

# Supporting Information: Time Series Modeling of Solute Transport in an H<sub>II</sub> Phase Lyotropic Liquid Crystal Membrane

Benjamin J. Coscia

Michael R. Shirts

December 12, 2019

## S1 Setup and analysis scripts

All python and bash scripts used to set up systems and conduct post-simulation trajectory analysis are available online at [https://github.com/shirtsgroup/LLC\\_Membranes](https://github.com/shirtsgroup/LLC_Membranes). Documentation for the `LLC_Membranes` repository is available at <https://llc-membranes.readthedocs.io/en/latest/>. Table S1 provides more detail about specific scripts used for each type of analysis performed in the main text.

Script Name	Section	Description
<code>/setup/parameterize.py</code>	2.1	Parameterize liquid crystal monomers and solutes with GAFF
<code>/setup/build.py</code>	2.1	Build simulation unit cell
<code>/setup/place_solutes_pores.py</code>	2.1	Place equispaced solutes in the pore centers of a unit cell
<code>/setup/equil.py</code>	2.1	Equilibrate unit cell and run production simulation
<code>/analysis/solute_partitioning.py</code>	2.1	Determine time evolution of partition of solutes between pores and tails
<code>/timeseries/msd.py</code>	2.2	Calculate the mean squared displacement of solutes
<code>/analysis/sfbm_parameters.py</code>	2.2	Get subordinated fractional Brownian motion parameters by fitting to a solute's dwell and hop length distributions and positional autocorrelation function.
<code>/timeseries/ctrwsim.py</code>	2.2	Generate realizations of a continuous time random walk with the user's choice of dwell and hop distributions
<code>/timeseries/forecast_ctrw.py</code>	2.2	Combines classes from <code>sfbm_parameters.py</code> and <code>ctrwsim.py</code> to parameterize and predict MSD in one shot.
<code>/analysis/markov_state_dependent_dynamics.py</code>	2.3	Identify frame-by-frame state of each solute, construct a transition matrix and simulate realizations of the MSDDM model.

Table S1: The first column provides the names of the python scripts available in the `LLC_Membranes` GitHub repository that were used for system setup and post-simulation trajectory analysis. Paths preceding script names are relative to the `LLC_Membranes/LLC_Membranes` directory. The second columns lists the section in the main text where the output or usage of the script is first described. The third column gives a brief description of the purpose of each script.

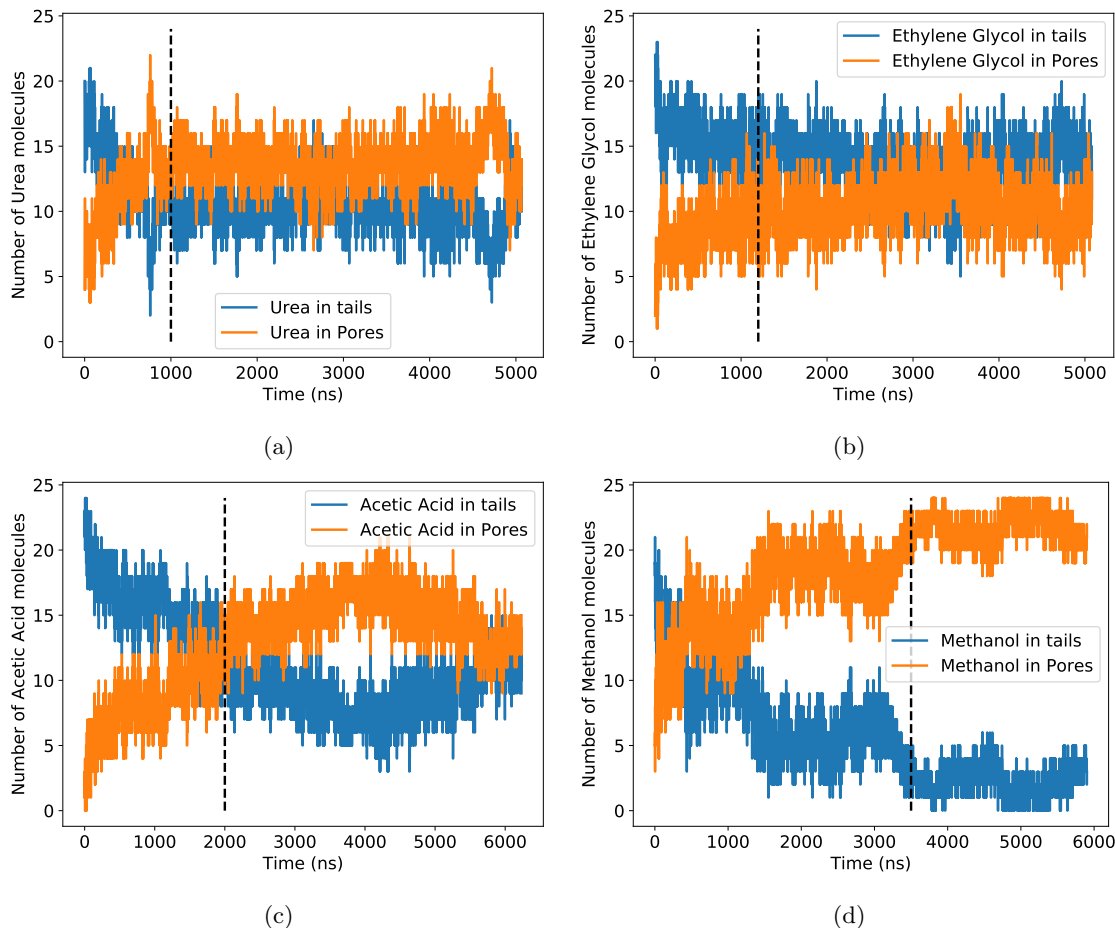


Figure S1: We considered a system to be equilibrated when the partition of solutes between the tails and pore plateaued. Our chosen equilibration point for each solute is indicated by the vertical black dashed line. (a) Urea equilibrates the fastest, after 1000 ns. (b) Ethylene glycol equilibrates after 1200 ns (c) Acetic Acid does not equilibrate until 4400 ns. (d) We considered methanol to be equilibrated after 3500 ns. Methanol nearly completely partitions into the tails.

## S2 Solute Equilibration

We collected all data used for model generation after the solutes were equilibrated. We assumed a solute to be equilibrated when the partition of solutes in and out of the pore region stopped changing. The pore region is defined as within 0.75 nm of the pore center. We've plotted the partition versus time in Figure S1 and indicated the chosen equilibration time points.

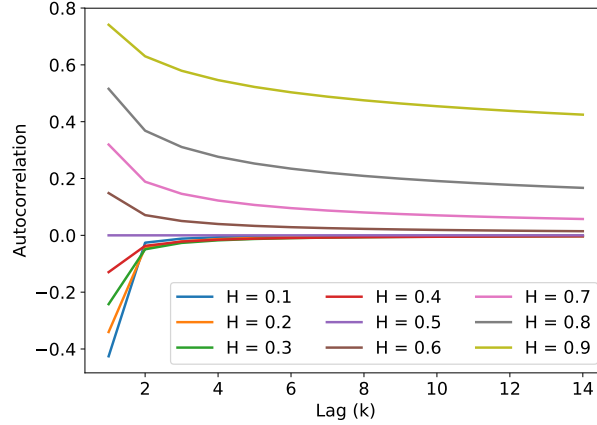


Figure S2: The analytical autocorrelation function of FBM decays to zero faster when  $H < 0.5$  compared to when  $H > 0.5$ .

### S3 Estimating the Hurst Parameter

We chose to estimate the Hurst parameter,  $H$  by a least squares fit to the analytical autocorrelation function for fractional Brownian motion (the variance-normalized version of Equation 6 in the main text):

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

In Figure S2, we plotted Equation 1 for different values of  $H$ . When  $H > 0.5$ , Equation 1 decays slowly to zero meaning one needs to study large time lags with high frequency in order to accurately estimate  $H$  from the data. Fortunately, all of our solutes show anti-correlated motion, so most of the information in Equation 1 is contained within the first few lags.

The autocovariance function of fractional Lévy motion is different from fractional Brownian motion (see Equations 6 and 8 of the main text), but their autocorrelation structures are the same. The autocovariance function of FLM is dependent on the expected value of squared draws from the underlying Lévy distribution,  $E[L(1)^2]$ . This is effectively the distribution's variance, which is undefined for most Lévy stable distributions due to their heavy tails. As a consequence, one should expect  $E[L(1)^2]$  to grow as more samples are drawn from the distribution with the autocovariance function responding accordingly. However, we are only interested in the autocorrelation function. In order to predict the Hurst parameter from the autocorrelation function, we must show that it has a well-defined structure and is independent of the coefficient in Equation 8 of the main text. In Figure S3, we plot the average autocorrelation function from an FLM process with an increasing number of observations per generated sequence. For all simulations we set  $H=0.35$  and  $\alpha=1.4$ . The variance-normalized autocovariance function, i.e. the autocorrelation function, does not change with increasing sequence length. Additionally, the autocorrelation function of FBM, with the same  $H$ , is the same.

### S4 Simulating Fractional Lévy Motion

#### S4.1 Truncated Lévy stable hop distributions

*Determining where to truncate the hop distribution:* A pure Lévy stable distribution has heavy tails which can lead to arbitrarily long hop lengths. Our distribution of hop lengths fits well to a Lévy distribution near the mean, but under samples the tails. In Figure S5 we compare the empirically measured transition emission distribution of the MSDDM for Urea to its maximum likelihood fit to a Lévy stable distribution. The ratio between the two distributions at each bin is nearly 1 close to the center, indicating a near-perfect fit, larger than 1 slightly further from the center, suggesting that we slightly over sample intermediate hop

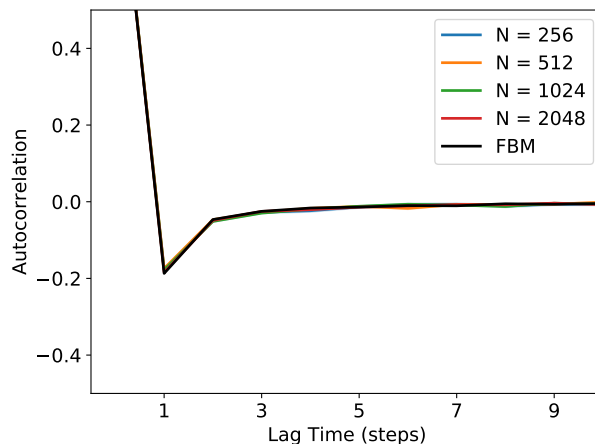


Figure S3: The autocorrelation function of an FLM process does not change with increasing sequence length ( $N$ ). It shares the same autocorrelation function as fractional Brownian motion (FBM). All sequenced used to make this plot were generated using  $H=0.35$  and, for FLM,  $\alpha=1.4$ .

lengths, and below 1 far from the center, indicating under sampling of extremely long hop lengths. Based on the plot, we chose a cut-off of 0.8 nm in order to compensate for over sampled intermediate hop lengths. We used the same procedure to determine truncation parameters for all solutes.

*Generating FLM realizations from a truncated Lévy distribution:* To generate realizations from an uncorrelated truncated Lévy process, one would randomly sample from the base distribution and replace values that are too large with new random samples from the base distribution, repeating the process until all samples are under the desired cut-off.

This procedure is complicated by the correlation structure of FLM. At a high level, Stoev and Taqqu use Riemann-sum approximations of the stochastic integrals defining FLM in order to generate realizations. They do this efficiently with the help of Fast Fourier Transforms. In practice, this requires one to Fourier transform a zero-padded vector of random samples drawn from the appropriate Lévy stable distribution, multiply the vector in Fourier space by a kernel function and invert back to real space. The end result is a correlated vector of fractional Lévy noise.

If one is to truncate an FLM process, one can apply the simple procedure above for drawing uncorrelated values from the marginal Lévy stable distribution, but after adding correlation, the maximum drawn value is typically lower than the limit set by the user. Additionally, the shape of the distribution itself changes. Analogous to the database used to correct the Hurst parameter, we created a database to correct the input truncation parameter (the maximum desired draw). The database returns the value of the truncation parameter that will properly truncate the output marginal distribution based on  $H$ ,  $\alpha$  and  $\sigma$  (the width parameter). Figure S5 shows the result of applying our correction. Note that generating this database requires a significant amount of simulation and still likely doesn't perfectly correct the truncation parameter. The output leads to a somewhat fuzzy, rather than abrupt, cut-off of the output distribution. This is likely beneficial since we observe a small proportion of hops longer the chosen truncation cut-off. When the cut-off value is close to the Lévy stable  $\sigma$  parameter, as it is in our anomalous diffusion models, we observed that the tails of the truncated distribution tend to be undersampled. In order to maintain the distribution's approximate shape up to the cut-off value we recommend ensuring that the cut-off value is at least 2 times  $\sigma$ . However, this may lead to a slight over-prediction of the MSD.

## S4.2 Achieving the right correlation structure

We simulated FLM using the algorithm of Stoev and Taqqu [1]. There are no known exact methods for simulating FLM. As a consequence, passing a value of  $H$  and  $\alpha$  to the algorithm does not necessarily result in the correct correlation structure, although the marginal Lévy stable distribution is correct. We applied a database-based empirical correction in order to use the algorithm to achieve the correct marginal distribution

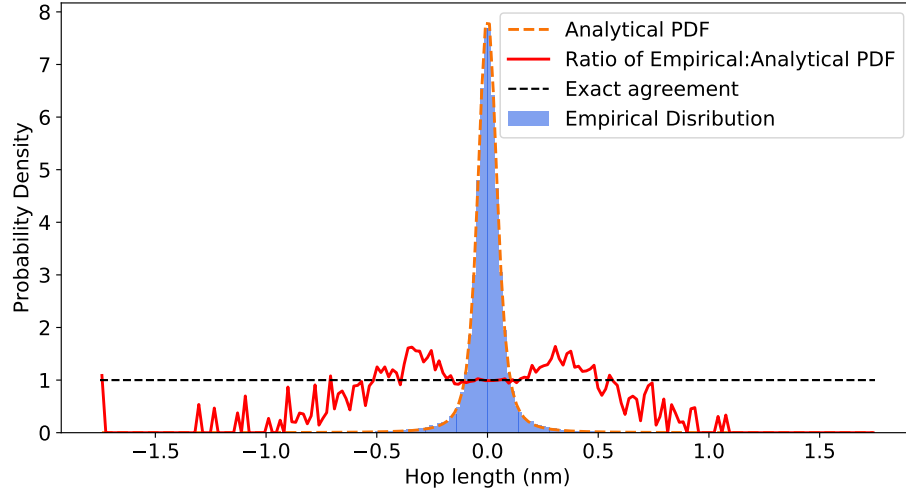


Figure S4: The ratio between the empirical and maximum likelihood theoretical distribution quantifies the quality of fit as function of hop length. The fit is near-perfect close to the mean. Intermediate hop lengths are over sampled, and the tails are under sampled. We used this type of plot to determine the appropriate place to truncate the Lévy stable distributions.

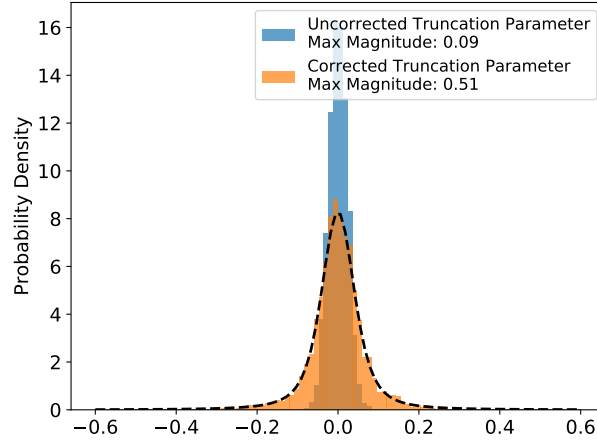


Figure S5: We can accurately truncate the marginal distribution of FLM innovations by applying a correction to the input truncation parameter. We generated FLM sequences and truncated the initial Lévy stable distribution (before Fourier transforming) at a value of 0.5. After correlation structure is added, the width of the distribution of fractional Lévy noise decreases significantly. We corrected the input truncation parameter with our database resulting in a distribution close to the theoretical distribution (black dashed line) with a maximum value close to 0.5.

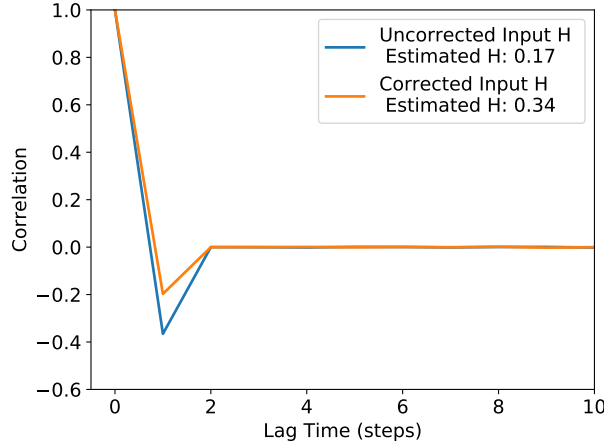


Figure S6: Correcting the Hurst parameter input to the algorithm of Stoev and Taqqu results in an FLM process with a more accurate correlation structure. We generated sequences with an input  $H$  of 0.35. We estimated  $H$  by fitting the autocorrelation function. Without the correction,  $H$  is underestimated, meaning realizations are more negatively correlated than they should be.

and correlation structure.

Stoev and Taqqu note that the transition between negatively and positively correlated draws occurs when  $H = 1/\alpha$ . When  $\alpha = 2$ , the marginal distribution is Gaussian and the transition occurs at  $H = 0.5$  as expected from FBM. We corrected the input  $H$  so that the value of  $H$  measured based on the output sequence equaled the desired  $H$ . We first adjusted the value of  $H$  by adding  $(1/\alpha - 0.5)$ , effectively recentering the correlation sign transition for any value of  $1 \leq \alpha \leq 2$ . This correction alone does a good job for input  $H$  values near 0.5, but is insufficient if one desires a low value of  $H$ . The exact correction to  $H$  is not obvious so we created a database of output  $H$  values tabulated as a function of input  $H$  and  $\alpha$  values. Figure S6 demonstrates the results of applying our correction. Without the correction, FLM realizations are more negatively correlated. This would result in under-predicted mean squared displacements when applying the model.

## S5 Verifying Markovianity

We verified the Markovianity of our transition matrix,  $T$ , in two ways. First we ensured that the process satisfied detailed balance:

$$T_{i,j}P_i(t = \infty) = T_{j,i}P_j(t = \infty) \quad (2)$$

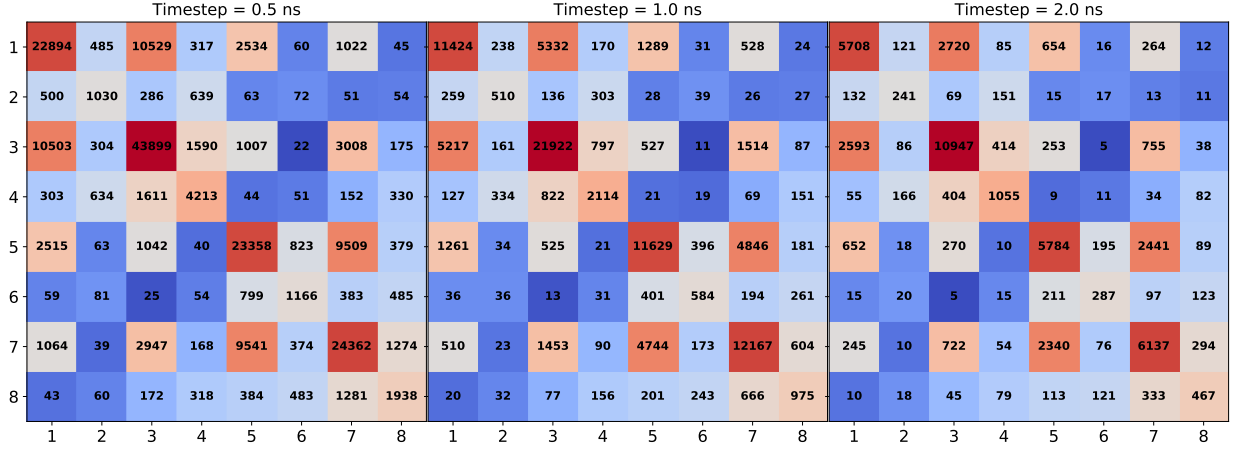
where  $P$  is the equilibrium distribution of states. This implies that the number of transitions from state  $i$  to  $j$  and from state  $j$  to  $i$  should be equal. Graphical representations of the count matrices show that this is true in Figure S7.

Second, we ensured that the transition matrix did not change on coarser time scales. In Figures S7 and S8, we show that increasing the length of time between samples does not change the properties of the count or probability transition matrices.

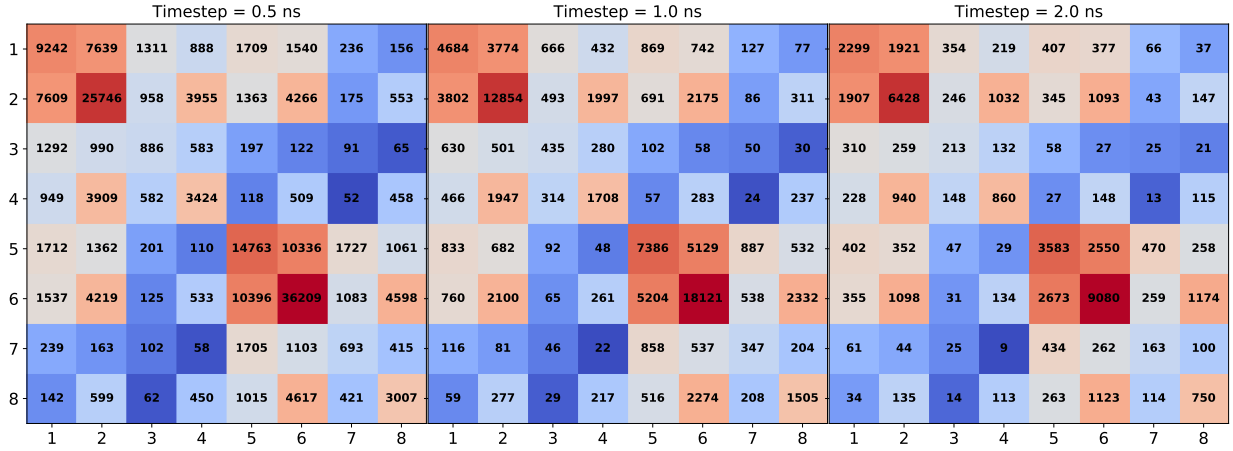
## S6 Derivation of Passage Time Distributions

To derive an analytical equation for the mean first passage time as calculated in the main text, first consider an initial pulse spreading out over time with a fixed mean. We can solve for the time-dependent probability density of particle positions by solving the one dimensional diffusion equation:

$$\frac{\partial p}{\partial t} = D \frac{\partial^2 p}{\partial z^2} \quad (3)$$

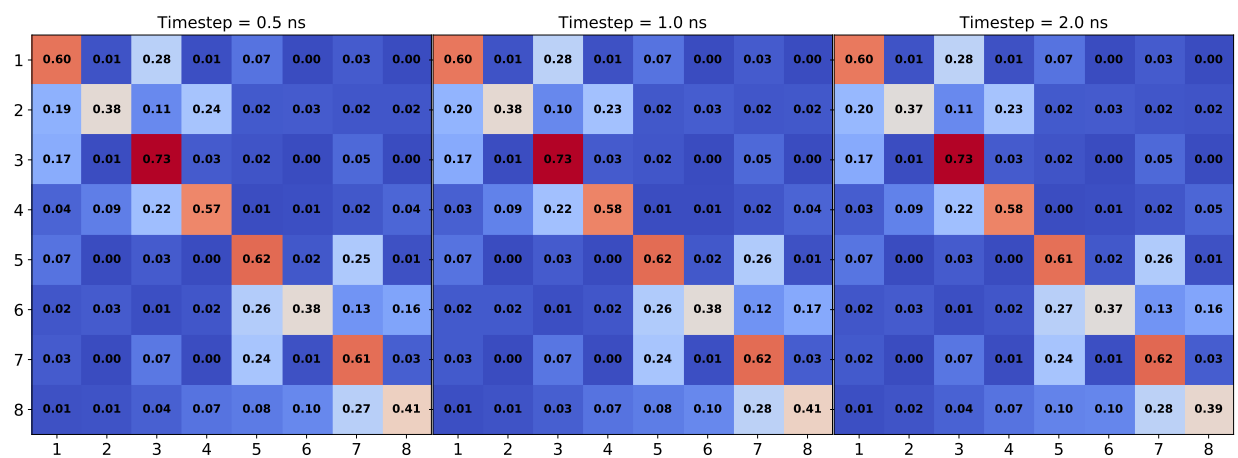


(a) Urea

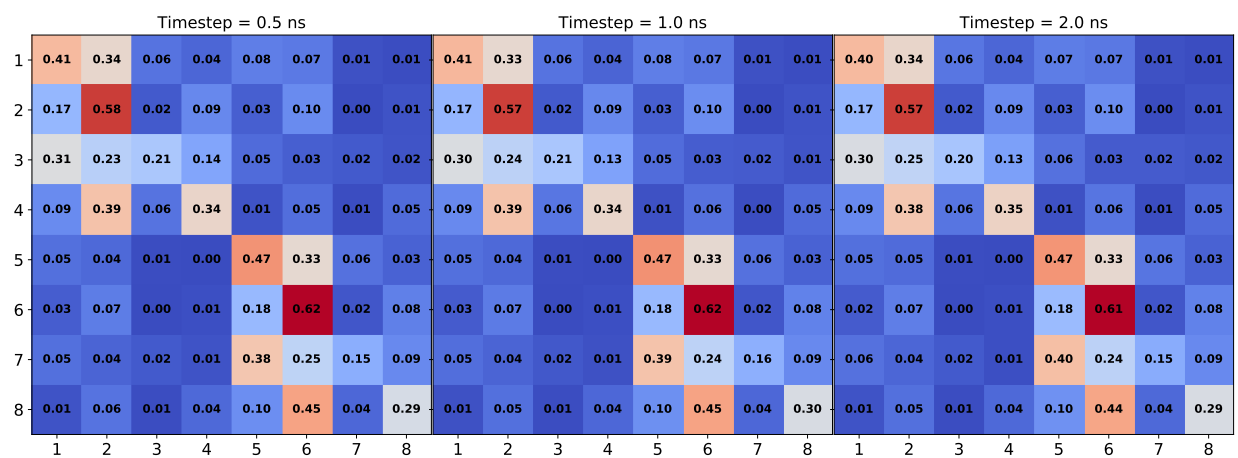


(b) Ethylene Glycol

Figure S7: The number of transitions from state  $i$  to  $j$  and  $j$  to  $i$  are very close indicating that our process obeys detailed balance. Detailed balance is conserved for different sized time steps.



(a) Urea



(b) Ethylene Glycol

Figure S8: As the timestep between observations increases, the probability transition matrix does not change significantly.



The appropriate initial and boundary conditions are:

$$BC1 : t > 0, z = \infty, p = 0$$

$$BC2 : t > 0, z = 0, \frac{\partial p}{\partial z} = 0$$

$$IC : t = 0, c = M\delta(z)$$

where  $M$  is the total number of particles in the system. It has been shown elsewhere that the solution to this equation is: [?]

$$p(z, t) = \frac{1}{\sqrt{4\pi Dt}} \exp\left(\frac{-z^2}{4Dt}\right) \quad (4)$$

We can make the substitution  $z = z - vt$ , where  $v$  represents a constant average velocity, in order to linearly shift the mean as a function of time:

$$p(z, t) = \frac{1}{\sqrt{4\pi Dt}} \exp\left(\frac{-(z - vt)^2}{4Dt}\right) \quad (5)$$

One can track the fraction of particles,  $F$ , that have crossed the pore boundary by integrating:

$$F(t) = \int_L^\infty p \, dz = \text{erfc}\left(\frac{L - vt}{2\sqrt{Dt}}\right) \quad (6)$$

where  $L$  is the pore length. This represents the cumulative first passage time distribution so we take its derivative in order to arrive at the first passage time distribution:

$$P(t) = -\frac{1}{\sqrt{\pi}} e^{-(L-vt)^2/(4Dt)} \left( -\frac{D(L-vt)}{4(Dt)^{3/2}} - \frac{v}{2\sqrt{Dt}} \right) \quad (7)$$

where the only free parameters for fitting are  $v$  and  $D$ . We calculated the expected value of Equation ?? in order to get the MFPT.

## S7 Solute hopping and trapping behavior

## S8 Tabular Anomalous Diffusion Parameters

1 Mode Model	Parameters	Urea	Ethylene Glycol	Methanol	Acetic Acid
Dwell Distributions	$P(\alpha_d)$	0.57	0.62	0.62	0.45
	$P_T(\alpha_d, \lambda)$	0.40, 0.0024	0.47, 0.0030	0.44, 0.0040	0.08, 0.0033
Hop Distributions	$\mathcal{N}(\sigma)$	0.33	0.34	0.35	0.27
	$L(\sigma, \alpha_h)$	0.21, 1.84	0.23, 1.92	0.22, 1.80	0.16, 1.72
Correlation	$\gamma(H)$	0.37	0.40	0.30	0.34

Table S2: To create a 1 mode model for each solute, we parameterized a pure power law ( $P(\alpha_d)$ ) and a truncated power law ( $P_T(\alpha_d, \lambda)$ ) distribution to describe solute dwell times. Lower values of  $\alpha_d$  lead to heavier power law tails and higher values of  $\lambda$  truncate the distribution at lower dwell times. We parameterized Gaussian ( $\mathcal{N}(\sigma)$ ) and Lévy stable ( $L(\sigma, \alpha_h)$ ) distributions to describe solute hop lengths. We assume the mean ( $\mu$ ) to be zero for these distributions and there to be no skewness ( $\beta = 0$ ) in the Lévy stable distributions. High values of  $\sigma$  and lower values of  $\alpha_h$  result in larger hops. Finally, we parameterized the hop autocorrelation function ( $\gamma(H)$ ) to describe the degree of correlation between hops. Higher values of  $H$  display closer to Brownian behavior.

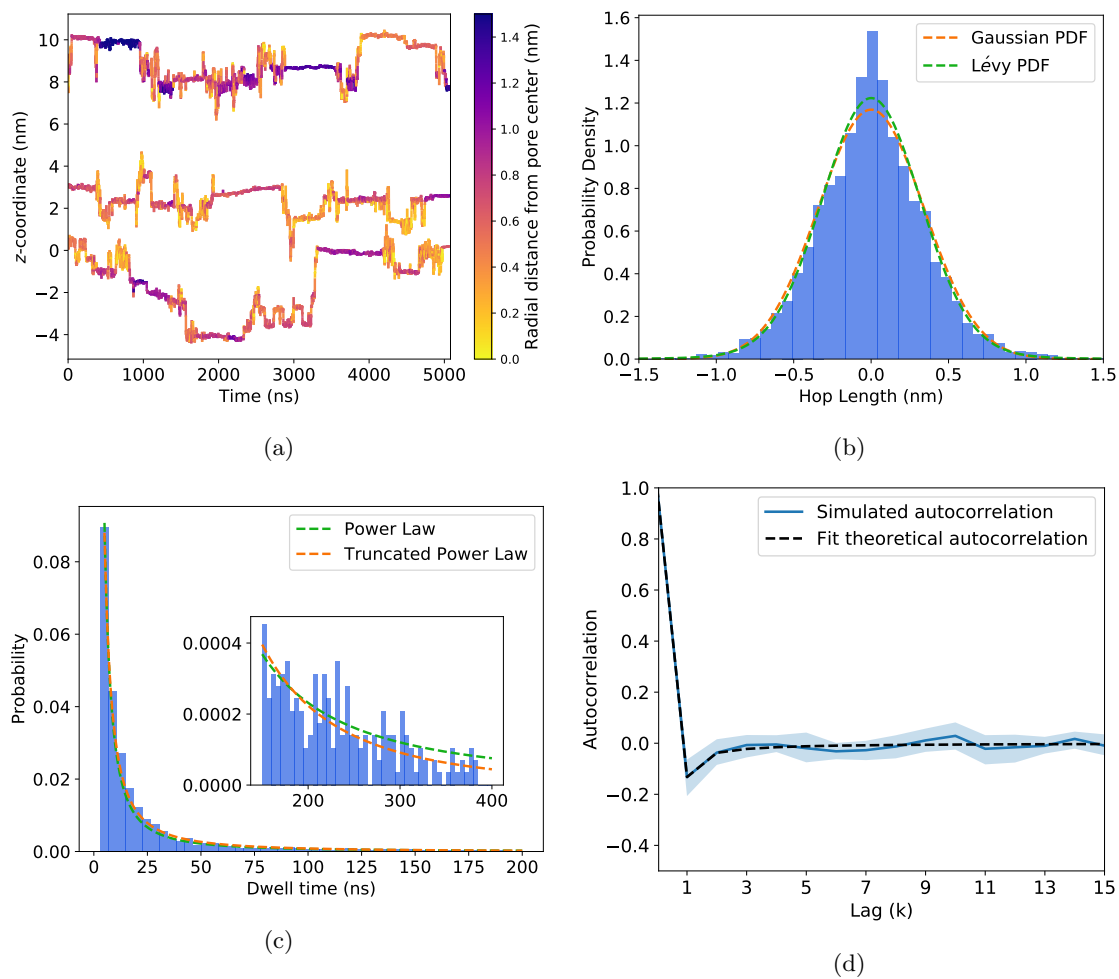


Figure S9: For ethylene glycol: (a) representative  $z$ -coordinate center of mass trajectories, (b) hop length distributions, (c) dwell time distributions and (d) hop autocorrelation functions for ethylene glycol. See Figure 2 and Section 3.1.1 in the main text for a more detailed discussion.

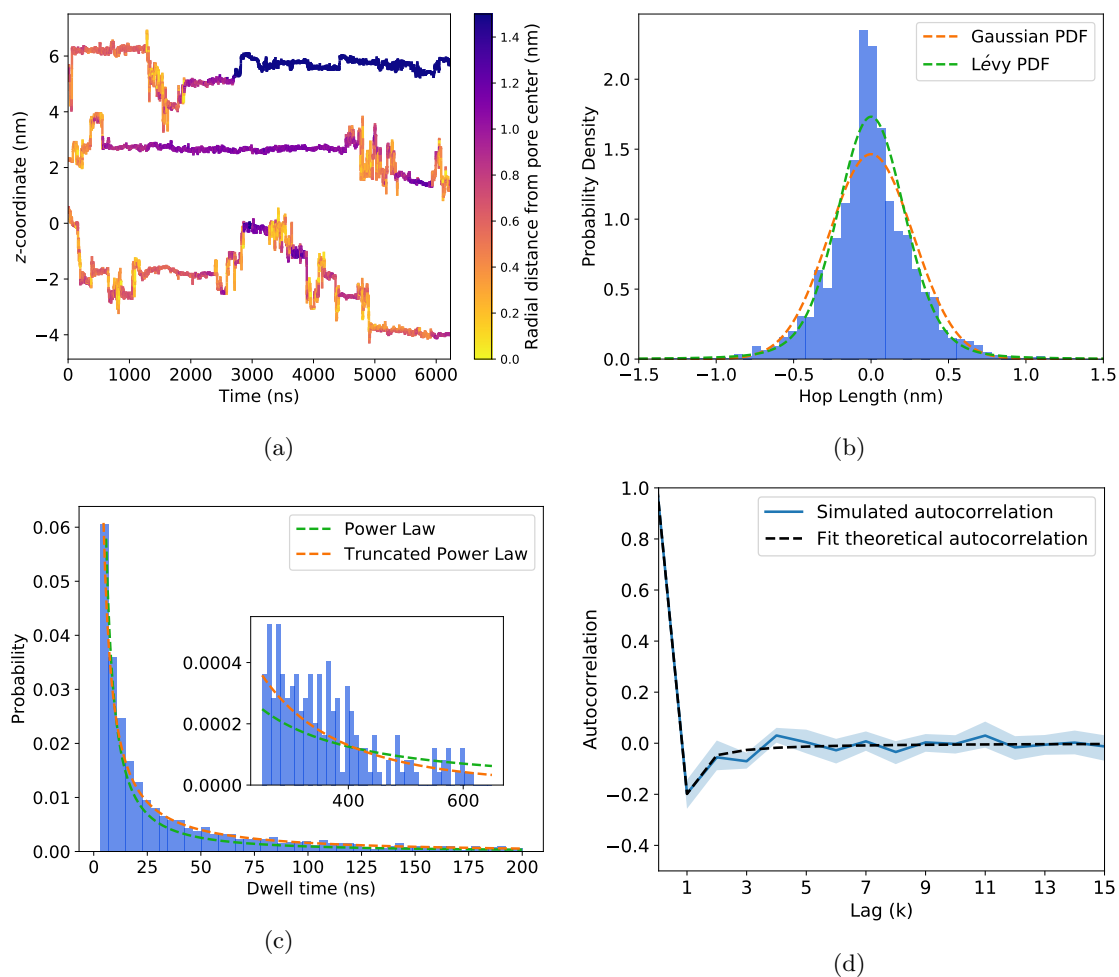


Figure S10: For acetic acid: (a) representative  $z$ -coordinate center of mass trajectories, (b) hop length distributions, (c) dwell time distributions and (d) hop autocorrelation functions for acetic acid. See Figure 2 and Section 3.1.1 in the main text for a more detailed discussion.

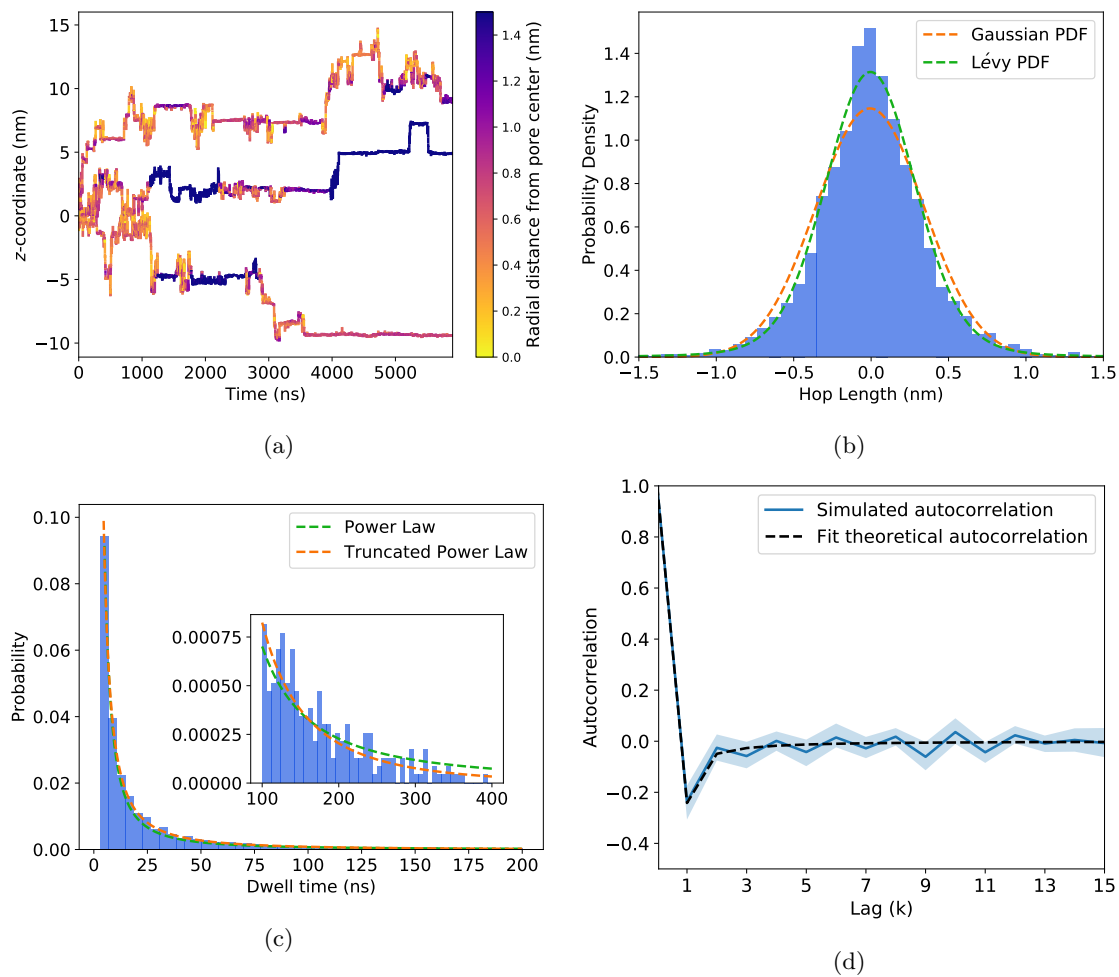


Figure S11: For methanol: (a) representative  $z$ -coordinate center of mass trajectories, (b) hop length distributions, (c) dwell time distributions and (d) hop autocorrelation functions for methanol. See Figure 2 and Section 3.1.1 in the main text for a more detailed discussion.

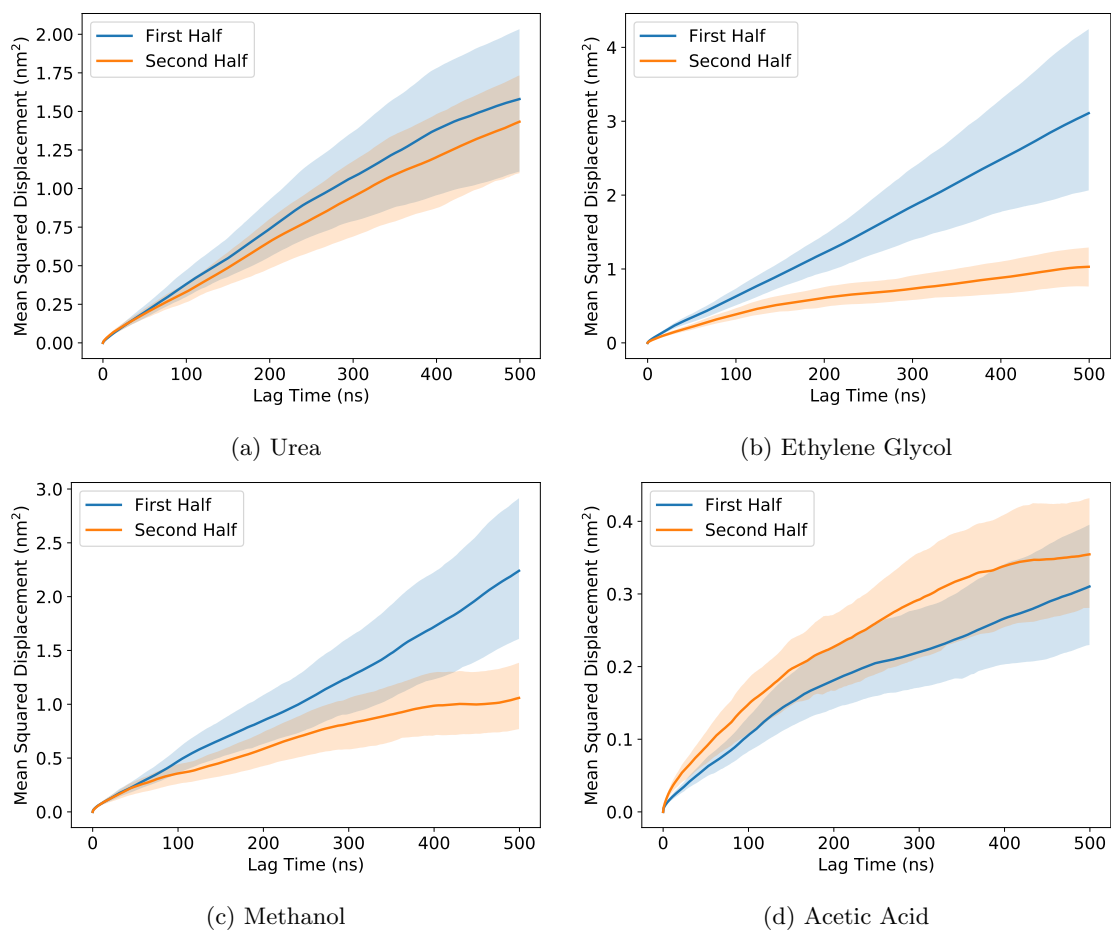


Figure S12: The ensemble of solute trajectories may be stationary if the MSD calculated from different portions of the trajectory are the same. Here we plot the MSD calculated up to a 500 ns time lag of the first and second halves of the equilibrated solute trajectories. Urea and acetic acid have similar MSDs, providing evidence of stationarity, while the MSDs of ethylene glycol and methanol are different.

2 Mode Model						
	Parameters	Mode	Urea	Ethylene Glycol	Methanol	Acetic Acid
Dwell Distributions	$P(\alpha_d)$	1	0.69	0.69	0.90	0.58
		2	0.38	0.48	0.58	0.33
	$P_T(\alpha_d, \lambda)$	1	0.56, 0.0037	0.62, 0.0026	1.04, 0.0006	0.41, 0.0026
		2	0.00, 0.0027	0.06, 0.0049	0.30, 0.0054	0.00, 0.0021
Hop Distributions	$\mathcal{N}(\sigma)$	1	0.35	0.38	0.45	0.32
		2	0.24	0.23	0.32	0.17
	$L(\sigma, \alpha_h)$	1	0.24, 1.91	0.26, 1.99	0.31, 1.97	0.21, 1.91
		2	0.12, 1.50	0.15, 1.90	0.20, 1.85	0.09, 1.50
Correlation	$\gamma(H)$	–	0.37	0.40	0.30	0.34

Table S3: The two model parameterizes solute behavior in the pores and tails separately. Generally, movement is much more restricted in the tail region. To create a 2 mode model, we generated a set of parameters based on solute behavior as function of distance from the pore center. Mode 1 corresponds to solute behavior within 0.75 nm of the pore center and mode 2 corresponds to behavior greater than or equal to 0.75 nm from the pore center. Note that we used the same Hurst parameter for both modes due to a low number of sufficiently long sequences of hops in each mode. See Table S2 for descriptions of the parameters.

## S9 Stationarity of Solute Trajectories

## S10 Tabular MSDDM parameters

State	Urea			Ethylene Glycol			Methanol			Acetic Acid		
	H	$\alpha_h$	$\sigma$	H	$\alpha_h$	$\sigma$	H	$\alpha_h$	$\sigma$	H	$\alpha_h$	$\sigma$
1	0.10	1.79	0.034	0.09	1.68	0.045	0.11	1.56	0.052	0.10	1.78	0.035
2	0.06	1.80	0.033	0.09	1.75	0.037	0.07	1.63	0.043	0.08	1.88	0.032
3	0.11	1.88	0.030	0.11	1.86	0.030	0.02	1.80	0.036	0.04	2.00	0.030
4	0.10	1.95	0.027	0.04	1.91	0.028	0.02	1.75	0.036	0.04	2.00	0.027
5	0.19	1.34	0.048	0.15	1.40	0.062	0.10	1.28	0.074	0.13	1.47	0.048
6	0.15	1.45	0.040	0.11	1.52	0.040	0.03	1.50	0.042	0.09	1.70	0.038
7	0.15	1.61	0.032	0.05	1.60	0.040	0.28	1.20	0.043	0.08	1.77	0.031
8	0.11	1.71	0.028	0.05	1.74	0.030	0.04	1.83	0.037	0.01	2.00	0.030
T	0.34	1.42	0.036	0.37	1.44	0.045	0.35	1.45	0.057	0.34	1.54	0.040

Table S4: We calculated values of  $H$ ,  $\alpha_h$  and  $\sigma$  from MD simulation trajectories and used them to generate realizations of our MSDDM model. The states are defined in Table ?? except state T which describes the transition emissions.

## S11 2-state MSDDM parameters

We used a simple 2-state model to illustrate how the self-transition probability effects simulated MSDs generated from the MSDDM. We used the following parameters:

## S12 Mean first passage time simulations

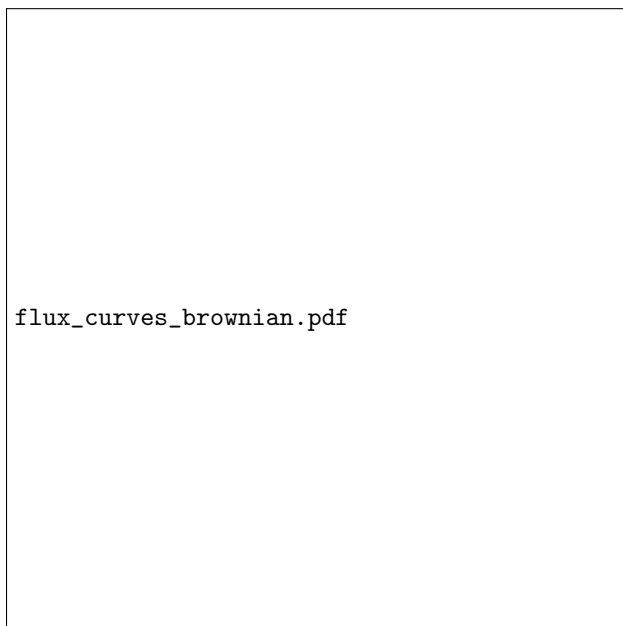


Figure S13: Fixed  $H = 0.5$

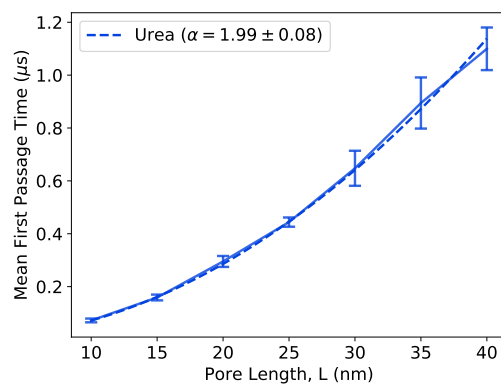


Figure S14

State	H	$\alpha_h$	$\sigma$
1	0.1	1.7	0.04
2	0.2	1.5	0.04
T	0.4	1.4	0.04

Table S5: Values of  $H$ ,  $\alpha_h$  and  $\sigma$  used to simulate realizations of the simple MSDDM whose MSDs are shown in Figure 12 of the main text. The values are chosen to be similar to those in Table ?? of the main text.



## References

- [1] S. Stoev and M. S. Taqqu, “Simulation methods for linear fractional stable motion and farima using the fast fourier transform,” *Fractals*, vol. 12, pp. 95–121, Mar. 2004.