Using Stochastic Modeling to Predict Long Timescale Transport Behavior of Solutes in an H_{II} Phase Lyotropic Liquid Crystal Membrane

Benjamin J. Coscia Michael R. Shirts

August 1, 2019

1 Introduction

We need highly selective membranes in order to perform efficient separations.

Amphiphilic molecules are capable of self-assembling into ordered nanostructures.

Lytropic liquid crystals are a class of amphiphilic molecules that can be cross-linked into mechanically strong membranes.

- H_{II} phase lyotropic liquid crystals have densely packed, uniform sized pores and 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.
- Q_I phase LLCs consist of a tortuous network of 3D interconnected pores.

We can only learn so much from experiment. MD 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 studied the transport of 20 small polar molecules in an H_{II} phase LLC membrane.

- In general, we observed subdiffusive transport behavior characterized by intermittent hops separated by periods of entrapment.
- We identified three mechanisms responsible for the solute trapping behavior: entanglement among monomer tails, hydrogen bonding with monomer head groups, and association with sodium counter ions.

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.

In this work, we explore three approaches to model the long timescale behavior of the four solutes fasted moving solutes studied in our previous work.

- Specifically, we study methanol, urea, ethylene glycol and acetic acid
- Our first approach is based on the anomalous diffusion literature.
- The second approach uses Markov state models (MSMs)
- The third uses an infinite hidden Markov model (iHMM), an unsupervised machine learning algorithm which groups solute behavior into different dynamical states.

2 Methods

We ran all MD simulations and energy minimizations using GROMACS 2018. We performed all post-simulation trajectory 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 explore as much structural space as possible.

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.[1]

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.

2.2 Modeling subdiffusion

Solutes in our H_{II} LLC membrane system exhibit subdiffusive behavior, a type of anomalous diffusion.

- During an anomalous diffusion process, the mean squared displacement (MSD) does not grow linearly with time, rather it is of the form:
$$\langle x^2(t)\rangle = K_\alpha t^\alpha \tag{1}$$
where α is the anomalous exponent and K_α is the generalized diffusion coefficient.
- A value of $\alpha < 1$ indicates a subdiffusive process, while values of $\alpha = 1$ and $\alpha > 0$ are characteristic of Brownian and superdiffusive motion respectively.

We analyzed both the ensemble-averaged and time-averaged MSDs of the simulated trajectories.

- The ensemble-averaged MSD measures the displacement of a particle from its initial position [2] and can be written as
$$\langle x^2(t)\rangle = \langle x(t) - x(0)\rangle \tag{2}$$
- The time-averaged MSD measures the displacement between all possible time lags and can be written as
$$\overline{x^2(\tau)} = \frac{1}{T-\tau}\int_0^{T-\tau} (x(t+\tau) - x(t))^2 dt \tag{3}$$
where τ is the time lag and T is the length of the trajectory [2].

Three common mathematical models for modeling anomalous subdiffusion processes include continuous time random walks (CTRW), fractional Brownian motion (FBM) and random walks on fractals (RWF).[2]

- FBM is common in crowded, viscoelastic environments where each step comes from a Gaussian distribution but is anti-correlated to its previous step. [3, 4, 5]
- A CTRW is characterized by a distribution of hop lengths and dwell times, where each each step is characterized by independent random draws from each distribution.[6, 7]
- An RWF is imposed by a system's geometry. Systems with tortuous pathways and dead ends cause anti-correlated motion.[2, 8]
- The processes described above can happen alone or in combination.

We believe that solutes in the system studied here exhibit subordinated fractional Brownian motion (sFBM) where the parent process is FBM and the leading process is a CTRW.

- The ensemble-averaged MSD differs from the time-averaged MSD, which is indicative of non-ergodicity, a trait inherent to CTRWs but not FBM or RWFs. [9]
- We also observe non-stationary z-coordinate traces of each solute's center of mass (COM).
- For a pure CTRW, the time-averaged MSD should be linear. [8, 10]
- However, a typical time-averaged solute MSD is sublinear (see supporting information), which suggests that there is another underlying subdiffusive mechanism.
- The hop lengths recorded after each dwell period are anti-correlated (See supporting information)
- Given the viscoelastic nature of the monomers in our system, we believe the hop lengths can be modeled with FBM.
- For subordinated FBM, it can be shown that

$$\langle x^2(t) \rangle \simeq t^{\alpha\beta} \tag{4}$$

  where α is the anomalous exponent characteristic of the leading CTRW process and β is the anomalous exponent characteristic of the parent FBM process.

We can characterize a CTRW process using the parameters which describe its dwell time and hop length distribution.

- We used the `ruptures` python package in order to identify changepoints in solute trajectories.[11] (See Supporting Information for more details on chosen parameters. i.e. type of cost function, cost function penalty tolerance, number of dimensions used)
- We used the corresponding hop lengths and dwell times between break points to construct empirical distributions.

For solutes in our system, the distribution of hop lengths appears to be well-represented by a Gaussian distribution. [12, 13, 14]

- We are most interested in the standard deviation, σ, of the hop length distribution.

The distribution of dwell times is expected to fit a power law (or heavy-tailed) distribution proportional to $t^{-1-\alpha}$. [2]

- Because we are limited to taking measurements at discrete values dictated by the output frequency of our simulation trajectories, we fit the empirical dwell times to a discrete power law distribution whose maximum likelihood α parameter we calculated by maximizing the following likelihood function:

$$\mathcal{L}(\beta) = -n \ln \zeta(\beta, x_{min}) - \beta \sum_{i=1}^{n} \ln\, x_i \tag{5}$$

  where $\beta = 1 + \alpha$, x_i are collected dwell time data points, n the total number of data points, and ζ is the Hurwitz zeta function where x_{min} is the smallest measured value of x_i. [15]

- We obtained distributions of the hop length standard deviations, σ, and α using statistical bootstrapping.[16]

FBM processes can be described using the Hurst parameter, H, where $H = \beta/2$.

- Brownian motion is recovered for $H = 0.5$
- The autocovariance function of hop lengths has the analytical form: [3]

$$\gamma(k) = \frac{1}{2}\left[|k-1|^{2H} - 2|k|^{2H} + |k+1|^{2H}\right] \tag{6}$$

  where k is the number of increments between hops.

- We obtained H by performing a least squares fit of Equation 6 to the empirically measured autocovariance function.
- We used statistical bootstrapping to generate a distribution of H values.

In general, we observe different dynamical behavior when solutes move inside the pore versus in the tail region.

- Therefore, we created 2 sets of parameters based on the solute's radial location.
- Solutes in the pore region are located less than 0.75 nm from the pore center and all else are in the tail region.
- We determined this cut-off and described how we calculated radial distance from the pore center in our previous work [1]

For each solute, we simulated 10000 5 μs sFBM trajectories.

- We constructed trajectories by simulating sequences of dwell times and correlated hop lengths generated based on parameters randomly chosen from our bootstrapped parameter distributions.
- We propagated each trajectory until the total time reached 1 μs, and truncated the last data point so that the total time exactly equaled 5 μs.
- Valid comparisons are only possible between fixed length sFBM simulations. The power law dwell time behavior gives rise to the aging phenomenon, embodied by a decrease in MSD with measurement time. [8, 13]
- We reported the MSD after 5 μs with corresponding 95 % intervals

2.3 Markov State Models

A Markov state model decomposes a time series into a set of discrete states with transitions between states defined by a transition probability matrix, T, which describes the conditional probability of jumping to a specific state given the previously observed state. [17, 18]

- In the context of molecular simulations, MSMs are frequently used to study systems with slow dynamics, such as protein folding. [19, 20]
- Researchers typically aim to come up with a low dimensional representation of the system based on features which preserve the process kinetics. This facilitates the identification of discrete states from which T is generated.
- Software packages such as MSMbuilder [21] and pyEMMA [22] provide workflows capable of featurization and dimensional reduction

This typically requires one to identify states by featurization which involves dimensional reduction of complex data into features with the most dynamical variance and identifying which independent components contribute the most to that variance. The identified features are often too abstract to be easily discernible. We have already featurized the system in our previous work by identifying three different trapping states.

2.4 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 only on a previous unobserved, or hidden, state. [23]

- 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.
- Unfortunately, standard HMMs require the number of hidden states to be known a priori.

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:

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k} \quad \theta_k \sim H, \beta \sim GEM(\gamma) \tag{7}$$

  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.

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k} \quad \pi_j \sim DP(\alpha, \beta) \tag{8}$$

- This hierarchical specification ensures that the transition probabilities in each row share the same support points $\{\theta_1, ..., \theta_k\}$.

We describe the dynamics of each state using a vector autoregressive (VAR) model.

- A VAR process is characterized by a vector of observations in a time series that are dependent on r previous values of the time series vector, weighted by a coefficient matrix A_i in addition to a white noise term $\mathbf{e}_t$:

$$\mathbf{y}_t = \sum_{i=1}^{r} A_i \mathbf{y}_{t-i} + \mathbf{e}_t \qquad \mathbf{e}_t \sim N(0, \Sigma) \tag{9}$$

- We assumed multivariate Gaussian noise and limited our analysis to an autoregressive order of $r = 1$.
- We used a conjugate matrix-normal inverse-Wishart prior on parameters A and Σ in order to analytically draw from the posterior.

Based on the VAR parameters and matrix of transition probabilities, we calculated the most likely sequence of hidden states.

- We repeated this process iteratively until we reached convergence
- Our python implementation of this process is heavily adapted from the MATLAB code of Fox et al. [24]
- We refer the interested reader to much more extensive descriptions of this process and its implementation. [23, 25, 26, 27, 28]

3 Results and Discussion

3.1 Subordinated Fractional Brownian Motion Modeling

We calculated the MSD of each solute in the set over the course of 1μs MD simulations.

- We extracted values of σ, α and H for each solute and then simulated 10000 sFBM trajectories of the same length as our MD simulations, as described in Section **??** of the Methods.
- The final MSDs of the sFBM trajectories are compared to those calculated directly from MD simulations in Figure 1.
- The three sFBM parameters, calculated based on MD solute trajectories are presented in Table 1.
- We would like to emphasize that we rely on the MD MSD values in order to define trends in the total MSD, while the sFBM trajectories and parameter values allow us to speculate as to the reasons for the observed trends.
- There is a non-negligible amount of error in the calculation of each parameter which prevents us from reliably portraying our sFBM MSDs as reduced uncertainty MD MSDs.

System	σ (nm)	α	H
Methanol	0.46	0.85	0.40
Urea	0.33	0.64	0.40
Ethylene Glycol	0.35	0.64	0.36
Acetic Acid	0.28	0.51	0.44

Table 1: We calculated values σ, α and H from MD simulation trajectories and then computed the average ensemble-averaged MSD of 10000 simulated trajectories.

Our simulated sFBM MSDs qualitatively reproduce the MD trends.

- sFBM generally lower
- There are a few noticeable discrepancies.
- We will discuss reasons for these as we proceed with the discussion.

It most cases, it is easy to relate σ, α and H to the simulated MSD values presented in Table 1.

- Higher values of σ indicate larger average hop lengths.
- Higher values of α mean that there will be less sampling of long dwell times.
- Values of H near the Brownian limit of 0.5, indicate a lower degree of anti-correlation.
- All of which contribute to an overall increase in the simulated MSD

We fit a CTRW model (see Section **??**) to these trajectories and predicted macroscopic diffusion coefficients which are presented in Table TBD.

- We confirmed that our molecules exhibit anomalous diffusion by fitting a power law to the MSD curve. The exponent is less than 1, indicative of anomalous subdiffusion.
- We used the decision making process given by Meroz and Sokolov in order to identify the appropriate subdiffusion model to use based on our time series (See Section S-TBD of the Supporting info for more details).
- The motion of the solutes is non-ergodic and the steps are uncorrelated which tells us the system is likely well-described by a CTRW.

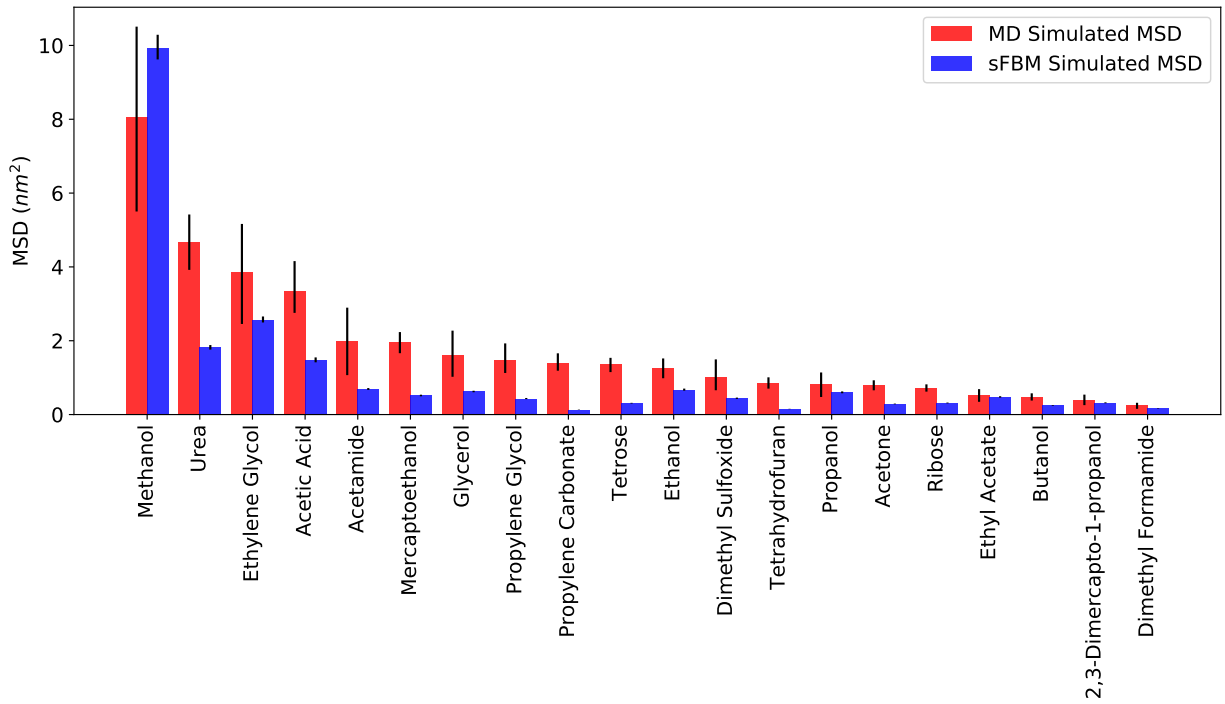

Figure 1

4 Markov Modeling

5 Conclusion

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

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] 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.

[2] Y. Meroz and I. M. Sokolov, "A Toolbox for Determining Subdiffusive Mechanisms," *Phys. Rep.*, vol. 573, pp. 1–29, Apr. 2015.

[3] B. Mandelbrot and J. Van Ness, "Fractional Brownian Motions, Fractional Noises and Applications," *SIAM Rev.*, vol. 10, pp. 422–437, Oct. 1968.

[4] J.-H. Jeon and R. Metzler, "Fractional Brownian Motion and Motion Governed by the Fractional Langevin Equation in Confined Geometries," *Phys. Rev. E*, vol. 81, p. 021103, Feb. 2010.

[5] D. S. Banks and C. Fradin, "Anomalous Diffusion of Proteins Due to Molecular Crowding," *Biophys. J.*, vol. 89, pp. 2960–2971, Nov. 2005.

[6] E. W. Montroll and G. H. Weiss, "Random Walks on Lattices. II," *Journal of Mathematical Physics*, vol. 6, pp. 167–181, Feb. 1965.

[7] G. T. Morrin and D. K. Schwartz, "Three Regimes of Polymer Surface Dynamics under Crowded Conditions," *Macromolecules*, vol. 51, pp. 1207–1214, Feb. 2018.

[8] T. Neusius, I. Daidone, I. M. Sokolov, and J. C. Smith, "Subdiffusion in Peptides Originates from the Fractal-Like Structure of Configuration Space," *Phys. Rev. Lett.*, vol. 100, p. 188103, May 2008.

[9] F. Thiel and I. M. Sokolov, "Weak Ergodicity Breaking in an Anomalous Diffusion Process of Mixed Origins," *Phys. Rev. E*, vol. 89, p. 012136, Jan. 2014.

[10] Y. Meroz, I. M. Sokolov, and J. Klafter, "Subdiffusion of Mixed Origins: When Ergodicity and Nonergodicity Coexist," *Phys. Rev. E*, vol. 81, p. 010101, Jan. 2010.

[11] C. Truong, L. Oudre, and N. Vayatis, "Ruptures: Change Point Detection in Python," *arXiv*, vol. arXiv:1801.00826, Jan. 2018.

[12] R. Metzler and J. Klafter, "The Random Walk's Guide to Anomalous Diffusion: A Fractional Dynamics Approach," *Phys. Rep.*, vol. 339, pp. 1–77, Dec. 2000.

[13] R. Metzler, J.-H. Jeon, A. G. Cherstvy, and E. Barkai, "Anomalous Diffusion Models and Their Properties: Non-Stationarity, Non-Ergodicity, and Ageing at the Centenary of Single Particle Tracking," *Phys. Chem. Chem. Phys.*, vol. 16, pp. 24128–24164, Oct. 2014.

[14] T. Neusius, I. M. Sokolov, and J. C. Smith, "Subdiffusion in Time-Averaged, Confined Random Walks," *Phys. Rev. E*, vol. 80, p. 011109, July 2009.

[15] A. Clauset, C. Shalizi, and M. Newman, "Power-Law Distributions in Empirical Data," *SIAM Rev.*, vol. 51, pp. 661–703, Nov. 2009.

[16] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. CRC Press: Boca Raton, May 1994.

[17] V. S. Pande, K. Beauchamp, and G. R. Bowman, "Everything you wanted to know about Markov State Models but were afraid to ask," *Methods*, vol. 52, pp. 99–105, Sept. 2010.

[18] C. Wehmeyer, M. K. Scherer, T. Hempel, B. E. Husic, S. Olsson, and F. No, "Introduction to Markov state modeling with the PyEMMA software [Article v1.0]," *Living Journal of Computational Molecular Science*, vol. 1, no. 1, pp. 5965–, 2018.

[19] C. D. Snow, E. J. Sorin, Y. M. Rhee, and V. S. Pande, "How Well Can Simulation Predict Protein Folding Kinetics and Thermodynamics?," *Annu. Rev. Biophys. Biomol. Struct.*, vol. 34, pp. 43–69, May 2005.

[20] J. D. Chodera, N. Singhal, V. S. Pande, K. A. Dill, and W. C. Swope, "Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics," *J. Chem. Phys.*, vol. 126, p. 155101, Apr. 2007.

[21] K. A. Beauchamp, G. R. Bowman, T. J. Lane, L. Maibaum, I. S. Haque, and V. S. Pande, "MSMBuilder2: Modeling Conformational Dynamics on the Picosecond to Millisecond Scale," *J. Chem. Theory Comput.*, vol. 7, pp. 3412–3419, Oct. 2011.

[22] M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Prez-Hernndez, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz, and F. No, "PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models," *J. Chem. Theory Comput.*, vol. 11, pp. 5525–5542, Nov. 2015.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

TOC Graphic