Capturing Subdiffusive Solute Dynamics and Predicting Selectivity in Nanoscale Pores with Time Series Modeling

Benjamin J. Coscia and Michael R. Shirts

Department of Chemical and Biological Engineering,

University of Colorado Boulder, Boulder, Colorado 80309, United States

(Dated: April 16, 2020)

Abstract

Mathematically modeling complex transport phenomena at the molecular level can be a powerful tool for identifying transport mechanisms and predicting macroscopic properties. We use two different stochastic time series models, parameterized from long molecular dynamics (MD) simulation trajectories of a cross-linked H_{II} phase lyotropic liquid crystal (LLC) membrane, in order to predict solute mean squared displacements (MSDs) and solute flux, and thus solute selectivity, in macroscopic length pores.

First, using anomalous diffusion theory, we show how solute dynamics can be modeled as a fractional diffusion process subordinate to a continuous time random walk. From the MD simulations, we parameterize the distribution of dwell times, hop lengths between dwells and correlation between hops. We explore two variations of the anomalous diffusion modeling approach. The first variation applies a single set of parameters to the solute displacements and the second applies two sets of parameters based on the solute's radial distance from the closest pore center.

Next, we generalize Markov state models, treating the configurational states of the system as a Markov process where each state has distinct transport properties. For each state and transition between states, we parameterize the distribution and temporal correlation structure of positional fluctuations as a means of characterization and to allow us to predict solute MSDs. We show that both models reasonably reproduce the MSDs calculated from MD simulations. However, qualitative differences between MD and Markov state dependent model-generated trajectories may limit its usefulness.

Finally, we demonstrate how one can use these models to estimate flux of a solute across a macroscopic-length pore and, based on those quantities, the membrane's selectivity towards each solute. This work helps to connect microscopic chemically-dependent solute motions that do not follow simple diffusive behavior with macroscopic membrane performance.

I. INTRODUCTION

Highly selective separations membranes are desirable in numerous applications. The ability to efficiently separate ions from saline water sources using membranes has been actively pursued for years in an effort to create potable water for people in water-scarce regions. [**?**] Even in relatively safe municipal water supplies, there is a need for membranes

that can specifically separate potentially harmful organic micropollutants such as fertilizers, pesticides, pharmaceuticals and personal care products. [**?**] In the materials industry, there is a strong interest in creating breathable fabrics which selectively allow passage of water vapor. [**?**]

Lyotropic liquid crystals (LLCs) are a class of amphiphilic molecules that can be cross-linked into mechanically strong and highly selective nanoporous membranes. [**?**] They may provide a promising alternative to conventional membrane separation techniques by being selective based not only on solute size and charge, but on solute chemical functionality as well. [**?**] One can tune the functionality of LLC monomers in order to enhance or weaken specific solute-membrane interactions. [**?**] Two general morphologies of LLC membranes are being actively developed. The inverted hexagonal, H_{II}, LLC phase has densely packed, uniform-sized pores on the order of 1 nm in size. A perfectly synthesized H_{II} phase has an ideal geometry for high throughput separations. Unfortunately, aligning the microscopic hexagonal mesophases on a large scale as required for high permeability has shown limited success. [**?** **?**] Bicontinuous cubic, Q_I, phase LLC membranes share the uniform and nm-sized chemically complex pores of the H_{II} phase but its geometry consists of a tortuous network of 3D interconnected pores which do not require alignment. [**?**] While its pore structure may decrease a Q_I phase membrane's permeability relative to the H_{II} phase, it is considerably easier to produce at scale. [**?**]

Molecular modeling may make it possible to efficiently evaluate solute-specific separation membranes using the available chemical space of LLC monomers. To date, only a limited subset of LLC monomers have been studied experimentally in membrane applications. [**?** **?** **?** **?** **?**] Even within this small subset, they have demonstrated selectivities that could not be explained beyond speculation and vague empirical correlations. [**?**] Atomistic molecular modeling can provide the resolution necessary to identify molecular interactions that are key to separation mechanisms, allowing us to move beyond Edisonian design approaches.

In our previous work, [**?**] we used molecular dynamics (MD) simulations to study the transport of 20 small polar molecules in an H_{II} phase LLC membrane. We chose to study the H_{II} phase because it is a simpler geometry to model than the Q_I phase and significantly more structural data is available to help ensure molecular models are consistent with experiment. 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 the monomer's sodium counter ions.

Up through our previous studies, our molecular models have provided valuable qualitative mechanistic insight with some quantitative support. This insight already allows us to speculate about new LLC monomer designs. However, they would be of greater value to a larger set of researchers if we could provide quantitative predictions of macroscopic observables such as solute flux and selectivity. Due to the size of the types of systems we are studying (at least (62,000 atoms) it is prohibitively expensive and time-consuming to run simulations longer than those performed for this work ($\sim$5 μs). Even with the relatively large size of our system, we only simulate 24 solute trajectories per unit cell in order to minimize solute-solute interactions. This results in relatively high uncertainties in observables such as mean squared displacement (MSD), preventing reliable long timescale predictions. We are in need of a way of studying the collective motion of a much larger set of solutes over much longer timescales.

Mathematical descriptions of transport in complex separations membranes are a powerful way to understand mechanisms and formulate design principles. [**?** **?** **?**] The complexity of a well-fit model generally parallels the complexity of the transport mechanism being studied, as well as the transport information the model conveys. In dense homogeneous membranes, the solution-diffusion model can extract diffusion and partition coefficients and has successfully predicted solute transport rates. [**?**] Analogously, pore-flow models yield predictions of diffusion coefficients and solute transport rates in nanoporous membranes. [**?**] Modern single particle tracking approaches have taken researchers beyond continuum modeling allowing them to characterize complex diffusive behavior. [**?**] At the molecular level, one can use molecular dynamics (MD) simulations to study both single particle dynamics and bulk transport properties with atomic-level insight. [**?** **?**] All of these approaches facilitate generation of hypotheses about the molecular origins of separations by attempting to give a more intuitive understanding of how solutes move as a function of their environment, in turn suggesting experiments that could be performed.

Using a bottom-up modeling approach, we can parameterize single particle behavior by extracting particle trajectories from MD simulations and studying the properties that have the greatest influence on solute dynamics. With this information, we can learn how to

construct an ensemble of characteristic single solute trajectories with these properties which would be useful for making long timescale predictions with computational ease and lower uncertainties. Fortunately, there is an abundance of information encoded in the complex solute trajectories. We can incorporate fluctuations and time-dependent correlations in solute position into transport models, as well as the length of time solutes are trapped. We can add further detail by integrating this time series analysis with our knowledge of the primary trapping mechanisms as well as the solute's changing chemical environment within the heterogeneous membrane.

In this work, we use the output of our MD simulations to construct two classes of mathematical models which aim to predict membrane performance while providing quantitative mechanistic insights. The functional forms of these models are driven by mechanistic observations from our previous work and their inputs are parameterized using a substantial amount of data generated by long (≥ 5 μs) MD simulations.

We constructed our first model by applying the existing rigorous theoretical foundation which describes the motion of particles that exhibit non-Brownian, or anomalous, transport behavior. [**?** **?**] The tools introduced by fractional calculus are instrumental to this theory. [**?**] They allow us to generalize the normally linear diffusion equation to fractional derivative orders, providing descriptions of a much more diverse set of behavior, including subdiffusion, a type of anomalous diffusion exhibited by solutes in this study. [**?**]

Three well-known classes of behavior leading to anomalous subdiffusion are continuous time random walks (CTRWs), fractional Brownian motion (FBM) and random walks on fractals (RWFs).[**?**] These types of motion are frequently used alone or in combination to describe single particle trajectories. [**?** **?**] A CTRW is characterized by a distribution of hop lengths and dwell times, where trajectories consist of sequential independent random draws from each distribution.[**?**] FBM is common in crowded, viscoelastic environments where each jump comes from a Gaussian distribution but is anti-correlated to its previous steps. [**?** **?** **?**] An RWF is imposed by a system's geometry. Systems with tortuous pathways and dead ends cause anti-correlated motion.[**?** **?**]

We treated our system as an FBM process subordinate to a CTRW, or subordinated FBM (sFBM) for short. We fit sFBM models in two ways: First, we parameterized solute motion using a single set of parameters fit to a hypothesized anomalous diffusion (AD) model. Second, we used a two-state approach by extracting two sets of model parameters

calculated based on a solute's radial distance from the closest pore center. This allows us to include the different dynamical behavior of solutes previously observed within the pore region versus within the monomer tails.

Where sufficiently close to observed data, these fitted sFBM models provide a way to propagate information about solute trajectories gathered at the microsecond time scale to realistic experimental time scales. They can also confirm the importance of assumed solute-membrane interactions used to define the model's functional form, and suggest missing components. They can help researchers formulate hypotheses to explain why a certain set of parameters is characteristic to a specific solute.

If one has knowledge of the primary mechanisms leading to anomalous diffusion behavior, one may gain additional insight by formulating a state-based model, such as Markov state models (MSMs). MSMs are a popular class of models used to project long timescale system properties based on molecular simulation trajectories by identifying different dynamical modes and quantifying the rates of transitions between them. MSMs are frequently used to study systems with slow dynamics, such as protein folding. [**?** **?**] Researchers typically aim to come up with a low dimensional representation of the system based on features which preserve the process kinetics. This still often results in hundreds to thousands of distinct states. [**?**]

Our second modeling approach applies an extended MSM framework to a relatively small set of known states based on the three previously observed solute trapping mechanisms. Standard MSMs are typically applied to determine equilibrium populations of states and the kinetics of transitions between those states. [**?**] We extend the framework to include state-dependent fluctuations and correlations in solute position. The magnitude of a solute's fluctuations from its average trapped position and the degree of correlation with previous fluctuations is determined by the current state. To distinguish our approach from standard MSMs, we have named it the Markov state-dependent dynamical model (MSDDM).

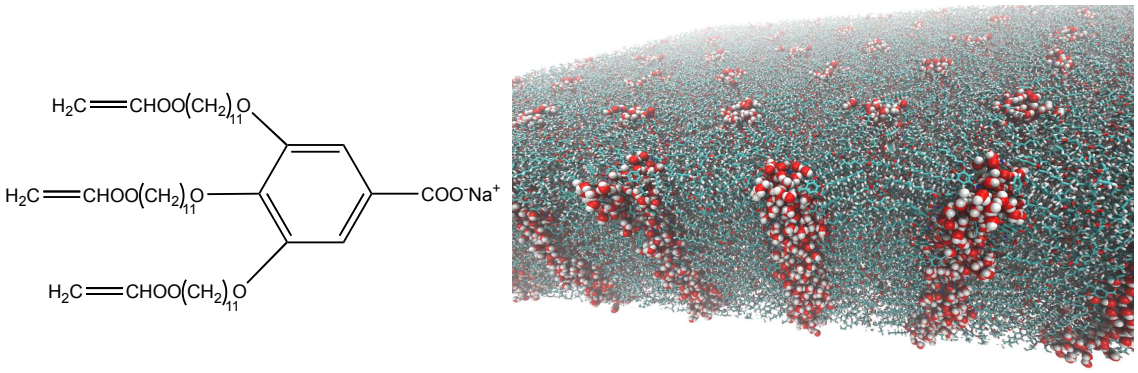
We determine the degree of success of our modeling approaches in two ways. First, if we can closely reproduce solute MSDs measured from MD simulations with realizations of our models, then it is likely that the model sufficiently captures solute dynamics and can be used in a predictive capacity. Even if a model fails to reproduce the MD MSDs, there is value in uncovering the cause of the deviation. The second measure of success is based on the qualitative comparison between individual realizations of solute trajectories generated

by our models and those observed from MD simulations. Even if we can reproduce the solute MSDs based on realizations of our models, the absence or inappropriate reproduction of hopping and trapping behavior may indicate underlying model issues.

The goal of this work is not to definitively determine which modeling approach is better but to evaluate their performance independently because they both have potential value dependent on a given research study's goals. The AD approach provides a systematic way to compare the dynamical behavior of different solutes based solely on the time series of their center of mass positions. The process of choosing the correct AD approach provides its own mechanistic insight since it requires a thorough analysis of solute time series behavior. One can simulate different solutes, compare the fit model parameters and relate them back to differences in solute size and chemical composition. The MSDDM characterizes explicitly defined trapping mechanisms, providing a clear picture of solute behavior while in each trapped state as well as the equilibrium occupation of each trapping state. It is possible to use the two modeling approaches in tandem, the AD approach to identify mechanisms, and the MSDDM to characterize mechanisms.

We evaluate the two modeling approaches by using them to characterize the dynamical behavior of the four fastest moving solutes studied in our previous work: methanol, urea, ethylene glycol and acetic acid. In addition to moving more quickly than other solutes studied previously, allowing them to extensively explore membrane structural space, these solutes have a range of chemical functionality and experience each of the three trapping mechanisms to different extents.

Finally, we use both models in order to predict solute flux and selectivity in pores of macroscopic length, thus achieving a better understanding of macroscopic properties on the basis of microscopic dynamics. We show that anti-correlated solute hopping behavior severely reduces flux relative to uncorrelated behavior. We use the relative solute fluxes to estimate the membrane's selectivity between all pairs of solutes and demonstrate that, when solutes display different degrees of anti-correlated hopping behavior, selectivity becomes a function of pore length. With this improved understanding of macroscopic behavior, we can begin to think more critically about how to design membranes in order to selectively pass or reject specific solutes.



(a) (b)

FIG. 1: (a) The wedge-shaped liquid crystal monomer Na-GA3C11 will form the inverted hexagonal phase in the presence of water where the carboxylate head groups occupy the pore centers. (b) A cross-section of a periodically replicated atomistic unit cell used for simulations in this study reveals the membrane's aqueous, hexagonally packed, straight and uniform sized pores. Water molecules (red and white spheres) present in the tail region are omitted for clarity.

II. 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`. The appropriate scripts to use for subsequent calculations are summarized in Table S1 of the Supplemental Material.

A. 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 (see Figure **??**). Approximately one third of the water molecules occupy the tail region with the rest near the pore center. We chose to focus our study on the 10 wt% water system because solutes move significantly faster than in the 5 wt% system studied previously, thus allowing more statistically valid analyses.

We chose to study a subset of the 4 fastest moving solutes from our previous work: methanol, acetic acid, urea and ethylene glycol. For each solute we created a separate system and to each system we added 6 solutes per pore for a total of 24 solutes. On the time scales which we simulate, this number of solutes per pore provides a sufficient amount of data from which to generate statistics. It also maintains a low degree of interaction between solutes since, at present, we are primarily interested in solute-membrane interactions. Further details on the setup and equilibration of these systems are described in our previous work.[**?**]

We extended the 1 μs simulations of our previous work to 5 μs in order to collect ample data. We ran MD simulations using the leapfrog integrator with hydrogen bonds constrained by the LINCS algorithm. We simulated a time step of 2 fs at a pressure of 1 bar and 300 K controlled by the Parrinello-Rahman barostat and the v-rescale thermostat respectively. We recorded frames every 0.5 ns.

We considered each system to be equilibrated when the solute partitioning between the pore and tail region reached apparent equilibrium and use only data after that point in our analysis of solute kinetics. We plot the solute partition versus time in Figure S1 of the Supplemental Material in order to justify our choice of equilibration times for each solute.

B. The Anomalous Diffusion Model

Solutes in this system very clearly exhibit subdiffusive behavior, a type of anomalous diffusion. During an anomalous diffusion process, the mean squared displacement (MSD) does not grow linearly with time, but instead it follows a power law of the form:

$$\langle z^2(t) \rangle = K_\alpha t^\alpha \tag{1}$$

where α is the anomalous exponent and K_α is the generalized diffusion coefficient. In this work, we only consider the MSD with respect to the solutes' center of mass z-coordinate, which is oriented along the pore axis. A value of $\alpha < 1$ indicates a subdiffusive process, while values of $\alpha = 1$ and $\alpha > 1$ are characteristic of Brownian and superdiffusive motion respectively.

While it is theoretically possible to extract the value of α by fitting Equation **??** to an MSD generated directly from MD simulations, we do not simulate enough independent

solute trajectories to obtain a reliable estimate. We can obtain higher precision estimates of α using the dwell time distributions described in subsequent sections.

In this study, we primarily use the MSD as a tool for characterizing the average dynamic behavior of solute trajectories. Rather than using them to calculate diffusion constants or to relate our simulations to experimental measurements, we compare MSDs calculated from MD simulations to those generated from our models in order to validate those models. Therefore, it is only important that we use a consistent definition for calculating the MSD between modeled trajectories and directly observed MD trajectories.

One can measure MSD in two ways. The ensemble averaged MSD measures displacements with respect to a particle's initial position:

$$\langle z^2(t) \rangle = \langle z(t) - z(0) \rangle^2 \tag{2}$$

Fits to the ensemble averaged MSD will always reproduce the form of Equation **??**. The time-averaged MSD measures all observed displacements over time lag τ:

$$\overline{z^2(\tau)} = \frac{1}{T-\tau} \int_0^{T-\tau} (z(t+\tau) - z(t))^2 dt \tag{3}$$

where T is the length of the trajectory.

The time averaged and ensemble averaged MSDs will give identical results unless a system displays non-ergodic behavior. For a pure CTRW, the power law distribution of trapping times leads to weak ergodicity breaking. In this case, the time-averaged MSD is linear while the ensemble averaged MSD has the form of Equation **??**. [**?**] With power law trapping behavior, the time between hops diverges so there is no characteristic measurement time scale of solute motion. In fact, as measurement time increases, the average MSD of a CTRW tends to decrease, a phenomenon called aging, because trajectories with trapping times on the order of the measurement time get incorporated into the calculation. [**?**]

We chose to use just the time-averaged MSD to compare MD trajectories with modeled trajectories, because, compared to the ensemble average, it is a more statistically robust measure of the average distance a solute travels over time. The ensemble MSD of only 24 solute trajectories would have much higher uncertainties.

1. Subordinated Fractional Brownian Motion

One can characterize the CTRW component of an sFBM process by the parameters which describe its dwell time and hop length distributions. We used the `ruptures` Python package in order to automatically identify mean shifts in solute trajectories, indicating hops.[**?**] We used the corresponding hop lengths and dwell times between hops to construct empirical distributions.

Dwell time distributions: For subdiffusive transport, the distribution of dwell times is expected to fit a power law distribution proportional to $t^{-1-\alpha}$. [**?**] Because we are limited to taking measurements at discrete intervals 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 log likelihood function:

$$\mathcal{L}(\beta) = -n \ln \zeta(\beta, t_{min}) - \beta \sum_{i=1}^{n} \ln t_i \tag{4}$$

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

In practical applications, the heavy tail of power law distributions can result in arbitrarily long dwell times that are never observed in MD simulations. In order to directly compare our anomalous diffusion model to finite-length MD trajectories we need to bound the dwell time distribution. A standard way of doing this, with easily estimated parameters, is by adding an exponential cut-off to the power law so the dwell time distribution is now proportional to $t^{-1-\alpha}e^{-\lambda t}$. [**?** **?**] We determine MLEs of α and λ by maximizing the log likelihood function: [**?**]

$$\mathcal{L}(\alpha, \lambda) = n(1-\alpha) \ln \lambda - n \ln \Gamma(1-\alpha, t_{min}\lambda) - \alpha \sum_{i=1}^{n} \ln t_i - \lambda \sum_{i=1}^{n} t_i \tag{5}$$

Correlated hop length distributions: The distribution of hop lengths by solutes undergoing an sFBM process is Gaussian, therefore we parameterize it by its standard deviation, σ. [**?** **?** **?**] The measured mean of the hop length distribution is always very close to zero so we assume that it is exactly zero in our time series simulations since we have no reason to expect drift in either direction based on pore symmetry.

sFBM implies that hops are correlated which we describe using the Hurst parameter, H.

The autocovariance function of hop lengths has the analytical form: [**?**]

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

where σ^2 is the variance of the underlying Gaussian distribution from which hops are drawn and k is the time lag, or number of increments between hops. The hop autocorrelation function is simply Equation **??** normalized by the variance. When $H < 0.5$, hops are negatively correlated, when $H = 0.5$ we recover Brownian motion and when $H > 0.5$, one observes positive correlation between hops.

There are many methods for estimating the Hurst parameter for a time series. [**?**] It can be a difficult task because Equation **??** decays slowly to zero, especially when $H > 0.5$, meaning one needs to study large time lags with high frequency. Fortunately (from a mathematical perspective), all of our solutes show anti-correlated motion, so most of the information in Equation **??** is contained within the first few time lags (see Figure S2a of the Supplemental Material). Any hope of fitting longer time lags to our data is lost in the noise since the analytical values are so close to zero. Therefore, we obtained H by performing a least squares fit of Equation **??** to the first ten 0.5 ns time lags of the empirically measured autocorrelation function.

2. Subordinated Fractional Lévy Motion

Because we also want to account for the possibility that the distribution of hops is not Gaussian, we can model them with the more general class of Lévy stable distributions. For independent and identically distributed random variables, the generalized central limit theorem assures convergence of the associated probability distribution function (PDF) to a Lévy stable PDF. [**?**] The characteristic equation which describes the Fourier transform of a Lévy stable PDF is:

$$p_{\alpha_h,\beta}(k;\mu,\sigma) = \exp\left[i\mu k - \sigma^{\alpha_h}|k|^{\alpha_h}\left(1 - i\beta\frac{k}{|k|}\omega(k,\alpha_h)\right)\right] \tag{7}$$

where

$$\omega(k,\alpha_h) = \begin{cases} \tan\frac{\pi\alpha_h}{2} & \text{if } \alpha_h \neq 1, 0 < \alpha_h < 2, \\ -\frac{2}{\pi}\ln|k| & \text{if } \alpha_h = 1 \end{cases}$$

α_h is the index of stability or Lévy index, β is the skewness parameter, μ is the shift parameter and σ is a scale parameter. The most familiar case, and one of three that can be expressed in terms of elementary functions, is the Gaussian PDF ($\alpha_h = 2$, $\beta = 0$). We assume symmetric distributions centered about 0 implying that β and μ are both 0.

For correlated hops, solute behavior may be described by subordinated fractional Lévy motion (sFLM). The Hurst parameter can again be used to describe hop correlations because they share the same autocorrelation structure. [**?**] The autocovariance function for FLM is:

$$\gamma(k) = \frac{C}{2}\left[|k-1|^{2H} - 2|k|^{2H} + |k+1|^{2H}\right],\ C = \frac{E\left[L(1)^2\right]}{\Gamma(2H+1)\sin(\pi H)} \tag{8}$$

where $E\left[L(1)^2\right]$ is the expected value of squared draws from the underlying Lévy distribution, effectively the distribution's variance. [**?**] In general, most Lévy stable distributions have an undefined variance due to their heavy tails. However, normalizing Equation **??** by the variance of a finite number of draws from a Lévy stable distribution results in the same autocorrelation structure as FBM. See Section S3 of the Supplemental Material for numerical simulations illustrating this point.

Analogous to power law dwell times, the heavy tails of Lévy stable hop length distributions result in rare but arbitrarily long hops. These long and unrealistic hops result in over-estimated simulated MSDs (see Figure **??** for example). We observe that the distribution of hops observed in our MD simulations are well approximated by Lévy stable distributions close to the mean, but they significantly under-sample the tails. We chose to truncate the Lévy stable distributions based on where the theoretical probability distribution function (PDF) starts to deviate from the empirically measured PDF (see Section S4.1 of the Supplemental Material). [**?**]

Multiple Anomalous Diffusion Regimes

We observe different dynamical behavior when solutes move while inside the pore versus while in the tail region. This suggests two anomalous diffusion models of varying complexity. We first create a simple, single mode model with a single set of parameters fit to solute trajectories. Our second, two mode model assigns a set of parameters to each of 2 modes based on the solute's radial location. We define the first mode as the pore region, defined as less than 0.75 nm from any pore center. Solutes outside the pore region are in the second

mode, the tail region. We determined this cut-off by maximizing the difference in dynamical behavior as described in our previous work. [**?**] Unfortunately, there were not enough sufficiently long sequences of hops in each mode to reliably calculate a Hurst parameter for each mode so we used the single, average Hurst parameter from the single mode model for both modes of the two mode model.

For the two mode model, we defined a transition matrix describing the rate at which solutes moved between the tail and pore region. We assumed Markovian transitions between modes, meaning each transition had no memory of previously visited modes. We populated a 2×2 count matrix by incrementing the appropriate entry by 1 each time step and then generated a transition probability matrix by normalizing the entries in each row of the count matrix so that they summed to unity.

Simulating Anomalous Diffusion

We simulated models with all combinations of the types of dwell and hop length distributions described above, summarized in Table **??**. All models include correlation between hops.

Dwell Distribution	Hop Length Distribution	Abbreviation
Power Law	Gaussian	sFBM
Power Law w/ Exponential Cut-off	Gaussian	sFBMcut
Power Law	Lévy Stable	sFLM
Power Law w/ Exponential Cut-off	Lévy Stable	sFLMcut

TABLE I: We tested four anomalous diffusion models with various modifications to the dwell and hop length distributions. We incorporate hop correlation into all models.

For each solute, we simulated 1,000 anomalous diffusion trajectories of length T_{sim} in order to directly compare our model's predictions to MD simulations. T_{sim} varied between solutes due to differing solute equilibration times. 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 equaled or exceeded T_{sim} μs, then truncated the last data point so that the

total time exactly equaled T_{sim} μs since valid comparisons are only possible between fixed length sFBM simulations.

We used Equation **??** to calculate the time-averaged MSD of the MD and AD model trajectories then estimated their uncertainty using statistical bootstrapping. For each bootstrap trial, we randomly chose n solute trajectories, where n is the number of independent trajectories, with replacement, from the ensemble of trajectories and then calculated the MSD of the subset. We reported the time-averaged MSD up to a 1000 ns time lag with corresponding 1σ confidence intervals.

When simulating 2 mode models, we determined the state sequence based on random draws weighted by the appropriate row of the probability transition matrix. We then drew hops and dwells based on the current state of the system. Since we calculated the transition probabilities from a finite data set, they have an associated uncertainty which we incorporated by re-sampling each row from a two dimensional Dirichlet distribution (which is also a beta distribution for the 2D case) with concentration parameters defined by the count matrix. [**?**]

We used the Python package `fbm` [**?**] to generate exact simulations of FBM and our own Python implementation (see Table S1 of the Supplemental Material) of the algorithm by Stoev and Taqqu to simulate FLM. [**?**] Note that, to our knowledge, there are no known exact simulation algorithms for generating FLM trajectories. However, the algorithm we used sufficiently approximates draws from the marginal Lévy stable distribution and reasonably approximates the correlation structure on MD simulation timescales. We added an empirical correction to enhance the accuracy of the correlation structure (see Section S4.2 of the Supplemental Material for validation of the approach).

C. The Markov State-Dependent Dynamical Model

A Markov state model (MSM) decomposes a time series into a set of discrete states with transitions between states defined by a transition probability matrix, T. T describes the conditional probability of moving to a specific state given the previously observed state. [**?** **?**]

In this work, we define a total of 8 discrete states based on the 3 trapping mechanisms observed in our previous work. Therefore, there is no need to apply any algorithmic ap-

proaches to identify and decompose our system into discrete states. The states we have chosen include all combinations of trapping mechanisms in the pore and out of the pore (see Table **??**). They assume that there are no significant kinetic effects resulting from solute conformational changes or pore size fluctuations. We use the same radial cut-off (0.75 nm) as in the AD approach to differentiate the pore and tail region. 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 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. [**?**]

Markov State-Dependent Dynamical Model State Definitions	
1. In tails, not trapped	5. In pores, not trapped
2. In tails and hydrogen bonding	6. In pores and hydrogen bonding
3. In tails and associated with sodium	7. In pores and associated with sodium
4. In tails, hydrogen bonding and associated	8. In pores, hydrogen bonding and associated

TABLE II: We defined 8 discrete states based on all combinations of previously observed solute trapping mechanisms.

We constructed the state transition probability matrix, T, based on observed solute trajectories. Using methods described in our previous work, we determined each solute's radial location and which, if any, trapping mechanisms affected it at each time step, then assigned the observation to a specific state according to the definitions in Table **??**. [**?**] Analogous to the mode transition matrix in Section **??**, and based on the current and previous state observation, we incremented the appropriate entry of an $n \times n$ count matrix by 1, where n is the number of states. We verified the Markovianity of state transitions as described in Section S5 of the Supplemental Material.

Adding to the standard MSM framework, we incorporated the dynamics of the solutes within each state as well as the dynamics of state transitions, which includes the overall configurational state of the solute and it's environments. While MSMs are often used to estimate equilibrium populations of various states, adding state-dependent dynamics allows us to simulate solute trajectories. Hence why we refer to them as Markov state-dependent dynamical models (MSDDMs).

We recorded the z-direction displacement at each time step in order to construct individual emission distributions for each state and transition between states. This results in 64 distinct emission distributions with some far more populated than others. We modeled all of the emission distributions as symmetric Lévy stable distributions in order to maintain flexibility in parameterizing the distributions.

We use the Hurst parameter to describe negative time series correlation. However, there is not sufficient data to accurately measure a Hurst parameter for each type of transition. We avoided this problem by combining all distributions associated with state transitions and treating all transitions as correlated emissions from a single Lévy stable distribution. This reduces the number of emission distributions from 64 to 9 (1 for each of the 8 states and 1 for transitions between states).

We simulated realizations of the MSDDM using the probability transition matrix and emission distributions. For each trajectory simulated, we chose an initial state randomly from a uniform distribution. We generated a full state sequence by randomly drawing subsequent states weighted by the rows of the probability transition matrix corresponding to the particle's current state. Again, because we are working with a finite data set, we incorporated transition probability uncertainties into the rows of the transition matrix by resampling them from a Dirichlet distribution. For each same-state subsequence of the full state sequence, we simulated an FLM process using the Hurst parameter of that state and the parameters of the corresponding emission distribution. Independently, we simulated the transition between each same-state sequence with an FLM process based on the Hurst parameter of transition sequences and the parameters of the single transition emission distribution. We used the same FLM simulation procedure described in Section **??**.

D. Estimating Solute Flux

We determine the rate at which solutes cross macroscopic-length pores based on the Hill relation: [**?**]

$$J = \frac{1}{MFPT} \tag{9}$$

where J is the single particle solute flux and MFPT refers to the mean first passage time. To account for input concentration dependence of the flux, assuming that particles are independent, one can multiply Equation **??** by the total number of particles to get the total

flux. In the context of our work, the MFPT describes the average length of time it takes a particle to move from the pore entrance to the pore exit.

We generated particle trajectories, parameterized with the above models, in order to construct a distribution of first passage times across a membrane pore of length L. For each pore length, we simulated 10,000 realizations of an AD approach model all released at the pore entrance ($z = 0$). In the case of uncorrelated hops, one can continuously draw from the hop length distribution until $z \geq L$ (or $-L$ for the sake of computational efficiency). The length of time between the last time the particle crossed $z = 0$ and the end of the trajectory gives a single passage time. When particle hops are correlated, as they are in all cases of this work, we cannot continuously construct the particle trajectories. Rather, we must generate trajectories of length n and measure the length of the sub-trajectory which traverses from 0 to L without becoming negative.

We calculated the expected value of analytical fits to the passage time distributions in order to determine the MFPT for a given solute and pore length. One should not use the mean of the empirical passage time distribution because it is highly likely that the true MFPT will be underestimated unless 100% of a very large number of trajectories reach L. If a trajectory does not reach L within n steps, it is possible that a very long passage time has been excluded from the distribution.

To derive an analytical equation describing the passage time distributions, one can frame the problem as a pulse of particles instantaneously released at the pore inlet ($z = 0$) which moves at a constant velocity, v, and spreads out as it approaches L. This spreading is parameterized by an effective diffusivity parameter, D. This approach gives results equivalent to if we had released each particle individually and then analyzed the positions of the ensemble of trajectories as a function of time. The analytical expression describing the distribution of first passage times is: [**?**]

$$P(t) = -\frac{1}{\sqrt{\pi}} e^{-(L-vt)^2/(4Dt)} \left(-\frac{D(L-vt)}{4(Dt)^{3/2}} - \frac{v}{2\sqrt{Dt}} \right) \tag{10}$$

where the only free parameters for fitting are v and D. A derivation of Equation **??** is given in Section S6 of the Supplemental Material. We calculated the expected value of Equation **??** in order to get the MFPT.

We used the ratio of solute fluxes in order to determine membrane selectivity, S_{ij}, towards solutes. Selectivity of solute i versus j is defined in terms of the ratio of solute permeabilities,

P: [**?**]

$$S_{ij} = \frac{P_i}{P_j} \tag{11}$$

We can relate this to solute flux using Kedem and Katchalsky's equations for solvent volumetric flux, J_v, and solute flux, J_s: [**?** **?**]

$$J_v = L_p(\Delta P - \sigma\Delta\pi) \tag{12}$$

$$J_s = P_s\Delta C + (1-\sigma)CJ_v \tag{13}$$

where L_p is the pure water permeability, ΔP and $\Delta\pi$ are the trans-membrane hydraulic and osmotic pressure differences, σ is the reflection coefficient, P_s is the solute permeability, ΔC is the trans-membrane solute concentration difference and C is the mean solute concentration. Since our simulations do not include convective solute flux, we eliminate the second term of Equation **??** which allows us to derive a simple expression for selectivity in terms of solute flux:

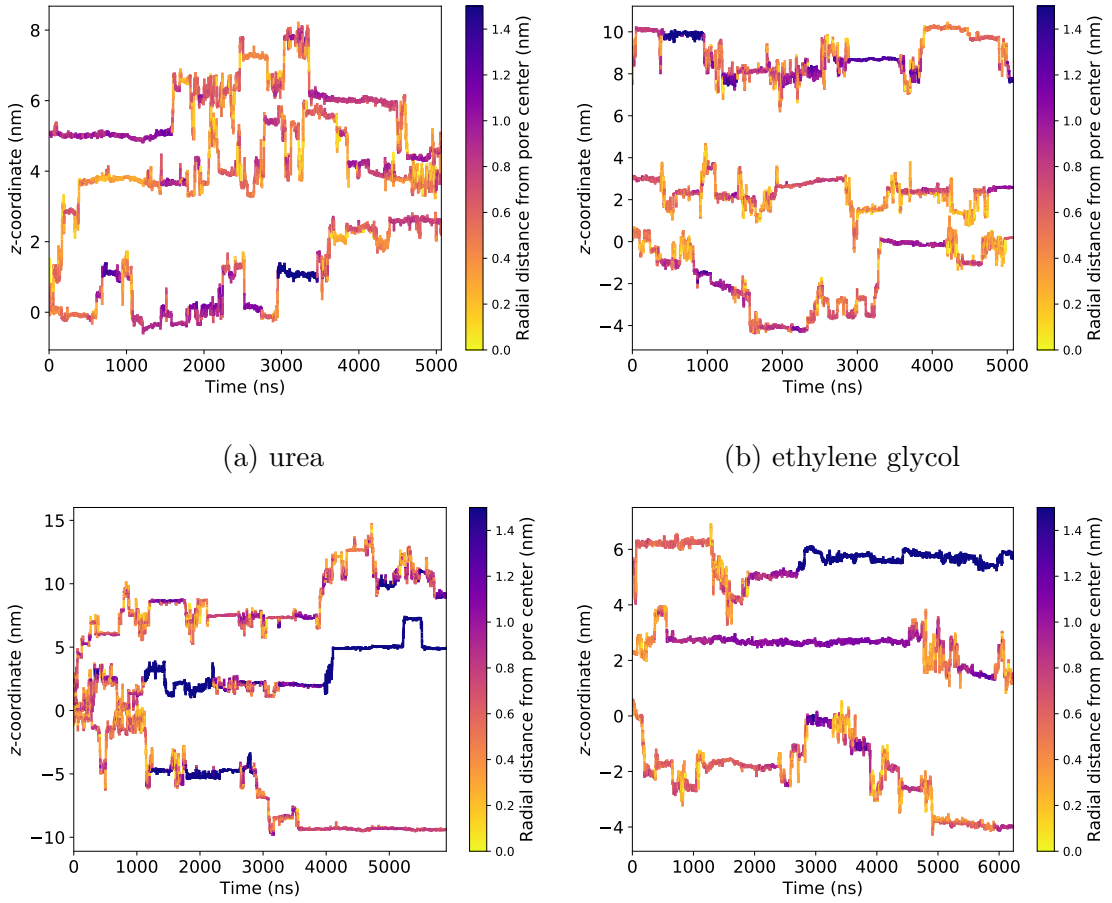
$$S_{ij} = \frac{J_i/\Delta C_i}{J_j/\Delta C_j} \tag{14}$$

III. RESULTS AND DISCUSSION

A. Anomalous Diffusion Modeling

1. Parameterizing Subordinated Fractional Brownian Motion

We find the data suggests that solute motion in this system can be well-modeled by subordinated fractional Brownian motion. In Figure **??**, we plotted representative trajectories for each solute. They are characterized by intermittent hops between periods of entrapment. The near-Gaussian distribution of jump lengths and power law distribution of dwell times are both characteristic of CTRWs (Figures **??** and **??**). The apparent anti-correlation between hops suggests a fractional diffusion process is subordinate to the CTRW. Of the different models, a process subordinated by FBM or FLM is best supported by the data because the analytical correlation structures of hop lengths are close to those observed in our simulations (Figure **??**). Fractional motion is common in crowded viscoelastic environments where motion is highly influenced by the movement of surrounding components, such as monomer tails in our case. [**?**]



(a) urea

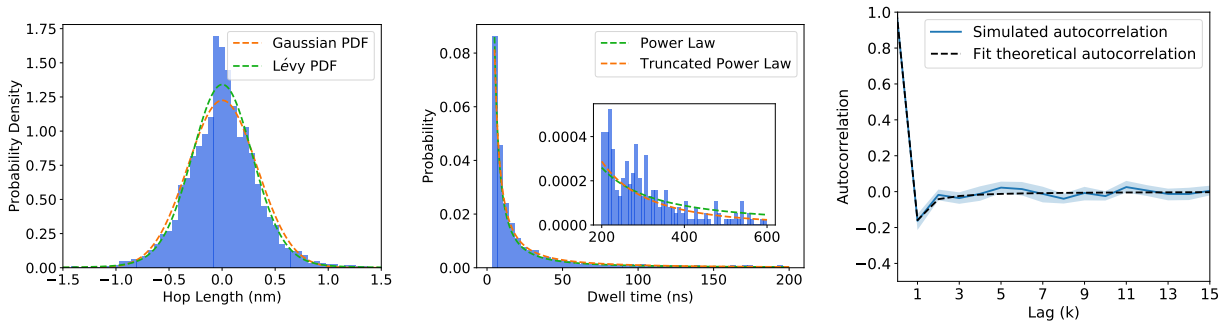
(b) ethylene glycol

(c) methanol

(d) acetic acid

FIG. 2: Three representative trajectories generated from each solute exhibit hops between periods of entrapment, characteristic of a CTRW. Solute dynamics show radial dependence, represented by the color at each time point. The longest periods of entrapment typically occur when solutes drift far from the pore center and into the tails.

We modeled the distributions of hop lengths in two ways. First, we assumed the distribution to be Gaussian since it is possible to exactly simulate realizations of fractional Brownian motion. Second, we fit the distributions to Lévy stable distributions since it is more general than the Gaussian distribution. We plotted the MLE fits of both on top of urea's hop length distribution in Figure **??**. The Lévy distribution does a better job capturing the somewhat heavy tails and high density near 0 of the hop length distribution. However, since there are no known exact simulation techniques for generating realizations of fractional Lévy motion, this more general fit may not be worthwhile. In fact, we typically observe very little



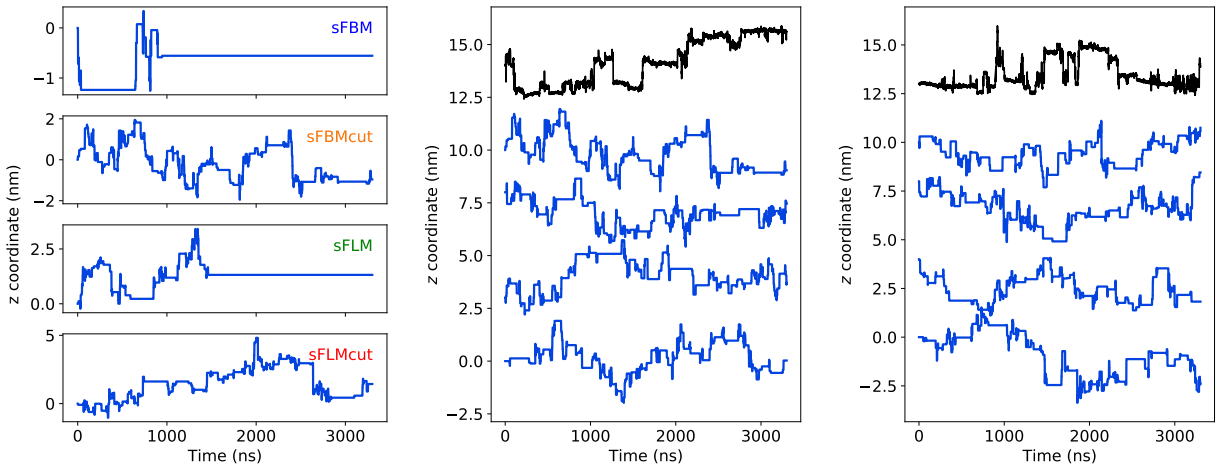
(a) (b) (c)

FIG. 3: Using urea as an example, (a) the distribution of hop lengths can be fit by a Gaussian in addition to a more general Lévy stable distribution, though the Lévy stable distribution does a better job of capturing the heavier tails and increased density near 0. We explore models fit to both distributions, as Gaussian hops are more convenient to model. (b) The distribution of dwell times is fit well by a power law but it over-estimates the probability density at long dwell times. A power law truncated with an exponential cut-off better describes the probability of long dwell times in our simulations. (c) The hops are negatively correlated to their previous hop. In combination, (a) – (c) support modeling solutes as either subordinated fractional Brownian or Lévy motion. All other solutes show similar distributions and autocorrelation functions (see Figure S8 of the Supplemental Material).

difference between model predictions parameterized with FLM versus FBM (see Figure **??**).

We also modeled the distribution of dwell times in two ways. First, we assumed pure power law behavior since it is consistent with most theoretical descriptions of CTRWs. The data fits well to this model at short dwell times but the density of long dwell times is over-estimated. In our second approach, we truncate the power law distribution with an exponential cut-off, lowering the probability of extremely long dwell times. We demonstrate this effect in the inset to Figure **??**, where the density of the truncated power law drops below that of the pure power law and tends towards 0 near a dwell time of 250 ns.

Several of the choices of AD models and parameters we have described yield qualitatively similar trajectories to those seen in our MD simulations. In Figure **??**, we plot representative sample trajectories for each combination of the dwell time and hop distribution, labeled according to Table **??**. The sFBMcut and sFLMcut in particular resemble the trajectories



(a) (b) sFBMcut (c) sFLMcut

FIG. 4: (a) Simulated urea trajectories generated by each of the four variations of the one mode AD model display qualitatively similar hopping and trapping behavior to that shown in Figure **??**. Dwell times are exaggerated in the sFBM and sFLM models because the power law dwell time distributions are not truncated and have infinite variance. In (b) and (c) we compare additional trajectories simulated with the sFBMcut and sFLMcut models (blue) to MD solute trajectories (black). Trajectories are vertically offset for visual clarity.

in Figure **??**. When we do not truncate the dwell time distribution, the trajectories tend to incorporate very long dwell times as shown by the sFBM and sFLM models. Predictions made with these models consistently under-predict the MD MSDs (see Figure S9 of the Supplemental Material). Therefore, we will not include the pure sFBM or sFLM models in the remainder of our analysis.

2. Stationarity of solute trajectories

We observe that in some cases, solute trajectories extracted from our MD simulations display non-stationary behavior. We defined the perceived equilibration time point for each solute based on the time at which the number of solutes inside the pores and tails stabilized (Figure S1). With this definition, we observe evidence of non-stationary solute behavior after the perceived equilibration point, on the μs timescale. In Table **??**, we compare the MSD of the solutes calculated using trajectory data from the first and second halves of the

"equilibrated" simulation time. Ethylene glycol and methanol show considerable differences between the MSDs of the two halves. Urea and acetic acid appear to demonstrate satisfactory stationarity. The shapes of urea and acetic acid's MSD curves are also very similar (see Figure S11 of the Supplemental Material).

Residue	MSD	
	First Half	Second Half
urea	0.31 (0.23, 0.40)	0.35 (0.28, 0.43)
ethylene glycol	3.11 (2.06, 4.25)	1.03 (0.76, 1.29)
methanol	2.24 (1.61, 2.92)	1.06 (0.77, 1.39)
acetic acid	1.58 (1.11, 2.03)	1.43 (1.10, 1.73)

TABLE III: In order to be considered stationary, the MSD of the ensemble of solute trajectories should be the same independent of the portion of equilibrated trajectory that is analyzed. The MSD values in this table are averages taken after a 500 ns time lag, calculated independently from the first and second halves of the equilibrated solute trajectories. Urea and acetic acid both appear to satisfy the stationarity criteria, while ethylene glycol and methanol show significantly smaller MSDs when calculated from the second half of the equilibrated trajectories.

Given stationary data, the AD approach is capable of predicting solute MSDs measured from MD simulations within or near statistical uncertainty. In Figure S12 of the Supplemental Material, we parameterized the AD models using the first half of the stationary solute equilibrated trajectories and then compared the model predictions to the MD MSD of the second half of the stationary trajectories. In most cases, the model predictions fall within the 1σ confidence intervals of the MD MSD.

This brief analysis suggests that we may be operating on the border of the minimum amount of data required to accurately parameterize AD approach models. Working with only half of the data we collected (~ 2 μs post-equilibration) may not always be sufficient for extracting reliable parameter estimates. Therefore, for the remainder of this work, we will employ parameters fit to the entire equilibrated portion of the solute trajectories. Even doubling the data might not be good enough for molecules with statistical non-stationarity, meaning the predictive and interpretive power of the time series modeling applied to these

trajectories will be lower.

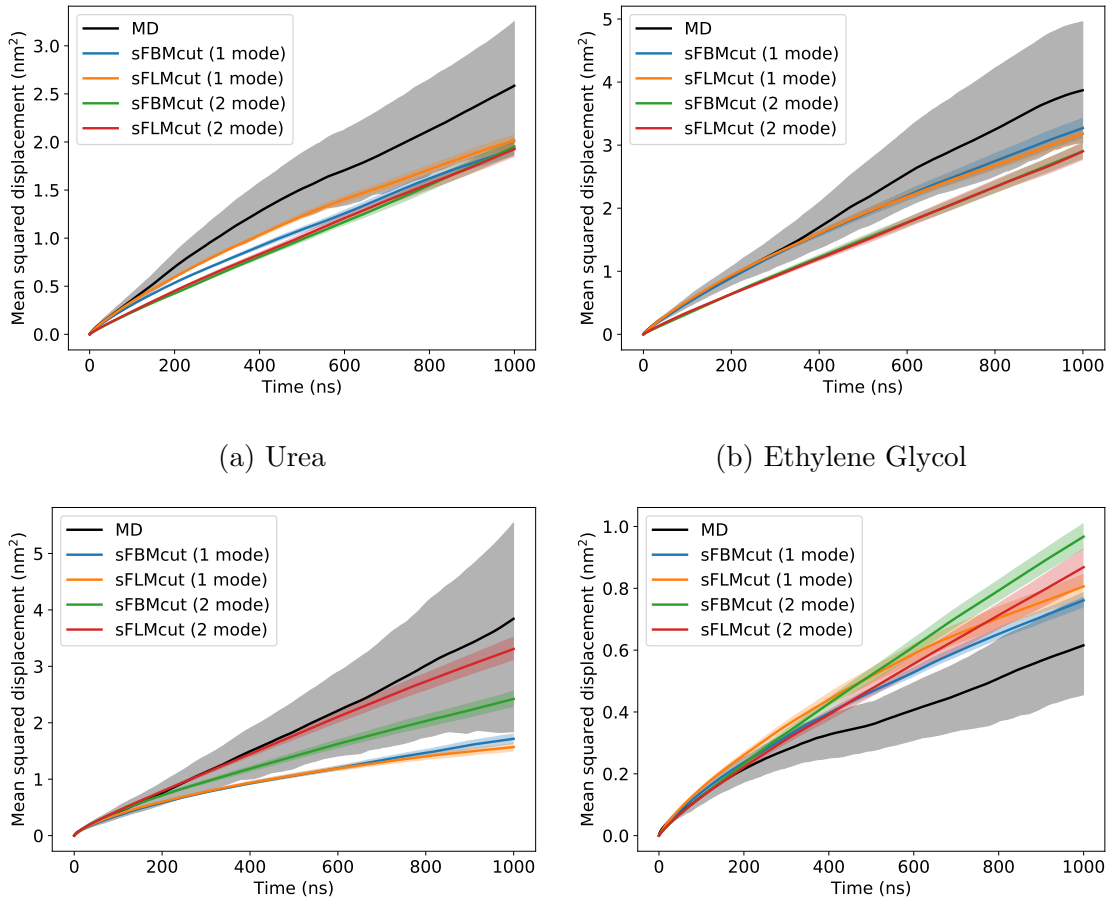
3. Model predictions

We obtain reasonable predictions of the MD simulated MSDs when we parameterize AD models with all available data after the perceived equilibration time. The MSD curves generated from both the one and two mode models are overlayed with the MD simulated MSDs for comparison in Figure **??**. The associated parameters for the one and two mode models are presented in Figures **??** and **??**.

The one and two mode AD models do a fairly good job of predicting the magnitude of the MD MSD curves examined up to a 1000 ns time lag. In most cases, both the Brownian (sFBMcut) and Lévy (sFLMcut) versions of both the one and two mode AD models give very similar results, most likely because their α_h parameter values are relatively close to 2, meaning the hop length distributions are nearly Gaussian even when fit to a more general Lévy distribution.

The deviation between the one mode MSD predictions and MD are primarily due to differences in their curvature at long time lags. The shape and magnitude of the predicted curves appears accurate relative to MD at short time lags. However, the modeled trajectories undershoot the mean MD MSD as the time lag increases. As discussed in the previous section, long time positional anti-correlation, on the order of hundreds of ns, may not exist in the MD system. Eventual loss of correlation would result in a shift from sub-linear to linear MSD behavior, as observed in the MD trajectories. Acetic acid exemplifies this point. At first glance, acetic acid's predicted MSD appears to match the curvature of MD quite well, but closer examination reveals that the MD MSD curve may actually shift from a sub-linear to a linear regime around 500 ns.

The two mode models display curvature more consistent with MD but for non-physical reasons. Every time a switch between the pores and tails occurs, the width of the distribution used to model hops changes. We are unaware of an appropriate technique which can correlate hops that come from different hop distributions, therefore every time a mode switch occurs, the correlation structure is broken. Solutes that switch between modes the least show predicted MSDs with the greatest curvature. Due to the much larger accessible volume that a smaller molecule has, methanol spends $> 90\%$ of its time in the hydrophobic tails (see



(a) Urea

(b) Ethylene Glycol

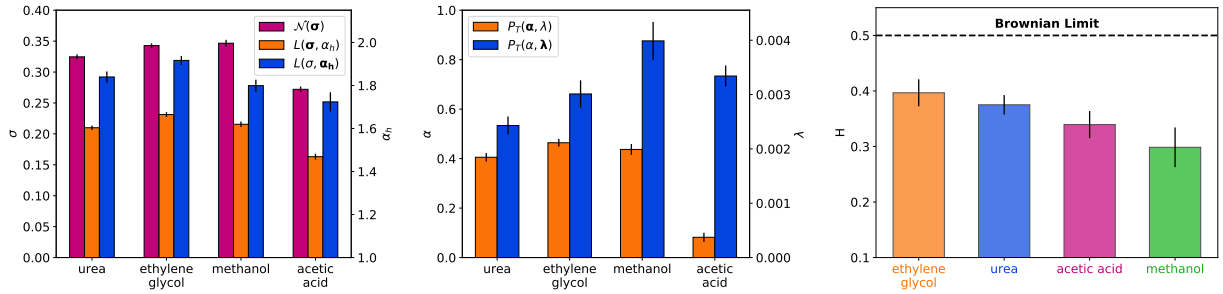
(c) Methanol

(d) Acetic Acid

FIG. 5: In most cases, MSDs generated from realizations of both the one and two mode AD models lie within or near the 1σ confidence intervals of MD-generated data. Drawing hops from a truncated Lévy stable distribution (sFLMcut) yields MSDs similar to when hops are drawn from Gaussian distributions (sFBMcut). In most cases, the one mode simulated MSDs under-predicted the mean at long timescales partially because they show pronounced curvature which the MD MSDs lack. The two mode predictions show less curvature than the one mode MSDs because the hop correlation structure is broken every time a transition between tails occurs.

Figure **??**), so mode transitions are relatively rare and the predicted MSDs have significant curvature. This artificially solves the problem of long timescale correlation, however we do not recommend this as it has no physical basis.

The model parameters for the one and two mode models tell stories about each solute's

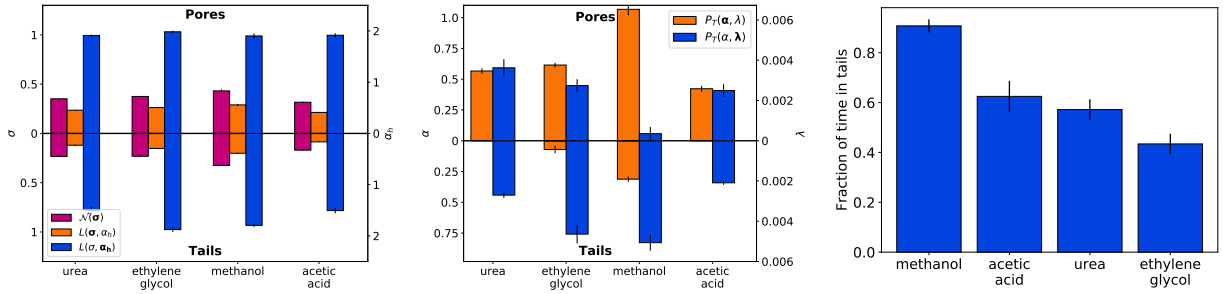


(a) (b) (c)

FIG. 6: The parameters of the one mode model reveal differences in dynamics between solutes. (a) We parameterized Gaussian, $\mathcal{N}(\sigma)$, and Lévy stable, $L(\sigma, \alpha_h)$, distributions to describe solute hop lengths. We assume the mean (μ) to be zero for these distributions and no skewness ($\beta = 0$) in the Lévy stable distributions. High values of σ and lower values of α_h result in larger hops. (b) We parameterized a pure power law, $P(\alpha)$, and a truncated power law, $P_T(\alpha, \lambda)$, distribution to describe solute dwell times. Lower values of α lead to heavier power law tails and higher values of λ truncate the distribution at lower dwell times. (c) Finally, we parameterized the hop autocorrelation function, $\gamma(H)$, to describe the degree of correlation between hops. Simulations with higher values of H display behavior closer to the Brownian limit.

behavior that help explain the difference between the MSDs of different solutes. Higher values of σ and lower values of α_h indicate larger average hop length magnitudes by increasing the hop length distribution's width and tail density respectively. Higher values of α indicate a lower probability of long dwell times. Higher values of λ truncate the power law distribution earlier preventing extremely long dwell times. Values of H closer to the Brownian limit of 0.5 indicate a lower degree of negative correlation between hops. All of these changes in physical behavior contribute to an overall increase in the predicted MSD.

Examining first the parameters of the one mode model, we can begin to break down the trends in solute MSDs. The parameters belonging to ethylene glycol and methanol are relatively similar, which is consistent with their similar MD MSDs and qualitatively similar MD timeseries. Relative to ethylene glycol, methanol tends to stay trapped for less time and takes larger hops but the most substantial difference is with respect to their Hurst parameters. Methanol has the lowest H of all the solutes studied because it spends the



(a) (b) (c)

FIG. 7: The two mode model parameterizes solute behavior in the pore and tails separately. We consider solutes to be within the pore region if they are 0.75 nm from a given pore center, otherwise they are in the tails. (a) Generally, movement is much more restricted in the tail region, parameterized by lower σ values (smaller hops) for the Gaussian and Lévy stable distributions. Values of α_h are significantly lower for urea and acetic acid meaning there is a larger probability that they will take large hops. (b) Dwell times are longer in the tails. Lower values of α correspond to power laws with heavier tails and thus higher probabilities of long dwell times. There is no easily discernible trend in λ of the truncated power law distribution. Note that we used the same Hurst parameter for both modes (shown in Figure **??**) due to a low number of sufficiently long sequences of hops in each mode. (c) Solutes spend various amounts of time in the tail and pore region dependent on their size, shape and chemical functionality. Methanol's small size favors occupation of the much larger accessible volume in the tails. Urea and acetic acid are fairly stable in both regions since they are small and hydrophilic. Ethylene glycol has a slight preference for the pores likely because it is a larger molecule with two hydrophilic hydroxyl groups.

majority of its time outside the pore region where collisions with tails are frequent. Urea has the third highest MSD which is primarily a consequence of more frequent and longer dwell times (lower α and λ). Urea's hop lengths (σ) and correlation (H) are comparable to ethylene glycol and methanol. Acetic acid has the smallest MSD among the solutes studied due to longer periods of entrapment and shorter hops. Its trapping behavior is parameterized by an α value significantly lower than other solutes, but an intermediate λ value, suggesting it experiences many medium-length periods of entrapment. Its hops are smaller but are

slightly compensated by a heavier tailed distribution (lower α_h) than the other solutes.

We can use the two mode model to gain an even deeper understanding of solute behavior in the pore versus in the tails. It is clear that solutes are significantly slowed while they are in the tail region where long dwell times are more probable (smaller α) and hops are smaller (smaller σ). Each solute spends a different amount of time in the tails (see Figure **??**). Urea and acetic acid spend slightly more than half of their time in the tails (56% and 62% of their time respectively) while ethylene glycol spends about 44% of its time in the tails. Urea and acetic acid's compact, flat structure allows it to more easily partition into the tails while ethylene glycol prefers the pore region due to its two hydrophilic hydroxyl groups. In contrast, methanol spends 91% of its time in the tails, likely due to its small size. The value of α_h for urea and acetic acid in the tails is 1.50, meaning its hop distribution is heavy tailed relative to ethylene glycol and methanol, whose α_h values are 1.90 and 1.85 respectively, which is more consistent with a Gaussian distribution (α=2). Acetic acid and urea are structurally similar molecules, both planar with two heavy atoms attached to a carbonyl group. Their small size and rigid shape may allow them to occasionally slip through gaps in the tails. Meanwhile, methanol is small enough that it does not need to make larger jumps to escape traps.

Overall, the AD approach does a reasonable job of predicting solute MSDs and its parameters can help us further understand solute dynamics. Hops appear to be well modeled as anti-correlated draws from either Gaussian or Lévy stable distributions. The data strongly suggests that one must truncate the power law dwell time distributions in order to obtain accurate MSD estimates. Trajectories generated by pure power laws are qualitatively non-physical. We can further understand solute dynamics by adding radially dependent parameter distributions as in the 2 mode model. A significant amount of solute trajectory data is necessary in order to achieve good parameter estimates.

B. The Markov State-Dependent Dynamical Model

The AD model is useful if one does not know exact transport mechanisms in a system since it only requires time series data. However, since we have already studied transport mechanisms in detail in our previous work, we can attempt to model transport as transitions between known discrete states, defined in Table **??**, with state-dependent positional

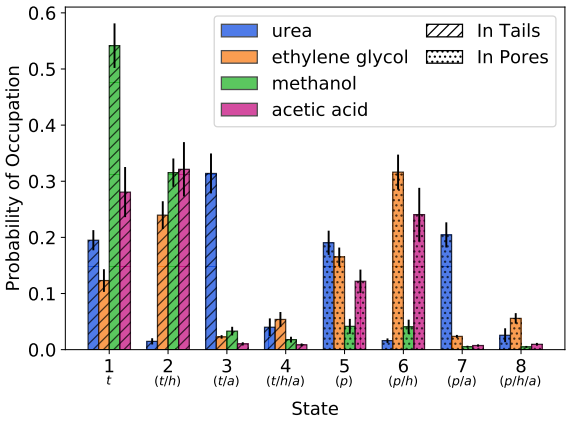


FIG. 8: Solutes spend varying amounts of time under the influence of each trapping mechanism. To aid the reader, we labeled each state with an abbreviation which identifies the combination of conditions to which each solute is subject in each state: t - tails, p - pores, h - hydrogen bonded, a - associated with sodium. Solutes tend to favor the same types of interactions (e.g. hydrogen bonding and/or associating with sodium ions) independent of whether they are in the pores or the tails.

fluctuations, which we refer to as the Markov state-dependent dynamical model (MSDDM).

1. Solute state preferences

Solute size and chemical functionality influence which states are visited most frequently. In Figure **??**, we plotted the probabilities of occupying a given state at any time. Solutes tend to favor the same types of interactions independent of which region they are in. We can relate these interactions to solute chemical functionality and use that intuition to hypothesize new designs for LLC monomers which control the transport rate of specific solutes.

Urea spends the largest fraction of its time trapped via association with sodium ions. It does so 31% of the total time while in the tails (state 3) and 21% of the total time while in the pores (state 7). Note that sodium does not drift significantly into the tails but sits close to the pore/tail region boundary. The electron-dense and unshielded oxygen atom of urea's carbonyl group is prone to associate with positively charged sodium ions. The nitrogen atoms of urea are only weak hydrogen bond donors. Therefore, the transport rate of urea is likely to be most significantly modified by removing or changing the identity of the counter-ion.

Ethylene glycol spends the largest fraction of its time trapped in a hydrogen bonded state. It does so 24% of the total time while in the tails (state 2) and 32% of the total time while in the pores (state 6). The two hydroxyl groups of ethylene glycol readily donate their hydrogen atoms to the carboxylate head groups and the ether linkages between the head groups and monomer tails. The transport rate of ethylene glycol might be modified by removing hydrogen bond accepting groups from the LLC monomers, especially those which stabilize the solute in the tails (i.e. the ether linkages).

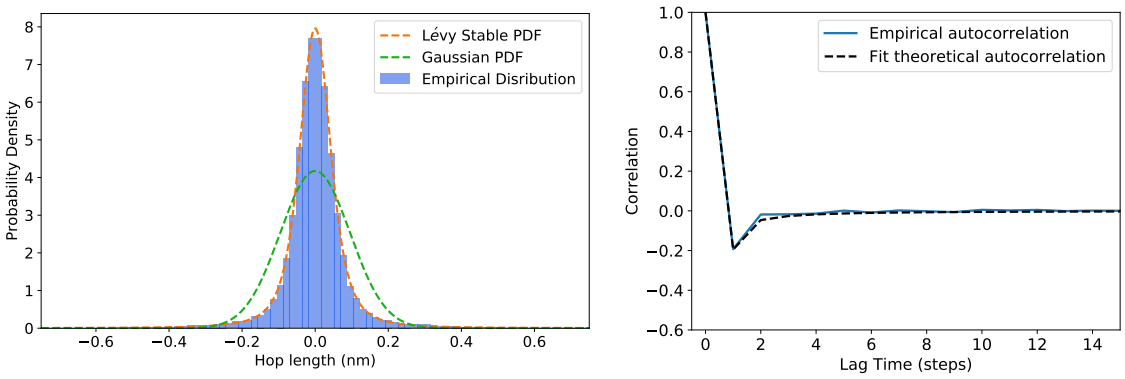
Methanol spends most of its time unbound in the tail region (state 1) and spends a significant portion of time hydrogen bonded while in the tail region. Tail region hydrogen bonds are donated from methanol to the ether linkages between the monomer head groups and tails. One might modify the rate of methanol transport by controlling its partition into the monomer tails. This might be achieved by adding cross-linkable groups near the monomer head groups.

Finally, acetic acid spends the majority of its time hydrogen bonding both in and out of the pore (states 2 and 6). Although it has an unshielded carbonyl group in its structure, association with sodium ions in this environment is apparently a much weaker interaction than hydrogen bonding. As with ethylene glycol, one might modify the transport rate of acetic acid by removing hydrogen bond accepting constituents of the LLC monomer. With this modification, we hypothesize that acetic acid might show similar transport rates to urea given their structural similarity.

2. Parameters of the MSDDM

To create an MSDDM for each solute, we determined the state sequence associated with each solute trajectory based on the geometric indicators of the states indicated in Table **??**. We then generated emission distributions of fluctuations within each state as well as transitions between states. In theory, one could parameterize separate transition distributions for those which occur in the tails versus in the pores, however this would lead to a broken correlation structure similar to that seen in the two mode AD models.

We observe correlated emissions drawn from Lévy stable distributions. The deviation of the emission distributions from Gaussian behavior is far more pronounced than that seen in the hop length distributions of the previous section. We therefore did not consider the



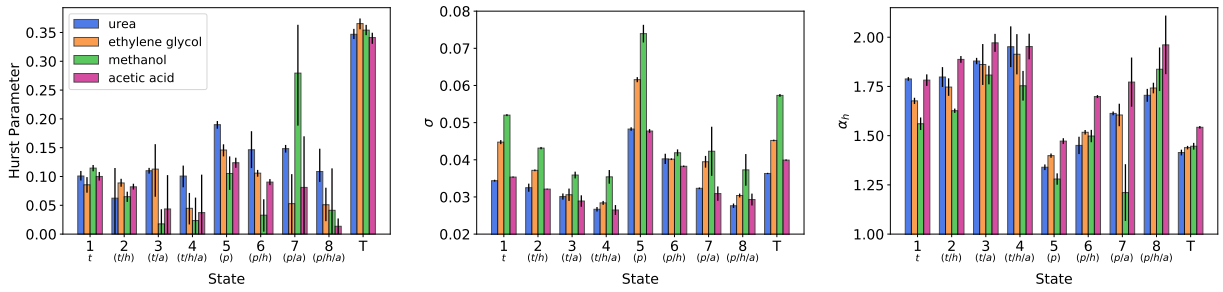
(a) (b)

FIG. 9: (a) The emission distributions of hop lengths are non-Gaussian and heavy-tailed. Shown here is the emission distribution for transitions between states. The maximum likelihood Gaussian fit severely underestimates the empirical density of hops near and far from zero while overestimating the density of hops at intermediate values. (b) Jumps drawn from the transition distribution are negatively correlated to each other. The normalized version of Equation **??** fits well to the data suggesting FLM is an appropriate way to model jumps.

Gaussian case (see Figure **??**). The correlation structure between hops is consistent with that of FLM (Figure **??**). The parameters of the Lévy stable distributions along with their Hurst parameters are visualized in Figure **??** (and tabulated in the Supplemental Material, Table S4).

Most of a solute's MSD is a consequence of transitions between the 8 states in Table **??**. Perfectly anti-correlated motion (H=0) results in no contribution to the solute's MSD. Motion while trapped in a state is highly anti-correlated as indicated by their consistently low Hurst parameters. There is a weak negative trend in the Hurst parameter values as the number of simultaneously influencing trapping mechanisms increases (Figure **??**). The Hurst parameters for transitional (T) emissions are up to 18 times higher than emissions from trapped states. The value of α_h for transition emissions is also relatively low giving higher probabilities to larger hops.

As solutes are influenced by more trapping interactions simultaneously (e.g. hydrogen bonding *and* association with sodium versus just hydrogen bonding), the width of the hop length distribution, σ, decreases while its Lévy index, α_h, increases. Treating states in the



(a) (b) (c)

FIG. 10: The parameters of the MSDDM are strong functions of trapping mechanisms. We observe different parameters but with similar trends between the tail and pore region. The states are defined in Table **??**. See Figure **??** for a description of the abbreviation under each state number. The legend in (a) applies to all subplots. (a) Motion is highly anti-correlated in trapped states. As the number of simultaneously influencing trapping mechanisms increases, the Hurst parameter, H, decreases. H is highest (closest to Brownian) during transitions between states (state T). (b) As more trapping mechanisms simultaneously influence solutes, the width of the hop length distribution (σ) decreases. The largest hops occur when solutes are unbound in the pores. (c) The weight of the hop length distribution's tails, parameterized by α_h, increases as more trapping mechanisms influence solutes simultaneously. The transitional hop distributions have among the heaviest tails.

tail and pore regions independently, σ is largest and α_h is smallest when solutes are not hydrogen bonding or associating with sodium (states 1 and 5). Solutes are free to move and take occasionally large hops. The smallest σ and highest α_h values are measured when solutes are hydrogen bonding and associating with sodium at the same time (states 4 and 8). Motion is restricted by multiple stabilizing forces which maintains a relatively narrow distribution of hop lengths.

3. Application of the MSDDM

Quantitatively, the dynamics of urea, ethylene glycol and, to a lesser extent, methanol appear to be well-captured by the MSDDM. We simulated 1000 MSDDM trajectory realizations for each solute, as described in Section **??**, then calculated their MSDs (see Figure **??**).

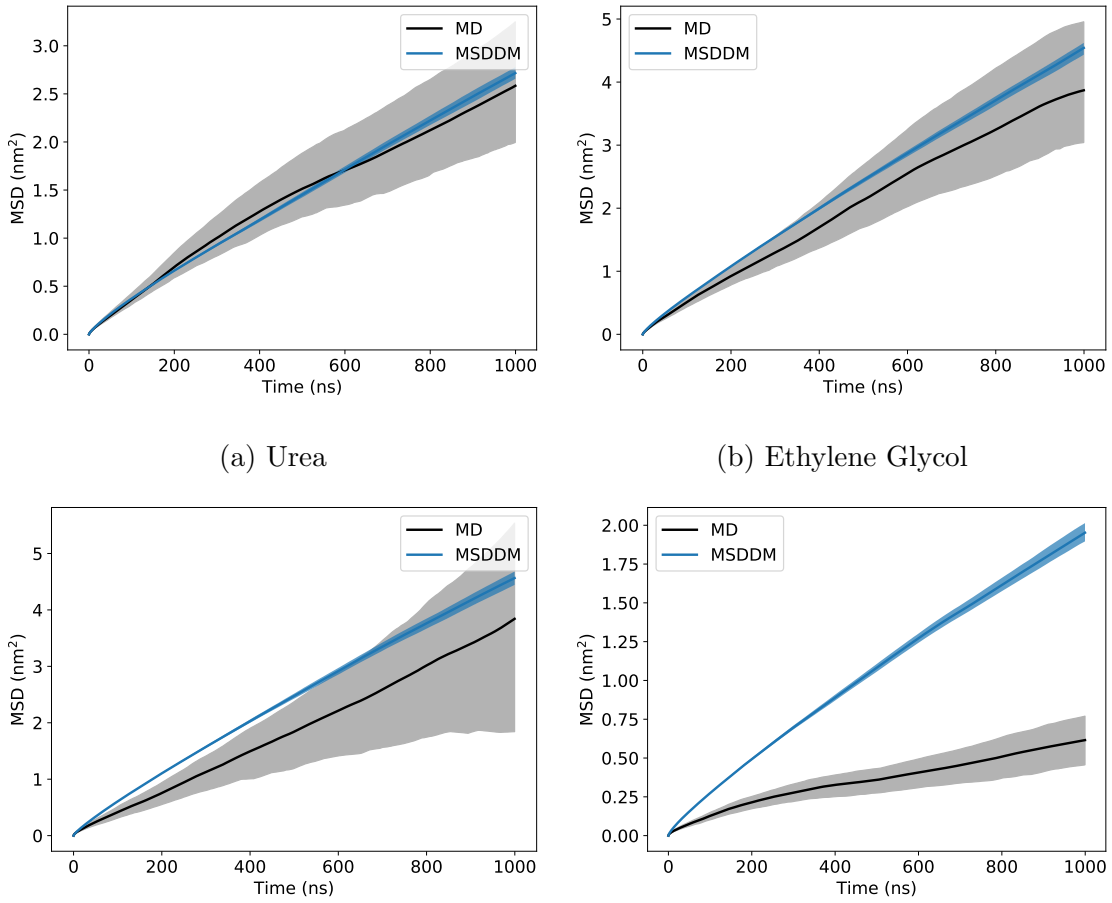
In most cases, the MSDDM predicts the magnitude of the MD MSDs within their 1σ confidence intervals.

The predicted MSD of acetic acid is severely over-estimated primarily due to under-estimation of hop anti-correlation (H too high) and an over-estimate of its hop lengths (σ and α_h too high). Acetic acid's predicted MSD actually lies just within the lower bound of urea's MD MSD 1σ confidence interval, but lower than the prediction of the MSDDM for urea. As stated earlier, most of a solute's motion modeled by the MSDDM is due to displacements during state transitions. Acetic acid has a transitional Hurst parameter close to urea's and transitional σ and α_h values that are higher than urea's. The primary reason that the MSDDM predictions for acetic acid's MSD are lower than urea's appears to be because of higher hop anti-correlation (lower H) in all states, including transitions. There is also the strong possibility that the over-estimate is a consequence of lumping together all of acetic acid's transitional hops into a single correlated distribution.

Despite relatively good predictions of the MD MSD, qualitative mismatch between simulated MSDDM and MD trajectories suggest that the MSDDM may be getting the right answers for the wrong reasons. We plotted typical realizations of the MSDDM for each solute and compared them to MD in Figure **??**. There is little evidence of trapping behavior or large hops. There are two reasons for this behavior. First, the width of hop length distributions are much smaller than those of the AD model. Closer examination of the characteristic MD trajectories shown in Figure **??** reveal that hops tend to be an accumulation of a series of hops in the same direction. All of the hops in the MSDDM are negatively correlated which prevents this from happening. The second reason is a consequence of using a single hop length distribution for transitions. This was necessary because we could not collect enough data to fit all of the possible transition distributions and because we could not correlate emissions coming from different hop distributions. Many transitions occur between two trapped states where the transitional hops are actually very small. Our model ignores this physical restriction which can cause the solute to drift rather than stay trapped.

C. Solute Flux and Selectivity

We used the one mode sFBMcut (see Table **??**) AD model in order to demonstrate how one can use its realizations in order to calculate the flux (see section **??**) of solutes given



(a) Urea

(b) Ethylene Glycol

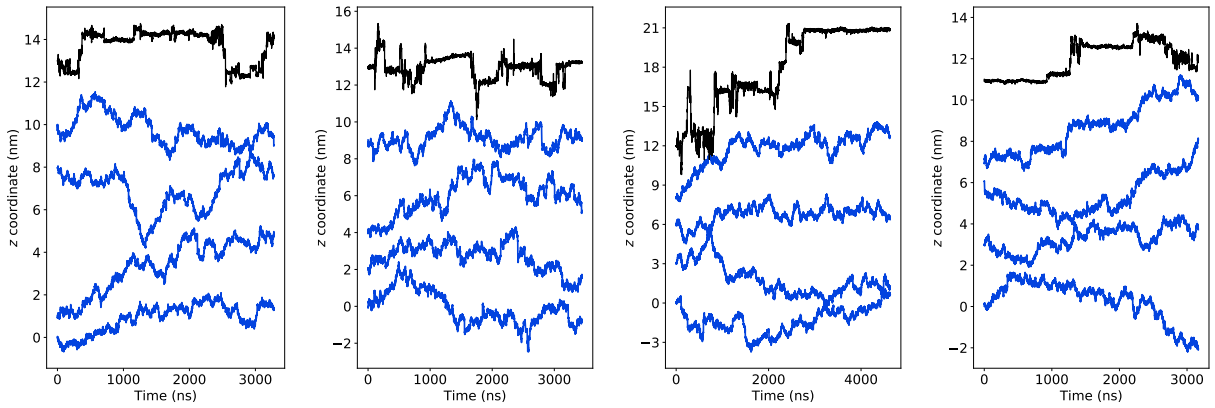
(c) Methanol

(d) Acetic Acid

FIG. 11: In most cases, the magnitude of the MSD curves predicted by the MSDDM agree well with those generated from MD simulations. The predicted MSD curves of urea and ethylene glycol lie within the 1σ confidence intervals of MD for all time lags. Methanol over-predicts the MSD at small time lags and acetic acid grossly over-predicts the MSD at all time lags. Like the AD approach models, the MSDDM doesn't fully capture the curvature of the MD MSD curves.

model parameters extracted from MD simulations. The one mode sFBMcut model generates predictions similar to the one mode sFLMcut model at a lower computational cost. We do not consider the two mode AD model because it has a broken correlation structure and we do not consider the MSDDM because its realizations do not display the expected hopping and trapping behavior.

It is computationally infeasible to simulate trajectories long enough that they traverse the

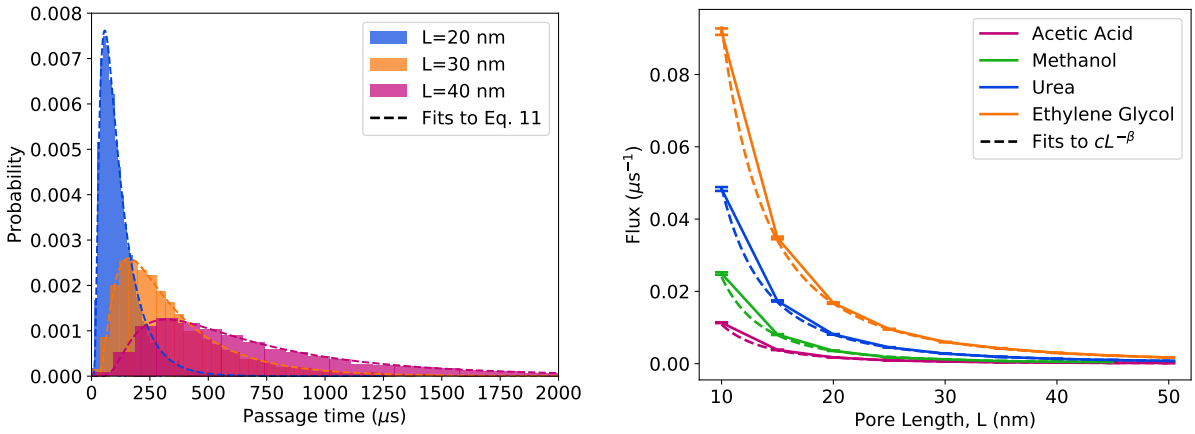


(a) Urea (b) Ethylene glycol (c) Methanol (d) Acetic Acid

FIG. 12: Realizations of the MSDDM for each solute (blue) do not reproduce the hopping and trapping behavior observed in our MD simulations (black). The trajectories are qualitatively similar to what one might expect for Brownian motion even though the MSDs are often similar to the atomistic systems.

length of a macroscopic pore. To date, the thinnest H_{II} LLC membrane synthesized with the monomer in this work was 7μm thick. [**?**] Using 24 cores to simulate trajectory realizations in parallel, it takes on the order of 1 day to simulate 10000 sFBMcut realizations of solutes traversing a 50 nm pore. The RAM requirements and performance scales greater than linearly and thus would take an infeasible amount of memory and time to simulate transport through a pore over 100 times longer. One could improve performance significantly by simulating less trajectories. In Figure S14 of the Supplemental Material, we determined that one can simulate as few as 100 sFBMcut realizations in order to parameterize Equation **??**. For better precision, we recommend simulating at least 1000 realizations. However, even with an order of magnitude decrease in number of trajectories, it is still infeasible to simulate experimental-length pores.

We used simulated trajectories which traverse computationally-reasonable length pores in order to construct an empirical model which one can use to estimate particle flux for arbitrary length pores. We fit Equation **??** to the empirical distribution of first passage times in Figure **??** and used the expected value of the analytical equation to calculate flux from Equation **??**. As shown in Figure **??**, the flux appears to scale according to a power



(a) (b)

FIG. 13: (a) The distributions of first passage times generated from the sFBMcut model fit well to Equation **??**. We show similar fits for the remaining solutes in Figure S13 of the Supplemental Material. (b) The single particle flux measured by the sFBMcut AD model decays with increasing pore length. The rankings of solute fluxes are consistent with the MSDs predicted by each model. We fit the single particle solute flux versus pore length, L, to a power law function of the form $AL^{-\beta}$ (dashed lines in (b) and (c)).

law of the form:

$$J(L) = AL^{-\beta} \tag{15}$$

The scaling of solute flux with pore length is primarily influenced by anti-correlation between solute hops. In Figure **??**a, we show that β is inversely related to the Hurst parameter. This makes intuitive sense since higher degrees of anti-correlation should slow the rate at which solutes cross the membrane pore. When we remove anti-correlation between hops (set H=0.5), the length dependence becomes the same for all solutes, dropping to a value just below 2. This also implies that hop lengths and dwell times do not affect length dependence since each solute exhibits different hopping and trapping behavior yet are all parameterized by the same value of β when H=0.5. We further verified this claim by removing hop anti-correlation and setting all dwell times equal to one timestep, effectively simulating Brownian motion. The length dependence remains unchanged.

Dwell times and hop lengths directly modify the rate at which solutes move through the membrane pores and are reflected in the scaling pre-factor of the solute flux curves, A. Comparison of Figure **??** with Figure **??** reveals that the ranking of the A parameters is

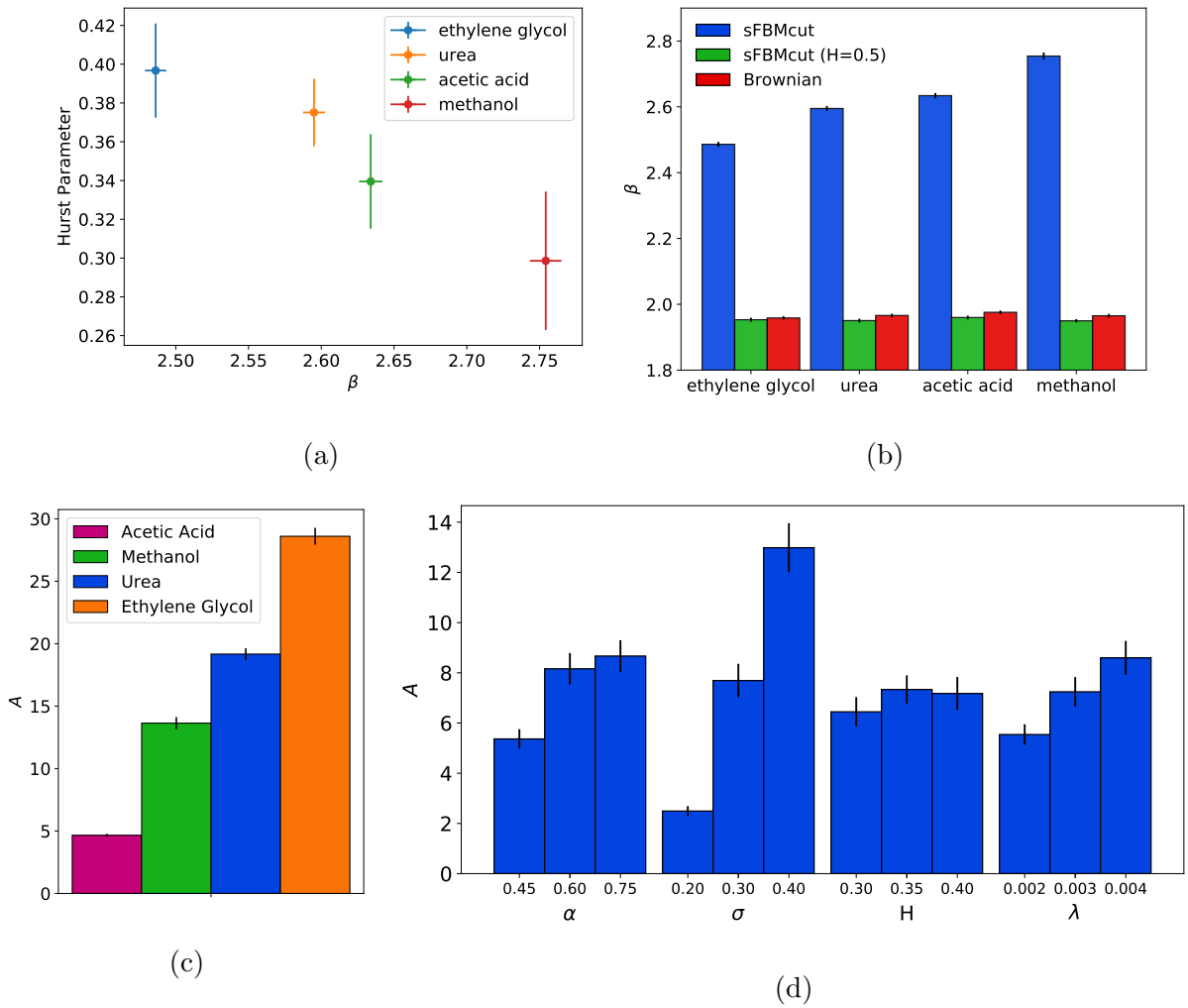


FIG. 14: (a) Increased anti-correlation between hops (decreased H) appears to increase the length-dependence of solute flux (higher β) for the sFBMcut model. (b) If we remove anti-correlation, by setting H=0.5, the length dependence parameter drops to a value near 2. When we remove hop anti-correlation and dwell times between hops, effectively simulating Brownian motion, the β parameter stays near 2. This suggests that β does not depend on dwell times (parameterized by α and λ). (c) As solute flux increases, the scaling of the flux curves, A, increases (compare ranking with Figure **??**). (d) Physical processes which increase the rate of solute displacement result in larger values of A. To test the dependence of A on α, σ, H and λ, we chose a single set of parameters, representative of solutes parameterized by the sFBMcut AD model, and generated realizations of the model by varying each parameter independently about the same base parameter set. Decreased dwell times (increased α), increased hop lengths (increased σ), and a lower cut-off to the dwell time distribution (increased λ) lead to increases in A. and non-linearly to α. The data suggests that the A parameters do not depend on hop anti-correlation (H).

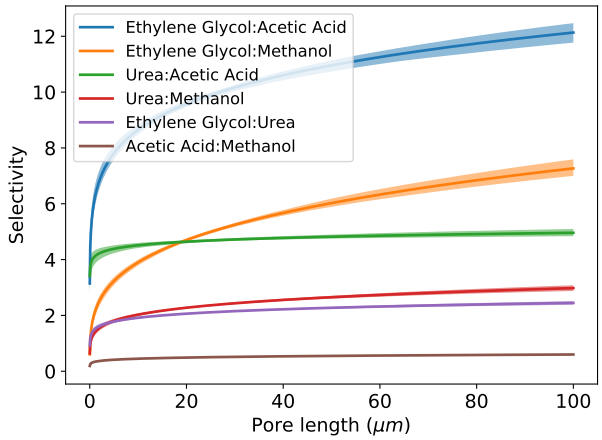


FIG. 15: The selectivity between pairs of species changes monotonically with pore length. The strength of dependence on pore length depends on the difference between β values. The largest differences in solute flux result in high selectivities at any pore length. This membrane may be a good candidate for the separation of ethylene glycol from acetic acid. Ethylene glycol has the lowest β value while acetic acid has the second highest, leading to strong length dependence. Ethylene glycol also has the highest flux and acetic acid has the lowest resulting in relatively high selectivities independent of pore length.

consistent with the ranking of solute flux. In Figure **??** we demonstrate that decreasing dwell times (increasing α), cutting off the dwell time distribution at shorter times (increasing λ), and increasing hop lengths (increasing σ) independently lead to an increase in A. Figure **??** also suggests that A is not dependent on hop anti-correlation (H).

The power law decay of the flux with pore length implies the following relationship for selectivity via substitution of Equation **??** into Equation **??**:

$$S_{ij}(L) = \left(\frac{A_i}{A_j}\right) L^{(\beta_i - \beta_j)} \tag{16}$$

In Figure **??**, we plot Equation **??** for pore lengths ranging from those studied in Figure **??** to macroscopic-length pores. For the same degree of hop anti-correlation, as is the case for uncorrelated motion, selectivity depends on solute hop lengths and dwell times (A). For $(A_i/A_j) = 1$, LLC membranes will be more selective towards passage of solutes with less anti-correlated hopping behavior (lower β). The selectivity towards solutes with less length dependent flux increases with membrane thickness. In most cases, the length dependence of selectivity plateaus near or within the range of experimentally-accessible membrane thick-

nesses. Therefore, from a practical standpoint, one should not expect significant changes in selectivity by varying LLC membrane thickness.

Of the solutes studied, the data suggests that this particular LLC membrane might be most useful for selectively separating ethylene glycol from methanol and acetic acid. Relative to acetic acid, ethylene glycol takes larger hops and is trapped longer, leading to a larger A. There is also an appreciable difference in the scaling of each solute's flux with pore length (β) which gives the selectivity significant length dependence. Relative to methanol, ethylene glycol takes similarly-sized hops and dwells which is why selectivity is relatively low for very short pore lengths. However, the large difference in hop correlation between the two solutes leads to the strongest pore length dependence of selectivity.

The insights into flux and selectivity provided by our time series modeling approach could not be drawn easily by simply observing solute motion, the structure of the membrane nanopores, or even the solute trajectories extracted from the MD simulations. Differences in solute MSDs alone do not fully explain the trends and magnitudes of the selectivities shown in Figure **??**. Complex interplay between membrane constituents and solutes with varying chemical functionality lead to diverse solute behavior. Even if our model is not perfect, it provides clear logic behind the mechanisms leading to selective behavior which could significantly help illuminate any design choices. We hope this type of analysis can be leveraged to explore new, interesting and complex separations problems.

IV. CONCLUSIONS

We have tested two different mathematical frameworks for describing solute motion by applying them to an H_{II} phase LLC membrane. The values obtained for the parameters when fitting the models to the time series data offer important mechanistic insight on the molecular details of transport. Subordinated fractional Brownian and Lévy motion have a strong theoretical foundation in the anomalous diffusion literature. Our single mode AD model quantifies and allows comparison of the hopping and trapping behavior among solutes. A two mode model that describes dynamics based on whether a solute is in or out of the pore region allows us to break down individual solute motion into the two distinct regimes and we showed that solute motion is clearly restricted while in the tail region. Our Markov state-dependent dynamical model uses explicitly defined trapping mechanisms and gives a

nice description of transitions between these observed states, the equilibrium distribution of solutes among states as well as the type of stochastic behavior shown in each state.

Although large portions of the MSDs predicted by our models fall close to or within the 1σ confidence intervals of MD, it is not always for physically accurate reasons. Qualitatively, MSDDM trajectories do not display the same hopping and trapping behavior shown by MD trajectories. Frequent transitions drawn from a single, relatively broad transition emission distribution prevent long periods of immobility. The most obvious solution to this would be to generate individual emission distributions for each type of transition, but this would require orders of magnitude more data or more prior assumptions about the distributions. This approach is further complicated by the need to make correlated draws from a fractional Lévy process with a frequently changing distribution width. A possible simplification could be to assume that correlation is lost every time a state transition occurs.

Although the AD approach generates qualitatively accurate trajectories and predicts MSDs near or within the 1σ confidence interval of our MD simulations, the curvature of the predicted MSDs does not appear to be consistent with the MD simulations. The MSD curves calculated from MD simulations appear to straighten out on long timescales while those predicted AD models continuously curve. This is because pure fractional Brownian motion features hop correlation that persists indefinitely. We may be able to address this by truncating the positional autocorrelation function, allowing correlation to diminish on the 100 ns timescale as suggested by the physical trajectories.

We demonstrated how one could use the one mode AD model in order to to determine macroscopic flux and selectivity. We showed that, when using the AD model, solute flux decreases with pore length at a rate faster than pure Brownian motion due to anti-correlation between hops. We used the ratio of solute fluxes generated from each model in order to calculate selectivity. Due to differences in hop anti-correlation, we observe length dependent selectivity. Based on these calculations we can hypothesize that this particular LLC membrane may be a good candidate for the selective separation of ethylene glycol from acetic acid or methanol.

Despite their shortcomings, our mathematical models help us think about how to design LLC monomers for solute-specific separations by forcing us to think in terms of controlling solute transport. The anomalous diffusion modeling approach is quite flexible because it does not require knowledge of specific trapping mechanisms. Screening a set of solutes and

applying the anomalous diffusion approach can help uncover trapping mechanisms by forcing the scientist to identify features common to solutes with long or short hop lengths and dwell times. The MSDDM is a powerful way to characterize explicit trapping mechanisms. It clearly partitions a trajectory into discrete mechanisms and quantifies their relative dominance in the real system. For example, it is clear that sodium ion association is primarily responsible for trapping urea while hydrogen bonding dominates ethylene glycol. If one redesigns an LLC monomer to eliminate one source of trapping, then higher selectivities may result. The modeling approaches presented in this paper facilitate new ways of approaching complex separation problems.

SUPPLEMENTAL MATERIAL

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

ACKNOWLEDGMENTS

We thank Richard Noble for helping us make the connection between the empirical mean first passage time distribution and the analytical equation to which we fit the distributions (Equation **??**).

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.