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

March 23, 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 crosslinked 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 μs simulations of our previous work to 5 μs 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 distributions over a countably infinite number of probability measures:

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

  where the θ_k are values drawn from the base distribution and the weights β_k come from a stick-breaking process parameterized by the concentration parameter γ (equivalently referred to as GEM(γ)).
- The concentration parameter expresses one's confidence in H relative to the posterior and is closely related to the number of data observations.
- Each row, G_j, of the transition matrix is produced by drawing from a DP specified using the β vector as a discrete base distribution and a separate concentration parameter, α.

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

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

We describe the dynamics of each state 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:

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

  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, Σ.
- We limited our analysis to an autoregressive order of $r = 1$.

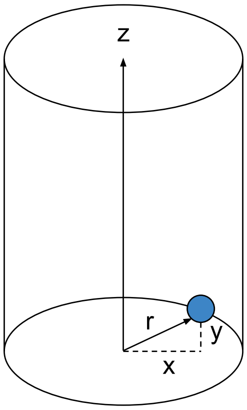

Figure 1: Using the nearest pore center as the origin, we represented the solute's location along the pore axis in terms of the z coordinates and their radial distance from the pore centers in 2 dimensions, x and y.

- We used a matrix-normal inverse-Wishart prior on parameters A and Σ 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 while simultaneously estimating 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 as shown in SI.
- We refer the interested reader to much more extensive descriptions of this process and its implementation. [4, 8, 9, 6, 10]

There are many ways one can apply the IHMM algorithm to timeseries data.

- We begin by trying to identify as many distinct states as possible in each solute trajectory and then take advantage of the system's cylindrical symmetry in order to cluster the parameter sets into an interpretable number of states.
- Having a large number of states gives more reliable clustering results.

We first applied the IHMM algorithm to 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 using 2 dimensions, rather than a single radial dimension, we increase the likelihood of finding more states since a single radial mean can be represented by an infinite number of x and y coordinate combinations.

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.

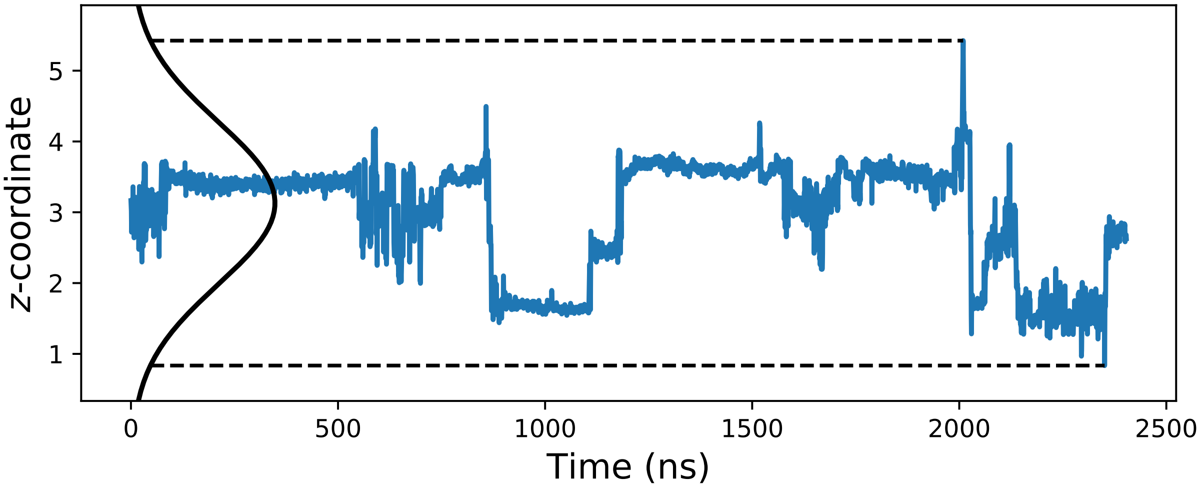

Figure 2: The parameters of the prior on c (black line) should be chosen such that the mean levels of each state identified in the trajectory (blue line) lie within regions of the prior with reasonable probability. We chose the mean of the prior as halfway between the maximum and minimum of each trajectory dimension. We chose σ of the prior by defining the maximum and minimum to be 2 standard deviations from the mean (dashed lines).

The states predicted by the IHMM are heavily influenced by the Gaussian prior placed on c in Equation 3.

- The entries of the A and $\mathbf{e}_t$ matrices 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 means of each state lie within a region of reasonable probability of the prior (see Figure 2).
- In each dimension, we defined the prior mean to be 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 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 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 averaged the parameters from each iteration recorded after the equilibration time point.
- We detected equilibration of the parameters using the module `pymbar.timeseries.detect_equilibration` on the time series of parameter estimates.
- Since all of the parameters are multidimensional, we found the equilibration time point of each component 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 the mean in r is a useful clustering variable.

- It will also be clearer to analyze the final parameter set in terms of r rather than x and y.
- We converted the x and y center-of-mass coordinates to r.
- In order to keep the same states found when we applied the algorithm in 3D, we reapplied the IHMM to the cylindrical trajectories with a fixed state sequence, only allowing inference on the VAR(1) parameters.

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

- For each solute studied, we identified between 200-300 independent sets of parameters.
- Many of these states exhibit very similar dynamical behavior except their mean levels are different, especially in the axial direction where solutes could get trapped along a broad and continuous range of z coordinates.
- Since we do not know the number of states with shared dynamical behavior beforehand, we used a non-parametric Bayesian Gaussian mixture model in order to group them.

We clustered based on 5 variables: the radial means of each state, the two eigenvalues of A and the two eigenvalues of e_t.

- Unfortunately, the number of independent parameter sets is too low for clustering in higher dimensions.
- Ideally, we could use all of the entries of A and e_t in addition to the radial means.
- It would not be helpful to use the means in the z direction because solutes are essentially unbound in that direction, while
- However, in our previous work, we observed radially dependent dynamical behavior. [3]

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.

We obtained $\mathbf{c}$ vectors of the clustered states by averaging each value of $\mathbf{c}$ assigned to the same cluster.

- We only care about the r component of $\mathbf{c}$ because solute trajectories are not bound in the z direction.

We used the IHMM algorithm in order to infer A and $\mathbf{e}_t$ 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 fluctations.
- We did this by subtracting the maximum likelihood estimate of the $\mathbf{c}$ vector from each same-state segement 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 weighted by the the rows of 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 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]

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

Flux, J, is simply 1 / 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:

$$S_{ij} = \frac{J_i/\Delta C_i}{J_j/\Delta C_j} \tag{5}$$

where ΔC_j is the trans-membrane concentration difference.

3 Results and Discussion

3.1 Inferring Solute Transport Mechanisms

Clustering parameters sets results in X distinct dynamical modes.

- In the figure below, we show time series simulations that qualitatively illustrate the difference in dynamical behavior between modes.

We can relate the identified states back to transport mechanisms.

- More detailed discussion of identified states
- How size of fluctuations, autoregressive parameters are influenced by trapping mechanisms
- How do these states compare to those identified in our previous work?
- Any new states?

3.2 Reproducing MD Trajectories and MSDs with the IHMM

Trajectory realizations qualitatively match MD simulation trajectories.

- Look for hopping and trapping behavior

MSDs generated from stochastic trajectories match those from MD.

- Look at curvature and 1-σ confidence intervals

3.3 Estimating Solute Flux and Selectivity

We can predict macroscopic flux and selectivity.

- Flux as function of pore length
- Selectivity as function of pore length (if flux scaling is length-dependent)

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

We can use the IHMM to predict macroscopic transport properties.

The IHMM is not limited to the H_{II} phase.

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. Saatci, 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.