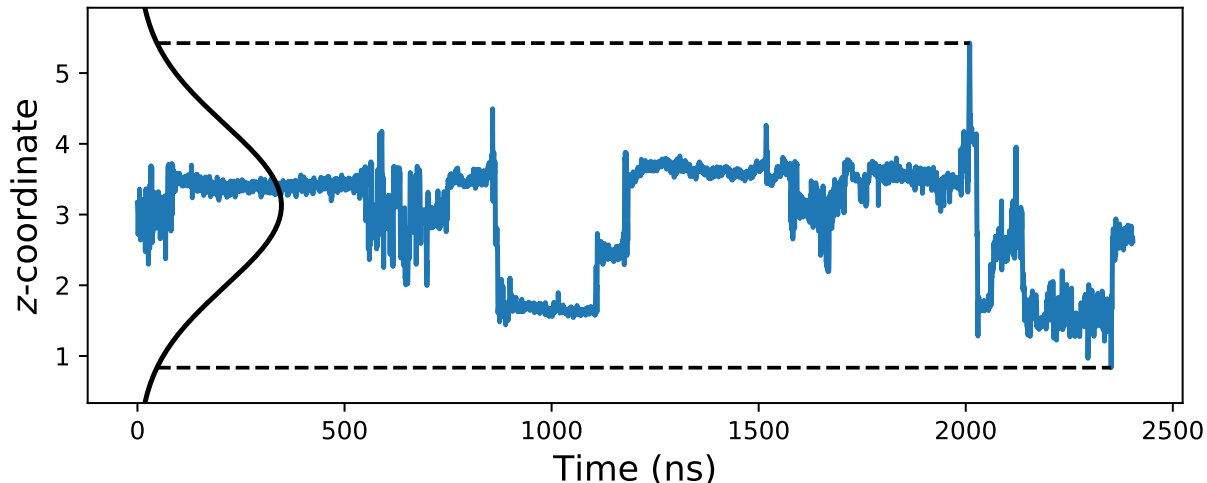


Figure 2: The parameters of the prior on the mean vector,  $c$ , (black line), which represents the coordinates at which solute are trapped, should be chosen such that the mean levels of each state identified in the trajectory (blue line) lie within regions of the prior with reasonable probability. We chose the mean of the prior as halfway between the maximum and minimum of each trajectory dimension. We chose  $\sigma$  of the prior by defining the maximum and minimum to be 2 standard deviations from the mean (dashed lines).

We clustered like parameter sets in order to reduce the state space to a more easily interpretable size.

- For each solute studied, we identified 200-325 independent states, each with separate VAR(1) parameters.
- Many of these states exhibit very similar dynamical behavior except their mean levels are different.

We reduced the parameter space used agglomerative clustering, a hierarchical clustering approach which uses a linkage criteria in order to successively merge similar clusters until a desired intracluster distance threshold or number of clusters is reached.

- We used the Ward linkage criteria, which works to minimize the sum of the squared differences within all clusters.
- We elected to choose the number of clusters rather than the distance threshold.
- For our data, non-parametric methods such as Bayesian Gaussian mixture models tend to delocalize the clusters in parameter space (see Supporting Information).

We clustered based on the two diagonal entries of  $A$  and  $\Sigma$  as well as the radial means,  $\mu_r$ , for a total of 5 variables.

- One can choose alternative clustering features, such as the eigenvalues of the  $A$  and  $\Sigma$  matrices, but we chose their diagonals because they can be easily translated to solute behavior.
- We also chose to cluster the diagonals of each matrix and the radial means independently before combining the clusters to assign final state labels.
- If we divide the total states into  $m$  clusters based on the diagonal entries of  $A$ ,  $n$  more clusters based on the diagonal entries of  $\Sigma$ , and  $k$  clusters based on  $\mu_r$ , then there are a total of  $mnk$  possible cluster combinations.
- This allows slightly more flexibility in the total number of states compared to clustering on all five features at once, since there can be up to, but not necessarily exactly,  $mnk$  states.

Choosing the number of clusters can be somewhat subjective so we attempted to add some structure to the selection process by following a set of qualitative and quantitative guidelines.

- We used the silhouette test in order to score the quality of clustering as a function of the number of clusters chosen.
- For our data, the silhouette test generally favors the lowest number of clusters possible (see Supporting Information).
- However, choosing too few clusters tends to result in inadequate distinguishment between visually obvious differences in dynamic behavior.
- This results in finalized parameter sets that are averages of distinct behavior which presents further problems with the predictive modeling that we discuss later on.
- We aimed to maintain the highest silhouette score, and thus lowest number of clusters, possible while verifying that visually distinct states stayed separated.
- Based on these guidelines, we decided to group the  $A$  and  $\Sigma$  parameters into five clusters each and the radial means into three clusters.
- Of the 75 distinct states allowed by this formulation, the solutes in this study clustered into 30–33 distinct clusters.
- While this may seem like a high number of states, many are sampled infrequently.

We remapped the state sequence based on the cluster assignments and generated a state transition probability matrix,  $T$ .

- The IHMM algorithm also produces an estimate of  $T$ , but since we fixed the state sequence, we decided to explicitly calculate  $T$  by counting the number of transitions between states compiled across all trajectories.

We used the parameter inference component of the IHMM algorithm in order to infer  $A$ ,  $\Sigma$  and  $\mu_r$  of the clustered states.

- We could not simply take the mean of the clustered  $A$  and  $\mathbf{e}_t$  parameters because it is not clear that this is a linear operation for this problem.
- To circumvent this problem, we modified the  $(r, z)$  solute trajectories so that they had a mean of zero, leaving only the fluctuations.
- We did this by subtracting the maximum likelihood estimate of the  $\mathbf{c}$  vector from each same-state segment of the unclustered trajectory.
- We used the IHMM algorithm on this modified trajectory to infer the clustered state parameters by fixing the clustered state sequence.

Finally, we generated stochastic trajectory realizations using the clustered parameter sets.

- We drew state sequences with transition probabilities given by  $T$ .
- While in a given state, we simulated motion according to the VAR(1) parameterization of that state.
- After each state transition, we reset the unconditional mean of each state based on the particle’s position immediately before the state transition occurred.



## 2.3 Calculation of Physical Interactions

We determined the number of hydrogen bonds between each solute and the membrane as a function of time.

- Based on the geometric criteria of Luzar and Chandler, we define a hydrogen bond to exist if the distance between donor, D, and acceptor, A, atoms is less than 3.5 Å and the angle formed by  $D - H \cdots A$  is less than 30 °. [?]

We estimated the lifetime of hydrogen bonds by recording the length of sequential frames where solute remained hydrogen bonded.

- We still counted sequences where hydrogen bonds were broken for a single frame before reforming.
- We reported the 95th percentile of hydrogen bond lifetimes.

We also measured the degree of association between solutes and sodium ions on a frame-by-frame basis.

- We define a sodium ion to be associated with an atom if they are within 2.5 Å of each other, as determined in our previous work. [3]

We developed a way to measure the local density of the membrane at arbitrary points in the unit cell.

- We histogrammed the positions in three dimensions
- Since our system is in a monoclinic unit cell, we periodically replicated the system in +/- x, y and z direction and then set the bounds on the histogram to create a rectangular box encompassing the unitcell with a 1 nm buffer between the histogram unitcell boundaries.
- We then used a regular grid interpolator in order to allow interpolation at arbitray points within the grid.

We quantified average solute movement using the time averaged mean squared displacment (MSD) of the solute center of mass trajectories.

- The time-averaged MSD measures all observed displacements over time lag  $\tau$ :

$$\overline{z^2(\tau)} = \frac{1}{T - \tau} \int_0^{T-\tau} (z(t + \tau) - z(t))^2 dt \quad (4)$$

where T is the length of the trajectory.

- We reported  $1\sigma$  confidence intervals based on the results of 200 bootstrap trials where calculated the mean of  $n$  randomly chosen trajectories with replacement from the pool of trajectories, where  $n$  is the number of trajectories.

## 2.4 Estimating Flux and Selectivity

We calculate first passage times by propagating stochastic trajectories until they reach distance  $L$ .

We determine the mean first passage time (MFPT) using the following equation: [11]

$$P(t) = -\frac{1}{\sqrt{\pi}} e^{-(L-vt)^2/(4Dt)} \left( -\frac{D(L-vt)}{4(Dt)^{3/2}} - \frac{v}{2\sqrt{Dt}} \right) \quad (5)$$

Flux,  $J$ , is simply  $1 / \text{MFPT}$  by the Hill relation. [12]

In our previous work, we showed that, in the absence of convective solute flux, selectivity towards solute  $i$  versus solute  $j$  can be calculated by:

$$S_{ij} = \frac{J_i/\Delta C_i}{J_j/\Delta C_j} \quad (6)$$

where  $\Delta C_j$  is the trans-membrane concentration difference.

### 3 Results and Discussion

Solutes in this system show a wide range of behavior influenced by the heterogeneity of the membrane’s nanostructure as well as the interactions between monomer and solute chemical functionality.

- Our application of the IHMM does a good job of identifying and distinguishing these behaviors.
- The following approach to analysis demonstrates how one can discover and explain the very complex behavior exhibited by solutes over long MD trajectories in terms of solute-membrane interactions.
- Further, we can use the IHMM in order to generate stochastic trajectories, with dynamics similar to MD, so that we can estimate macroscopic transport properties.

#### 3.1 Automatic Detection of Distinct Dynamical Modes

We applied the IHMM to all trajectories of each of the four solutes studied.

- Based on our choice of clustering, we identified over 30 distinct dynamical modes for each solute.
- Many of these modes are sampled very infrequently and thus do not merit significant study.
- In this section, we explore a subset of this data, focusing on the states and dominant interactions most common to solute trajectories.

Methanol exhibits a range of dynamic behavior.

- In Figure 3, we plot a representative 2D methanol trajectory in cylindrical coordinates.
- The coordinates are colored according to the cluster with which methanol’s dynamics are most consistent, which implies that this particular methanol trajectory exhibits 8 distinct dynamical behaviors (in order of their first appearance, they are colored dark yellow, red, light blue, light green, orange, blue, teal and lime green).
- The solute timeseries are coupled with color coded bars which describe the solute’s local environment and physical interactions that it undergoes at that trajectory frame.
- From the top, they describe whether the solute is associated with sodium, whether the solute is hydrogen bonding with the monomer as well as to what part of the monomer it is hydrogen bonding, and finally, the local number density surrounding the solute.

One can gain useful chemical intuition just by studying plots like Figure 3.

- It is clear that, for methanol, association with sodium is a much rarer interaction than hydrogen bonding with the LLC monomers.
- Methanol hydrogen bonds to all regions of the monomers dependent on its radial position.
- Its fluctuations tend to be smallest when hydrogen bonded or in areas of high local number density.
- For example, from approximately 75 to 525 ns, methanol appears to be trapped in a high density region of the tails.
- In the time that follows, it enters a region of relatively low density where fluctuations are quite large.
- During this time period, there is intermittent, short-lived hydrogen bonding with the oxygen atoms at the tail ends, as suggested by the frequent state changes during that time period and red slivers in the bar above.
- Significant hops in the  $z$  direction appear to be described by the dark blue state.

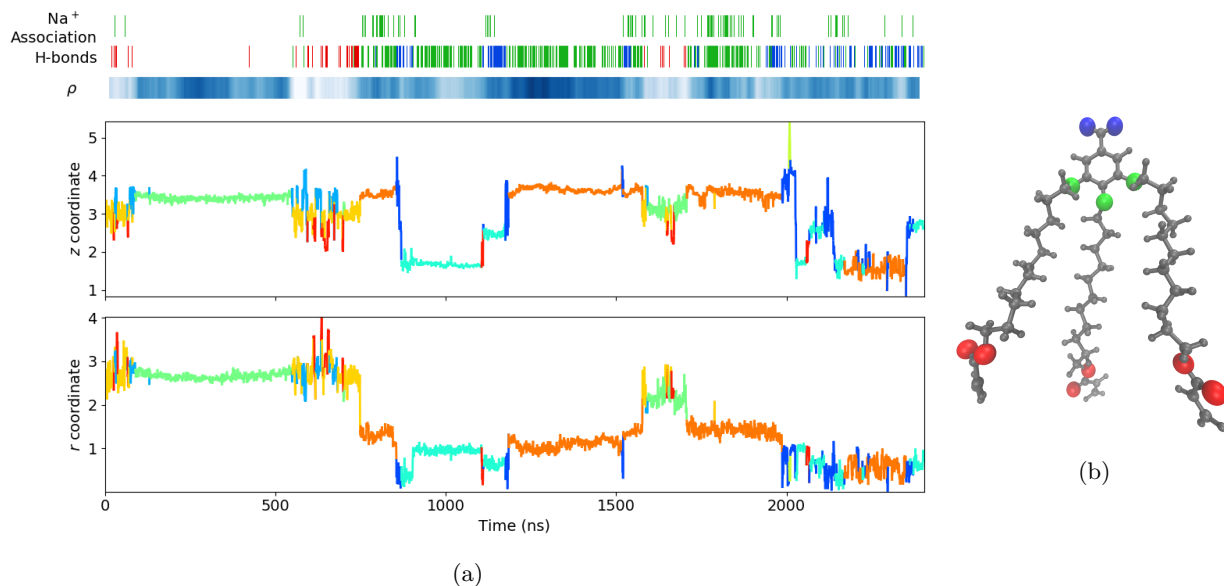


Figure 3: (a) We can learn a significant amount of detail about solute motion by viewing solute trajectories color-coded according to their clustered dynamical behavior alongside plots of the dominant molecular interactions and descriptions of the membrane’s local environment. In the plot above, we show the radial and axial coordinates of a single methanol molecule over the course of the equilibrated portion of its MD simulation. Disjoint segments of the same color imply the solute behaves similarly at different points of the trajectory because their VAR parameters place them in the same cluster. There are eight distinct dynamical modes exhibited by this trajectory. In order of their first appearance, they are colored dark yellow, red, light blue, light green, orange, blue, teal and lime green. On the top of the plot are bars which represent different physical interactions and the membrane’s local environment as a function of time. The top bar is colored green at time points where solutes are associated with sodium. Methanol associates with sodium ions relatively infrequently. The next bar down is colored when hydrogen bonds exist. The colors match the enlarged oxygen atoms highlighted in (b). Blue slivers correspond to hydrogen bonds donated to the monomer’s carboxylate head group. Green slivers correspond to hydrogen bonds donated to the ether linkages between the head groups and tails. Red slivers correspond to hydrogen bonds donated to the oxygens on the ends of the monomer tails. Methanol exhibits significant hydrogen bonding with all parts of the monomer. Which part of the monomer methanol hydrogen bonds to is heavily correlated to the solute’s  $r$  coordinate. Finally, the bar with the blue gradient represents the local membrane density based on the solute’s  $(r, z)$  coordinate. Darker shades of blue correspond to higher membrane densities. The density appears to be higher when methanol motion is restricted (e.g. from 75 to 550 ns). It is relatively low in areas where methanol’s position fluctuates significantly (e.g. from 550 to 700 ns).

- After entering the dark blue state ca. 850 ns, methanol appears to jump significantly in the  $z$  direction from a hydrogen bonded state with the ether linkages to a hydrogen bonded state with a head group in a monomer below it.

It is clear that there is a wealth of information in these plots, but they don't give the complete picture needed in order to formulate LLC membrane design principles.

- One can analyze each segment sequentially and form a clear picture of this instance of methanol's behavior without the need to visualize the trajectory with software like VMD.
- However, it is still possible to gain a more general understanding of methanol's average behavior.

We may learn the most by studying the dynamical modes common to the majority of the solute trajectories.

- Using the IHMM, we identified 33 total dynamical modes exhibited by methanol.
- Of those, only six states appear in a third (8 out of 24) or more methanol trajectories, and three of them appear in over 50 % of the methanol trajectories.
- 12 of the detected states are shared by two or less methanol trajectories.
- In Figure 4a, we plot representative dynamics of each of the six states which we determined to be prevalent.
- We also chose to study a seventh state, which only appears in three out of 24 trajectories, because methanol experiences very long periods of entrapment while in this state. Long trapping events, even if they are relatively rare, will have a significant influence on solute flux.
- No trajectory contains all seven states. However, the trajectory in Figure 3 exhibits all modes except states 6 and 7.
- The states in Figure 4 are colored to match states that also appear in Figure 3.

We can begin to dissect the dynamics of each state by studying their parameters.

- States 1, 3 and 5 have radial means near 1 nm or less (see  $y$  axis of  $r$  direction plot in Figure 4a), which suggest that they occur close to the pore centers.
- States 2, 4, 6 and 7 occur well within the tail region of the membrane.
- States 1 and 2 have higher covariances in both the axial and radial directions compared to all other states.
- Covariance in each dimension appears to have a strong correlation.
- States 3 and 5 (in the pore region) as well as states 4, 6 and 7 (in the tail region) are primarily distinguished based on their autoregressive coefficients,  $A$ .
- Higher values of  $A$  generally lead to more wandering about the state mean.
- One could argue that the parameters of states 3–7 are actually very close which suggests that solutes might behave alike in topologically distinct regions of the membrane.

Despite some similarity in their VAR parameters, the length of time spent in states 3–7 appears to be dependent on the solute's radial location.

- Based on the self-transition probabilities, we estimated the average time spent by each solute in each state (Figure 4c).
- States 3, 5 and 7 have long dwell times relative to the rest because they have high self-transition probabilities.

- Somewhat surprisingly, both states 3 and 5 correspond to solutes located close to a pore center, where one might expect solute motion to be less restricted.
- Although solutes in states 4 and 6 are in the viscous tail region, their dwell times are only intermediate. This could be due to inhomogeneity in the tail region which causes transitions between many different trapped states but with little net effect on solute MSD.
- However, there are instances of states within the tail region with dwell times that exceed 100–200 ns, as shown by our inclusion of state 7.
- There are no other states with dwell times longer than state 5 in the pore region.

States 1 and 2 exhibit the shortest dwell times of the states studied.

- This is consistent with the trajectory in Figure 3.
- State 2 is the shortest lived state, appearing to bridge other tail region states.
- State 1 appears to represent a significant hopping state.
- It has a high covariance and occurs near the pore center, where there is less obstruction to solute motion.
- Its dwell time of ca. 8 ns, suggests that solutes stay in this highly mobile state for long enough for their positions to displace significantly.

Most of what we have learned thus far has been based on the parameters of our model with some qualitative validation using a single methanol trajectory.

- While we may be able to further corroborate our arguments by analyzing more individual trajectories, it is more useful to correlate the state behavior with solute-membrane interactions.

The types of hydrogen bonding undergone by solutes is dependent on radial location (see Figure 5).

- States with radial means near 1 nm or less (states 1, 3 and 5) typically hydrogen bond with the monomer carboxylate groups and ether linkages.
- States with radial means greater than 1 nm (states 2, 4, 6 and 7) tend to hydrogen bond with the monomer tails.
- In the past we’ve shown that there is a gradual transition from the hydrophilic to hydrophobic region in these membranes, meaning the monomer location is not tightly bound in  $r$ .
- Therefore, there are instances of hydrogen bond interactions that one might consider uncharacteristic based on the state’s radial mean.

States which exhibit longer dwell times have longer hydrogen bond lifetimes (see Figure 5).

- States 3–7 have significantly longer hydrogen bond lifetimes than states 1 and 2.
- It makes physical and mathematical sense that states 1 and 2 should not stay trapped by hydrogen bonding for long periods because they typically take large hops as communicated by their large covariances.
- In the tail states, hydrogen bond lifetimes appear to increase with their dwell times.
- However, in the pores, the same linear relationship does not hold.
- For example, state 5 has the lowest hydrogen bond lifetime of states 3–7 despite having the second longest dwell times.

The local density experienced by solutes may have an impact on the length of solute entrapment.

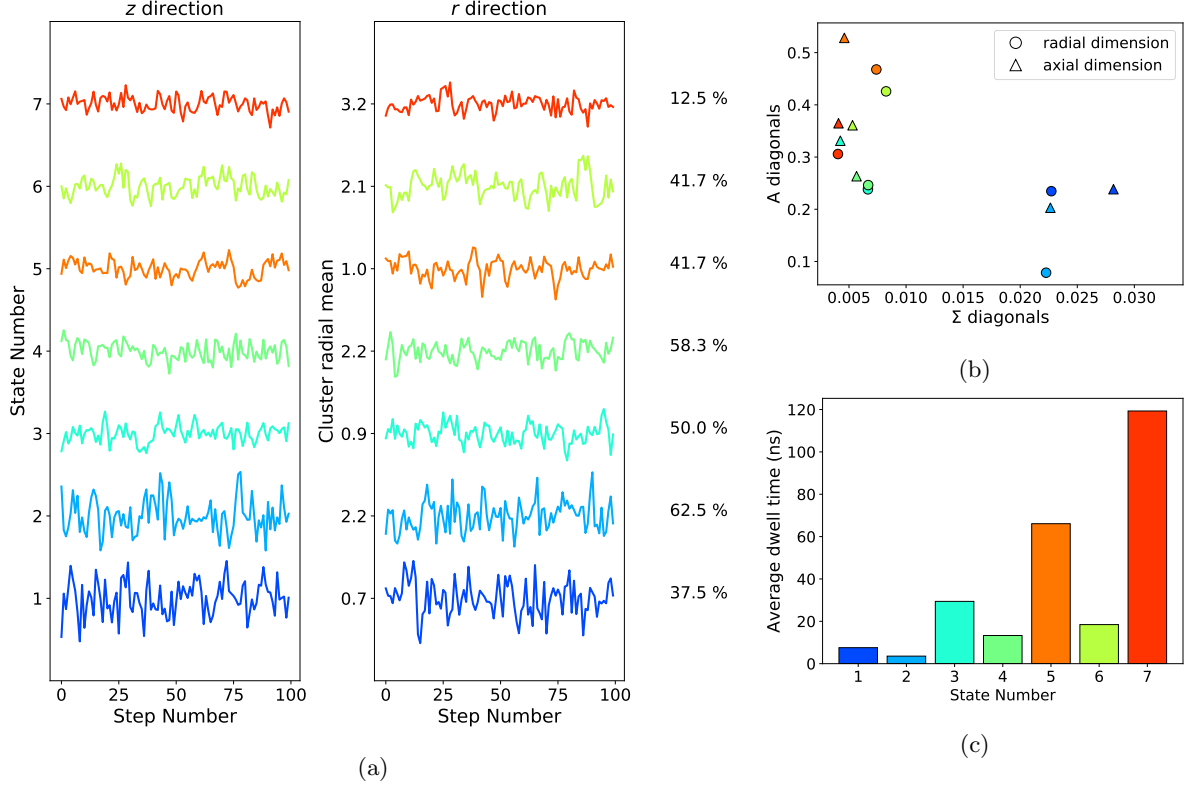


Figure 4: We can learn the most about methanol’s behavior by studying the dynamics of the states common to many of the methanol trajectories. In (a), we show representative dynamics of six states that are common to more than one third of the solute trajectories. The  $y$ -axis labels of the  $r$  coordinate plot specify the radial mean of each clustered state. The percentage of trajectories in which the states appear are labeled to the right of the  $r$  plots. States 1–5 appear in the trajectory shown in Figure 3 and are colored to match for easy visual comparison. All time series have a mean of zero and are shifted to put the mean in line with the  $y$ -axis labels for the purpose of visualizing them. (b) We can begin to understand the solute behaviors demonstrated in (a) by understanding their VAR parameters. States 1 and 2 have higher covariances in both the axial and radial directions compared to all other states. This should lead to significant solute displacement. Covariance in each dimension appears to have a strong correlation. States 3–6 all have relatively low covariances, which may be characteristic of trapped states. They are primarily distinguished based on the magnitude of their autoregressive coefficients,  $A$ , and radial means. Higher values of  $A$  imply motion which wanders more about the mean. States 1, 3 and 5 have radial means near 1 nm or less, which suggest that they occur close to the pore centers. States 2, 4 and 6 occur well within the tail region of the membrane. (c) Using the transition probabilities ( $p$ ), we estimated the average time spent within each of the states using 1000 simulations of a Bernoulli process where we continued to draw until transitioning out of the state with probability  $1 - p$ . Of the prevalent states, states 1 and 2 exhibit the shortest dwell times alongside their high covariances. States 5 and 7 have the longest dwell times in the pore and tail regions respectively.

- On average, solutes in state 5 are surrounded by a relatively high density of particles.
- Based on its radial location and preference towards hydrogen bonding with the ether linkages (see Figure 5a), it is reasonable to assume that while in state 5, methanol spends most of its time just outside the pore region, sandwiched between the monomer head groups.
- Although state 5 hydrogen bonds break more frequently than one might expect for a state with long dwell times, it is possible that the solute’s dense surroundings prevent it from wandering off before the hydrogen bonding interaction can be recovered.
- This may explain the behavior of methanol in state 5 of the trajectory in Figure 3 where the solute’s center of mass appears to move more freely about its mean than other trapped states.
- The data suggests that methanol molecules in state 3 lie closer to the pore center than state 5 because they tend to hydrogen bond most frequently with the carboxylate groups.
- Methanol molecules which experience state 3 are subject to a lower density local environment and rely more heavily on hydrogen bonding interactions, evidenced by their longer lifetimes, to stay trapped.

States describing behavior in the tail region show a direct relationship between hydrogen bond lifetimes and local density.

- Methanol molecules which experience state 2 spend their time in a low density region of the tails where they can easily break free of hydrogen bonds.
- Conversely, methanol molecules trapped in state 7 are in the densest region of all states studied. It is likely difficult for methanol to escape these high density regions and its position is further stabilized via hydrogen bonding.

### 3.2 Reproducing MD Trajectories and MSDs with the IHMM

We can use the IHMM in order to generate stochastic trajectory realizations which bear some qualitatively similar characteristics to those output by our MD simulations.

- Because we clustered the data, we are left with a single transition probability matrix which allows transitions between dynamical behaviors shared by the 24 MD trajectories.
- In Figure 6, we show some stochastic realizations generated by our models compared directly to MD.
- The trajectories show clear periods of entrapment with intermittent hops.

The MD trajectories show much broader variability in their overall motion than realizations of the IHMM model.

- Some trajectories span a wide range of  $z$  coordinates, while others stay trapped for nearly the entire simulation.
- Our model is very unlikely to generate trajectories with dwell times similar to the third MD trajectory shown in Figure 6.
- This is because the realizations generated by our model are representative of the average behavior of solutes studied.
- It is possible that certain states with long dwell times should have been clustered as their own, but the transition probabilities would have high uncertainties.
- It is also possible that we should not be using a static transition matrix in order to generate stochastic trajectories, but rather sample the rows of the transition matrix from a Dirichlet distribution since average dwell times are highly sensitive to self transition probabilities.

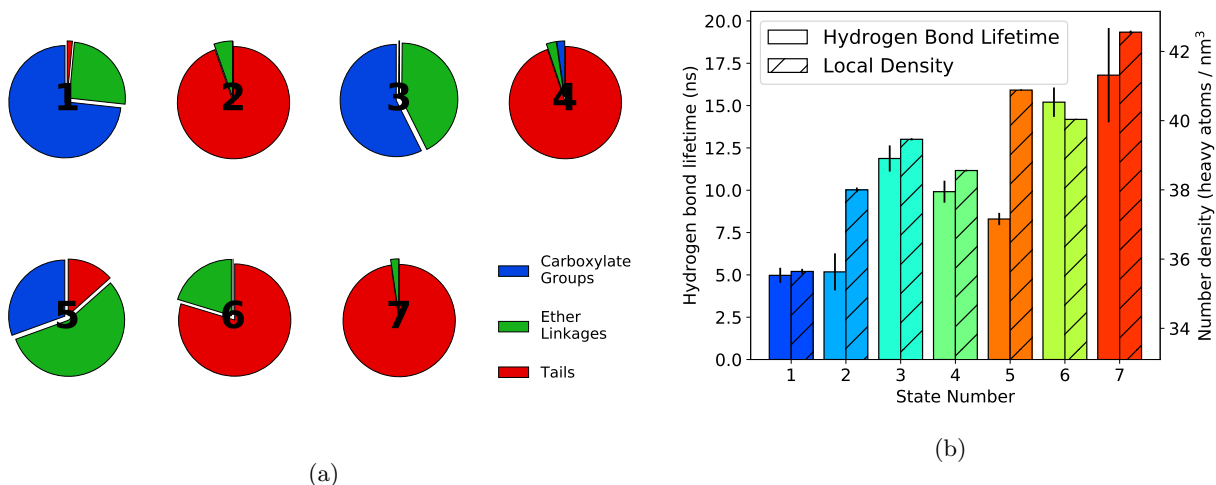


Figure 5: (a) Methanol hydrogen bonds with different segments of the LLC monomer dependent on the state. States with mean radial locations in the pore region tend to hydrogen bond with the head groups and ether linkages. States with mean radial locations in the tails tend to hydrogen bond with the tail oxygen atoms. (b) Hydrogen bond lifetimes and the local density experienced by methanol molecules can help us to understand dynamical states in terms of membrane interactions. States 1 and 2, which have large covariances, also have short hydrogen bond lifetimes and exist in relatively low density regions. The hydrogen bond lifetimes of methanol in states with radial means in the tail region tend to increase with local density and mean dwell times. This monotonic relationship does not exist for states in the pore region. The high local density experienced by state 5 methanol molecules may explain why hydrogen bond lifetimes are short but dwell times are high. State 5 methanol's preference towards hydrogen bonds with ether linkages suggests that the solute sits between LLC monomer head group which prevent it from wandering to far before recovering its previous hydrogen bond.



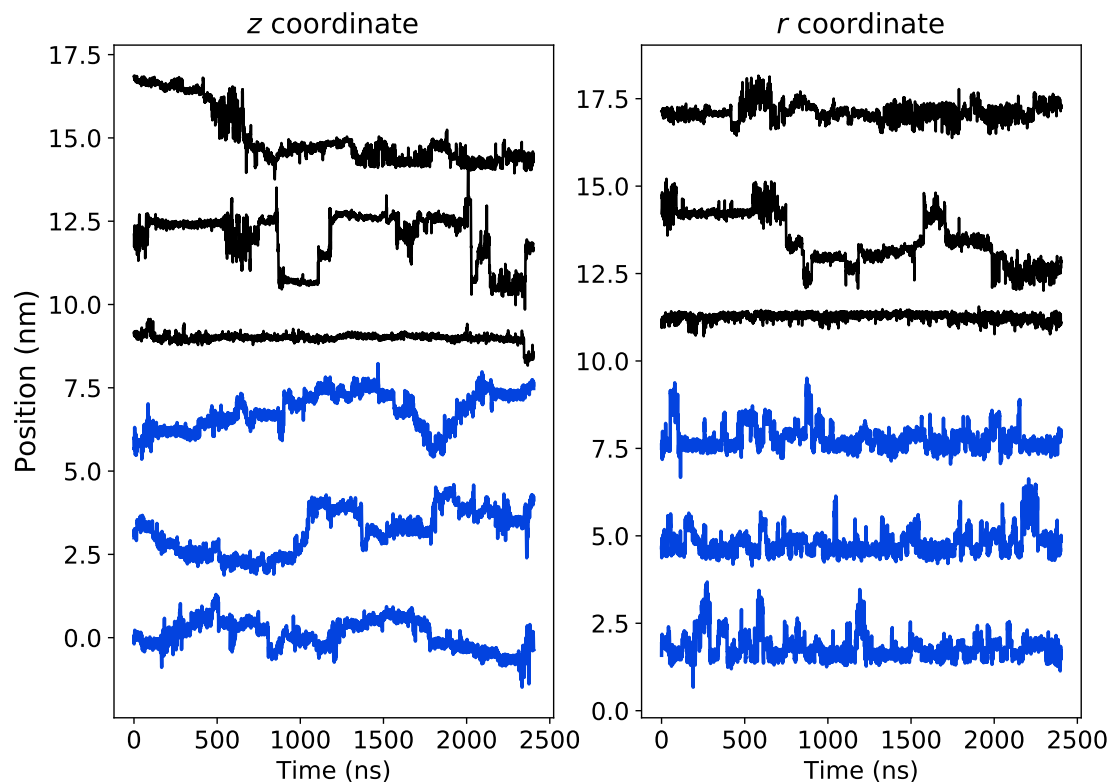


Figure 6: We can generate trajectories which bear some qualitative resemblance to the MD trajectories. Solute trajectories generated by our model (blue) show the same hopping and trapping behavior exhibited by MD (black). The behavior of individual MD trajectories tend to show wider variability than our model's realization because the model effectively represents the average behavior of the MD trajectories. This implies that much longer simulations or a different method of trajectory generation might be necessary in order to obtain stochastic realizations with the same range of behavior as MD.

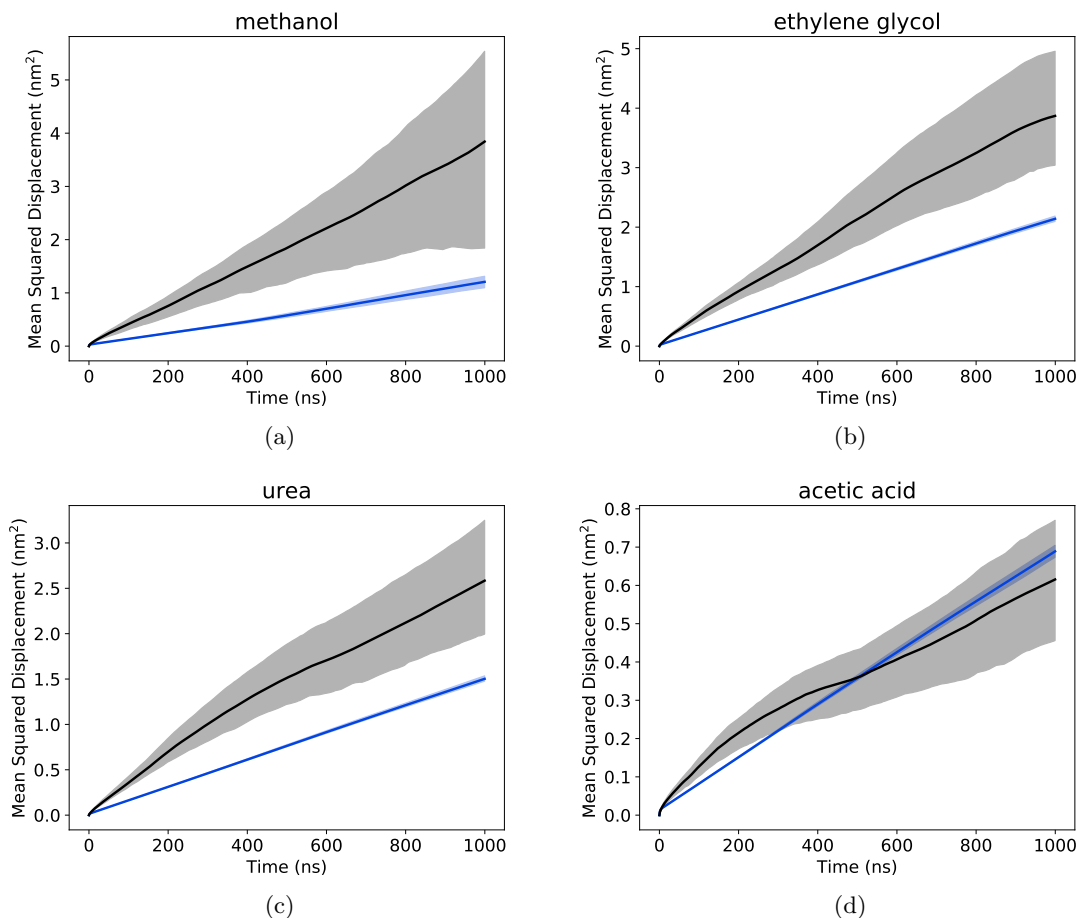


Figure 7: The MSDs predicted based on 1000 realizations of our model (blue) tend to underpredict the MSDs based on the MD trajectories (black). Due to the variability in solute behavior across MD trajectories, as demonstrated in Figure 6, there are several individual solute MSD curves that are significantly higher than the rest and drive up the mean of the MSD. If we remove those trajectories from the MSD calculation, we obtain estimates much closer to the IHMM estimate. (I haven't done it yet, but could plot a third curve with outlier removed). Acetic acid, has no obvious outliers which is consistent with our model's prediction which lies within the  $1\sigma$  confidence interval of MD after a 1000 ns time lag.

The MSDs calculated by stochastic realizations of our model are quantitatively similar to MD.

- For all solutes except acetic acid, our predicted MSD curve lie close to but below the  $1\sigma$  confidence intervals of MD.
- However, in these cases, the average MD MSDs are driven up by single trajectories with uncharacteristically large MSDs.
- It is possible that an improved trajectory generation approach may allow for this highly variable behavior.
- If we remove these uncharacteristic trajectories, the agreement between MD and our realizations improves.
- Acetic acid has no obvious outliers and shows good agreement between MD and IHMM realizations.

### 3.3 Estimating Solute Flux and Selectivity

We can use realizations of our model in order to predict macroscopic flux and selectivity.

- For various length pores, we constructed distributions of first passage times for each solute and fit them to Equation 4 in order to estimate the mean first passage time (MFPT).
- We plotted the inverse of MFPT, the solute flux, versus pore length in Figure TBD.
- Flux rapidly decays with pore length.
- Selectivity is simply the ratio of the fluxes for this system.

We can use the slopes of the MSD curves to estimate selectivity in a simpler way.

- As in our previous work, we fit the flux curves to a decaying power law function of the form:

$$Ae^{-\beta} \tag{7}$$

- We find that in all cases,  $\beta$  is equal to 2. This is consistent with Brownian motion which implies long term linear behavior of the MSD curves.
- Since the diffusion constant is linearly proportional to the slope of the MSD, and selectivity in this system can be estimated as the ratio of diffusion constants, we can estimate selectivity directly based on the slopes of our stochastically simulated MSD curves.

## 4 Conclusion

We have shown that the IHMM can be used to parameterize solute time series with an unknown number of latent dynamical modes.

We showcase this modeling approach by example, but it is important to recognize the generality of this analysis.

We can use the IHMM to help identify mechanisms by relating the latent states to observed solute behavior.

We show how one can use the IHMM to predict macroscopic transport properties.

## Supporting Information

Detailed explanations and expansions upon the results and procedures mentioned in the main text are described in the Supporting Information. This information is available free of charge via the Internet at <http://pubs.acs.org>.

## Acknowledgements

This work was supported in part by the ACS Petroleum Research Fund grant #59814-ND7 and the Graduate Assistance in Areas of National Need (GAANN) fellowship which is funded by the U.S. Department of Education. Molecular simulations were performed using the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. Specifically, it used the Bridges system, which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC). This work also utilized the RMACC Summit supercomputer, which is supported by the National Science Foundation (awards ACI-1532235 and ACI-1532236), the University of Colorado Boulder, and Colorado State University. The Summit supercomputer is a joint effort of the University of Colorado Boulder and Colorado State University.

## References

- [1] J. R. Werber, C. O. Osuji, and M. Elimelech, “Materials for Next-Generation Desalination and Water Purification Membranes,” *Nat. Rev. Mater.*, vol. 1, p. 16018, May 2016.
- [2] B. J. Coscia, J. Yelk, M. A. Glaser, D. L. Gin, X. Feng, and M. R. Shirts, “Understanding the Nanoscale Structure of Inverted Hexagonal Phase Lyotropic Liquid Crystal Polymer Membranes,” *J. Phys. Chem. B*, vol. 123, pp. 289–309, Jan. 2019.
- [3] B. J. Coscia and M. R. Shirts, “Chemically Selective Transport in a Cross-Linked HII Phase Lyotropic Liquid Crystal Membrane,” *J. Phys. Chem. B*, June 2019.
- [4] M. J. Beal, Z. Ghahramani, and C. E. Rasmussen, “The Infinite Hidden Markov Model,” in *Advances in Neural Information Processing Systems 14: Proceedings of the 2001 Neural Information Processing Systems (NIPS) Conference*, vol. 1, (Cambridge, MA, US), pp. 577–585, MIT Press, Sept. 2002.
- [5] J. D. Hamilton, *Time Series Analysis*. Princeton University Press, 1994.
- [6] E. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, “Nonparametric Bayesian Learning of Switching Linear Dynamical Systems,” in *Advances in Neural Information Processing Systems 21* (D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, eds.), pp. 457–464, Curran Associates, Inc., 2009.
- [7] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, “The Sticky HDP-HMM: Bayesian Nonparametric Hidden Markov Models with Persistent States,” *MIT Laboratory for Information and Decision Systems*, p. 60, Nov. 2007.
- [8] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, “Hierarchical Dirichlet Processes,” *Journal of the American Statistical Association*, vol. 101, pp. 1566–1581, Dec. 2006.
- [9] J. Van Gael, Y. Saatchi, Y. W. Teh, and Z. Ghahramani, “Beam Sampling for the Infinite Hidden Markov Model,” in *Proceedings of the 25th international conference on Machine learning - ICML '08*, (Helsinki, Finland), pp. 1088–1095, ACM Press, 2008.
- [10] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, “Bayesian Nonparametric Methods for Learning Markov Switching Processes,” *IEEE Signal Processing Magazine*, vol. 27, pp. 43–54, Nov. 2010.
- [11] E. L. Cussler, *Diffusion: Mass Transfer in Fluid Systems*. Cambridge University Press, 3 ed., 2009.
- [12] T. L. Hill, *Free Energy Transduction and Biochemical Cycle Kinetics*. Springer-Verlag, 1989.