Transient, often anomalous and heterogeneous, diffusive transport through Nature's favorite barrier fluid: Mucus

Greg Forest, UNC Chapel Hill

Mathematics, Biomedical Engineering, Community Organizer

May 19, 2014

Frontier Probability Days 2014

Tucson, Arizona

Acknowledgements

UNC Virtual Lung Project:

Faculty + Students from Chemistry, Computer Science, Mathematics, Physics, Pharmacology, the Cystic Fibrosis & Pulmonary Medicine Center

Primary Collaborators for this (and Scott's) Lecture:

UNC: D. Hill (Medicine), J. Mellnik (Bioinf & Comp Bio)

S. McKinley (Fla.), N. Pillai (Harvard), M. Lysy (Waterloo)

P. Vasquez (So. Carolina), J. Fricks (Penn St), C. Hohenegger (Utah)

L. Yao (Case Western)

UNC: T. Elston (Pharmacology), R. Superfine (Physics – Microscopy guru)

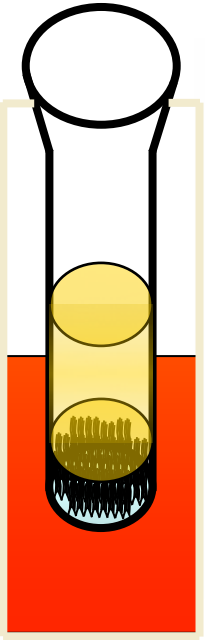
Research Support:

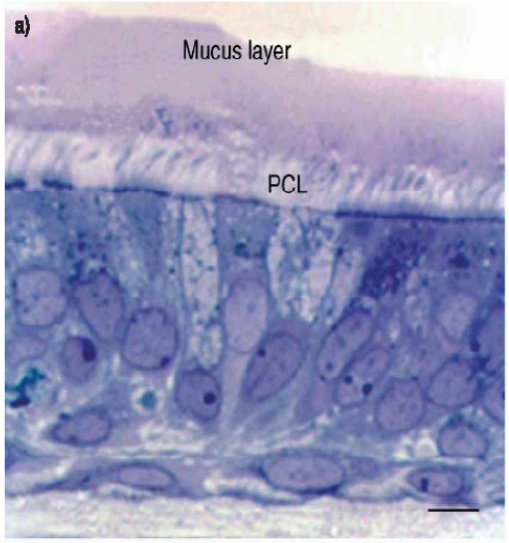
NSF-NIGMS ("mucus" grant w/ D. Hill), NIH

Mucus: the lung's primary defense mechanism

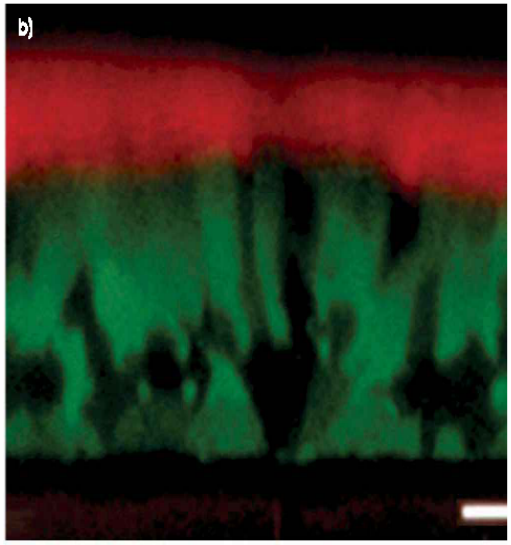
- 5-50 micron mucus layer; 7 micron liquid "PCL" layer for 8 micron cilia
- Oxygen-rich air must reach the alveoli to achieve oxygen exchange
- Price to pay: trap and clear all "foreign particles" to larynx to be swallowed
- Mucus is Nature's fly paper---inhaled particles land and stick, and estimates are that the **particle-laden mucus escalator transports up to 60 microns/sec**
- **Particles (viruses, bacteria, drug carriers, smoke, dust, pollen) diffuse in the mucus layer, many hoping to traverse the barrier to reach epithelial cells**
- ***Lungs are sterile iff flow of mucus dominates diffusive transport through the layer***

Human cell cultures





Biochemical composition
MUC5AC MUC5B



Our lungs are **chemically tuned to inhibit diffusion of inhaled pathogens** over a wide size spectrum and surface chemistry. It is called evolution ☺

Successful pathogens actively respond, e.g., viruses dress themselves with local proteins to go undetected and thereby diffuse freely.
It is called evolution ☺

Cartoon of mucus layers in lung pathways

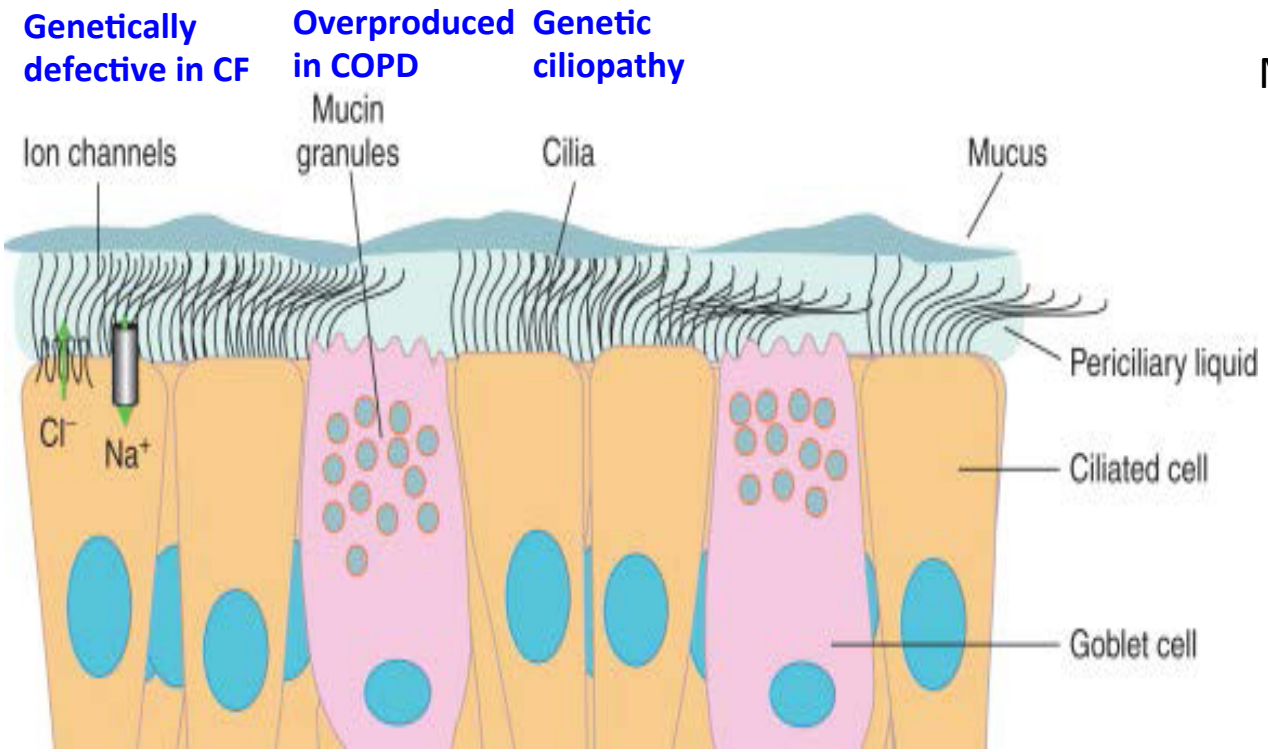
Mucus is a "hydrogel" of large molecular weight mucin molecules, salts & proteins together with trapped airborne assaults and immunological response agents

Mucus is driven by cilia (10-15 Hz), tidal breathing (.2 Hz), cough (freq. spectrum)

In healthy lungs the mucus barrier flows toward the larynx, and must be replenished

Mucin macromolecules are continuously injected and water volume is regulated such that transport properties yield efficient clearance—continuous mechano-chemical feedback.

In disease (CF, COPD, asthma) mucus morphs: ***we seek biophysical / rheological markers of disease progression and physical/drug therapies to steer mucus to a healthy state***

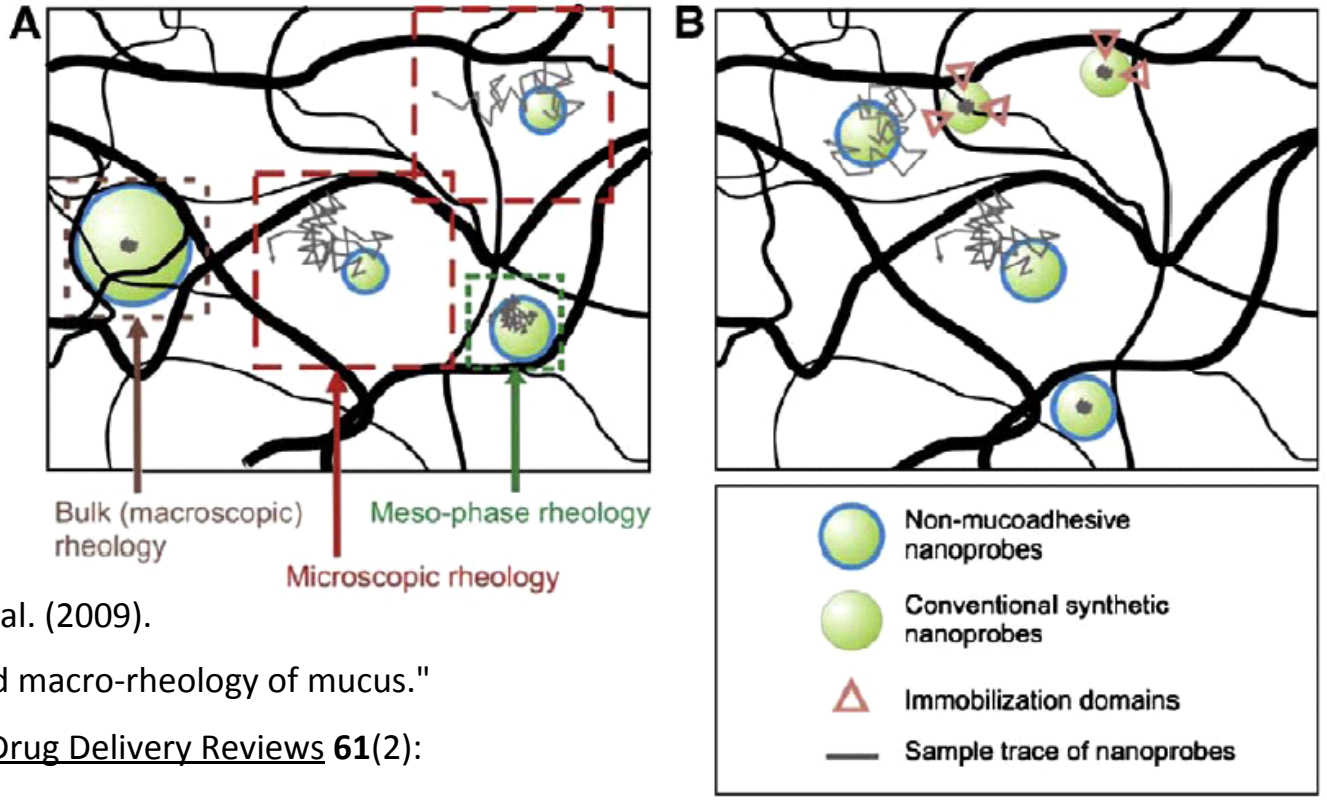


Mucus layer is not to scale here!

7 micron "PCL"
8 micron cilia
5-50 micron mucus layer

Microstructure Relative to Bead Dimensions & Surface Chemistry

For drug inhalation particle delivery (diabetes, cancer, COPD, CF): our target is to **control passage time distributions**. To do that, you need to understand what the stochastic path processes are, build inference tools applicable to path data, then derive or simulate PT distributions per particle per patient mucus. N.B. There is no physical theory of mucus that guides the diffusive process for one particle, much less all possible particles.

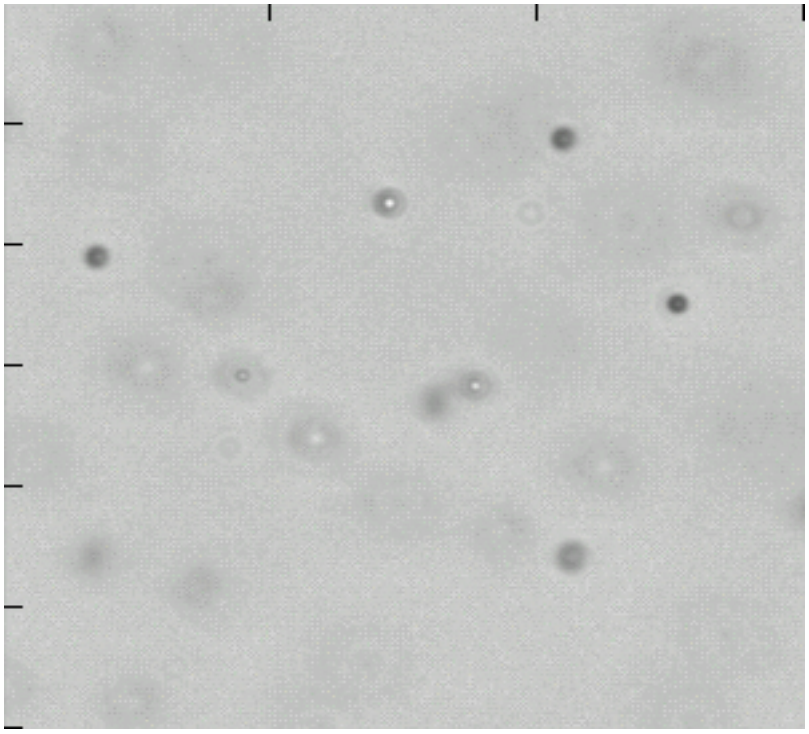


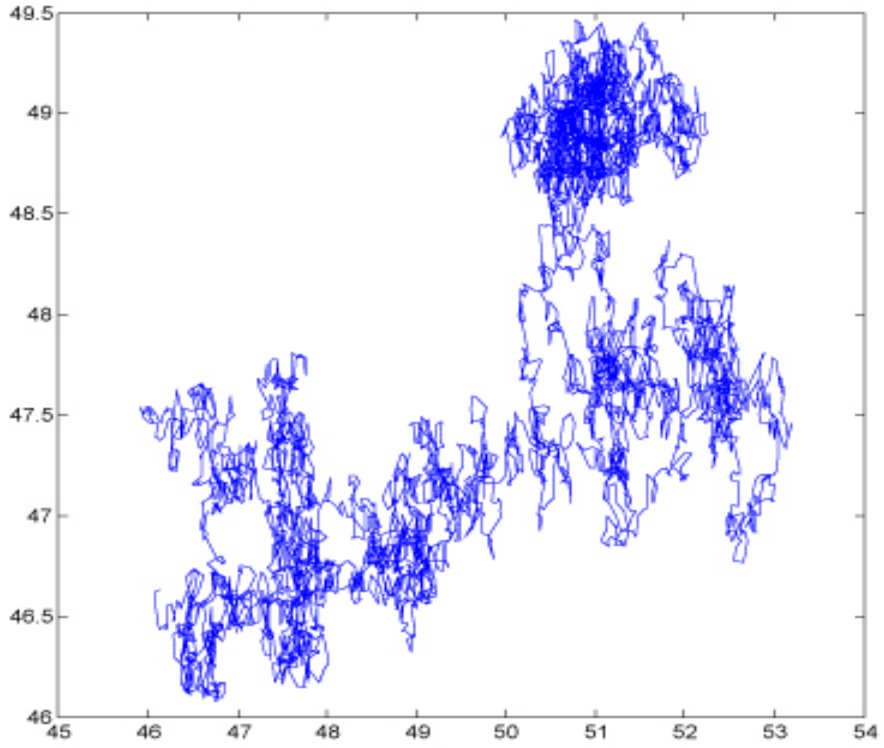
Lai, S. K. et al. (2009).

"Micro- and macro-rheology of mucus."

Advanced Drug Delivery Reviews **61**(2): 86-100.

Incredible microscopy tools for path data (~1 nm, 10^{-3} sec resolution)
CISSM @ UNC, R. Superfine, David Hill, Cystic Fibrosis Center
http://cismm.cs.unc.edu/





UPSHOT: we can observe real paths at a user-chosen sampling rate
With lung cell cultures, we can explore mucus from different cell lines or during disease progression with "bugs" or designer particles.

Pulmonary delivery

- COPD
- Asthma
- CF
- Vaccines
- Gene therapies
- Insulin treatments
- Cancer treatments

Marketing aside, our lungs are engineered with a mucus barrier to trap and clear everything that comes in with inhaled air, including drug carrier particles.

Drug protocols must account for the mucus barrier to calculate dosage – lung vs gut vs time.

The fundamental role of probability and statistics in lung biology

- Our challenge begins with **experimental path data on specific particles in pulmonary mucus**, in order to:

--detect disease, assess disease progression, guide and evaluate drug and physical therapeutics in a clinical setting, *based on an individual patient's mucus transport properties* – *the miners' canary*

How do we propose to do that from path data?

--**forecast first passage time distributions vs mucus layer thickness of any given particle species in a patient's mucus** *and*

--forecast the flow transport timescales of mucus in lung airways by coordinated cilia, and air drag from tidal breathing and cough

- **You have one week, should you accept this mission.**

Human Bronchial Epithelial Cell (HBE) Cultures

- Cells generate cilia, produce mucus & PCL
- **Non-invasive mucus source**: harvest and tune to 2-2.5% solids (normal) to 4-8% solids (CF, COPD) to mimic disease progression

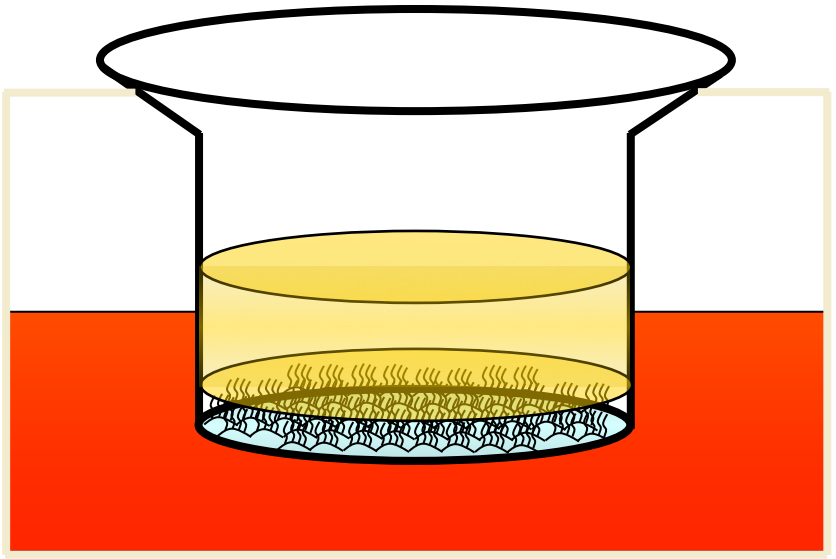
Transformational for lung biology: model system

1. Imaging of "living" lung tissues
2. Observe cilia beat cycles, measure stall forces
3. Observe mucin secretion, replenishment of mucus layer overnight after removal
4. Measure stress-dependent ATP release, biochemical cascades, dynamics of water flux (cf. Elston, Forest, et al. J Theor Biol **325** 2013)

Drawbacks: Flow: cultures are not faithful to lung airway geometry; some observations may be a result of the geometry.

Diffusion: cultures are not usually exposed to pathogenic assaults & immune responses.

Ongoing studies of chemical and bio assaults to examine the consequences for both transport properties.



- Grown on a collagen-coated sheet of Gore-Tex
- Cells receive nutrients from basal side, just as in body
- Cilia appear ~ 1 month after plating cells, beat, coordinate **(how?)** and create "Mucus Hurricanes"

Challenges to probe mucus & infer transport properties

- Availability (microliters)
- Mucus "yields" easily— nonlinear transitions depend on lengthscale, frequency & amplitude of applied forces
- Literature values for dynamic moduli and "diffusion rates" differ by orders of magnitude
- What factors are responsible?
- Can you tease apart physiological from experimental variability?

- Experimental limitations: typical rheometers need milliliter volumes, and possibly exceed mucus yield thresholds, conflating linear and nonlinear responses.
- So, we go small: microrheometry
- We "see" remarkable things with advanced microscopy…do we understand what we're seeing, in passive and active regimes of particle motion?

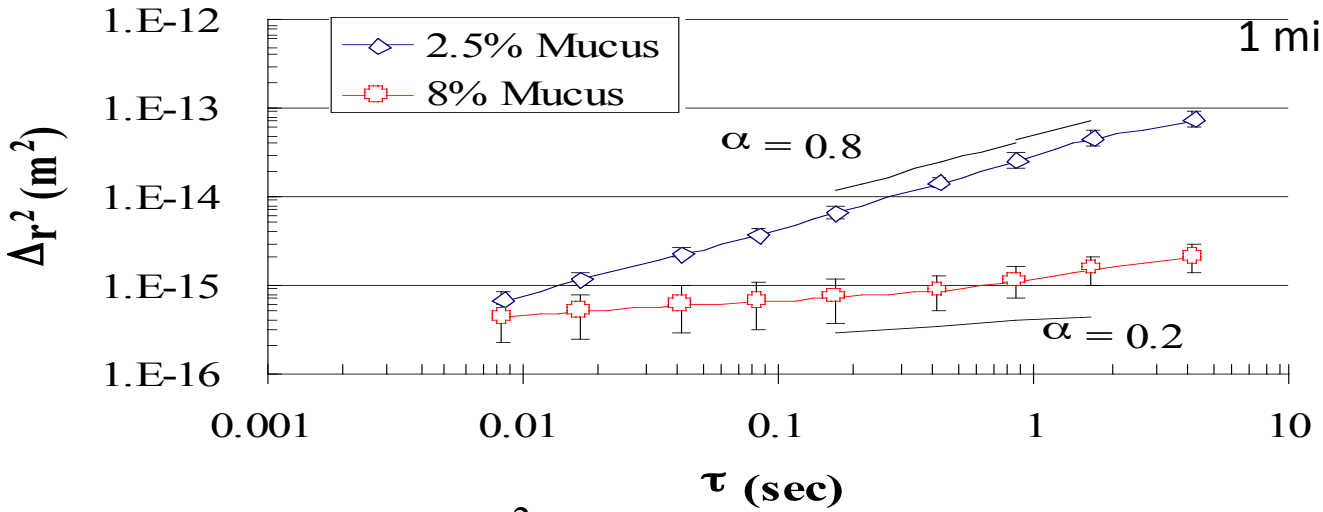
Little agreement on linear viscoelasticity of mucus at physiological frequencies, even less about nonlinear thresholds of mucus relative to physiological force scales, yet less about diffusive transport properties of any inhaled particle in your lung mucus.

Passive microrheology 1996, Mason & Weitz, exploits fluctuation-dissipation to learn linear viscoelastic moduli from fluctuation spectra, simultaneously across broad frequency spectrum

Not conceived for, yet ideal as, direct observation of diffusive transport of particles in mucus.
From HBE cultures, we can explore mucus from very different cell lines & diverse particles.

What statistical properties do path data possess in "(un)healthy" mucus?
What can we infer from the data about the underlying statistical processes?

Upshot: sub-diffusive scaling, sensitive to particle size, surface chemistry and mucus source.
Two plots of MSD below spanning mucus from "healthy" and "advanced disease" wt% solids
We will show later, we observe every exponent in MSD scaling between 0 and 1.

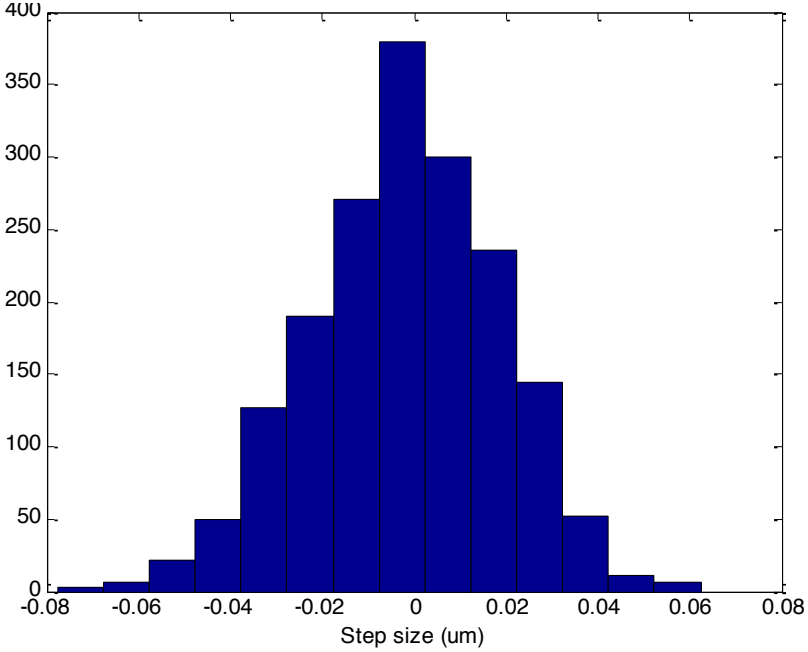


1 micron bead data

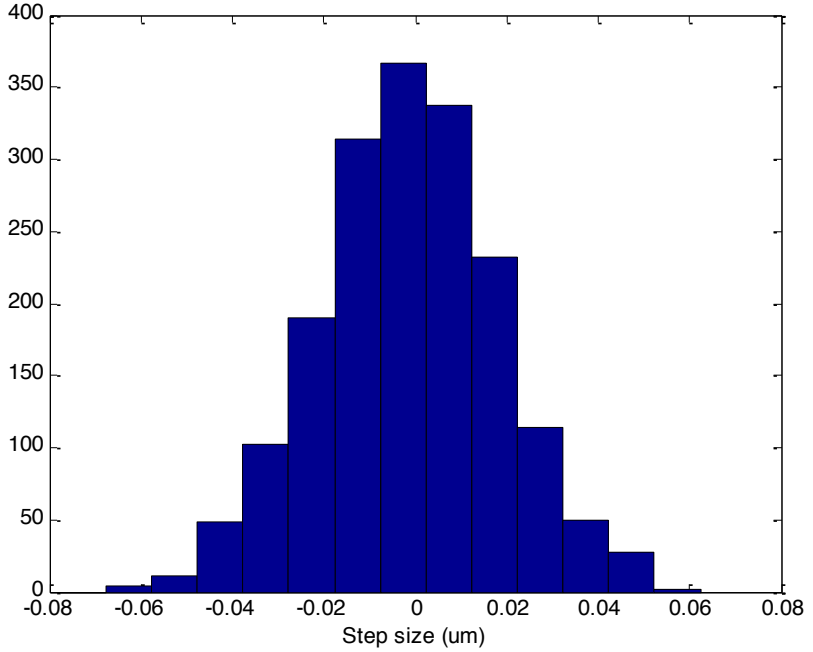
$$\Delta r^2 \sim t^{\alpha}, 0 < \alpha < 1$$

Increments

ΔX



ΔY



Good news: increment process is ~ Gaussian for 1 micron beads in 2.5 wt% mucus
But: of the *many* stochastic processes that have Gaussian increments,
what features discriminate among Gaussian processes, what models are consistent?

60 frames per second, for 30 seconds, is a typical dataset

Our first approach (w/ T. Elston, J. Fricks, L. Yao): *generalized Langevin equations*, "industry standard" in passive microrheology: nobody had yet attempted *inverse characterization & forecasting via simulation*
They were interested in viscoelasticity (F-transform of ensemble MSD)

- Existing exactly solvable models: Rouse and Zimm kernels for GLEs obey MSD scaling of $t^{1/2}$ and $t^{2/3}$ respectively, but **data spans** t^{α} for $0 < \alpha < 1$ for a finite window of lag times. Note: .5 & .66 are not dense in (0,1) ☺
- Existing direct simulation algorithms for GLEs: none (except Kremer-Grest molecular dynamics for monodisperse polymer chains)
- Existing inference methods of GLE kernel and memory spectrum from particle time series data: none Instead, rely on fits to MSD scaling exponent, declare victory for comparison with other complex fluids only
- **Put these three facts together: insufficient models to fit experimental data, so hopeless situation for simulating and forecasting passage time distributions**

Established principles and tools for the inverse problem and statistically accurate simulation of viscous diffusion that this audience already knows

Exploit Langevin (Ornstein-Uhlenbeck process) properties: Markovian and Gaussian

$$m\dot{v}(t) = -\xi v(t) + \tilde{f}(t)$$

where $\xi = 6a\pi\eta$ (Stokes-Einstein drag law)

The random force $\tilde{f}(t)$ satisfies, by fluctuation-dissipation:

$$\left\langle \tilde{f}(t)\tilde{f}(s)\right\rangle = k_B T\xi\delta(t-s)$$

A straightforward, 2 parameter (α and σ) stochastic ODE

$$\frac{dv(t)}{dt} = -\alpha v(t) + \sigma f(t)$$

With $\alpha = \xi/m, \sigma^2/\alpha = 2k_B T/m$ and $\left\langle f(t)f(s)\right\rangle = \delta(t-s)$

Next, the sensible way to infer α and σ from path data

Generation of discrete paths guaranteed to be statistically consistent with the process

$$v(t) = e^{-\alpha t} v(0) + \sigma \int_0^t e^{-\alpha(t-s)} f(s) ds$$

Ito quadrature provides an Auto-Regressive (AR) discrete process

$$v_n = e^{-\alpha\Delta} v_{n-1} + \varepsilon_n, \qquad \Delta \text{ is sampling rate}$$

where ε_n are independent normal random variables w/ variance

$$s = \sigma^2 \frac{1 - e^{-2\alpha\Delta}}{2\alpha}$$

Note: Euler discretization is just the 1st order approximation of this statistically exact discretization. This observation is exploited for both direct simulations & inversion/inference. **Time series are guaranteed to be statistically consistent with the O-U process, and not polluted by cumulative numerical error.**

Inversion: Maximum Likelihood Estimator

From observations $v_1,\ldots,v_N$, the *Likelihood Function* is:

$$
\begin{aligned}
L(\alpha,\sigma) &= g(v_1,\ldots,v_N,\alpha,\sigma) \\
&= \prod_{n=1}^{N} h(v_n | v_{n-1}, v_0, \alpha, \sigma) \\
&= (2\pi s(\alpha,\sigma))^{-n/2} \exp\left(-\sum_{n=1}^{N} \left(\frac{v_n - e^{-\alpha\Delta} v_{n-1}}{2s(\alpha,\sigma)}\right)^2\right)
\end{aligned}
$$

where $s(\alpha,\sigma) = \sigma^2 \dfrac{1-e^{-2\alpha\Delta}}{2\alpha}$

The standard Maximum Likelihood method minimizes the Likelihood Function to get the **best estimator of *α* and σ** and their **variance** (i.e., **normal distributions for α and σ**).

I lied a little bit: need to use Kalman filter with ML since one does NOT observe velocities, rather positions.

Fricks, Yao, Elston, Forest SIAP 09 has a "tutorial" for the unitiated

Viscoelastic diffusion: Generalized Langevin Equations (GLEs)
Zwanzig & Bixon; Mason & Weitz re-introduced in 1996

"Brownian" particles in a viscoelastic medium are modeled by:

$$\frac{dV(t)}{dt} = -\int_0^t \xi(t-s)V(s)ds + \sqrt{\frac{k_B T}{m}}F(t)$$

where the random force *F(t),* by virtue of the **fluctuation-dissipation** theorem, is colored consistent with the memory kernel *ξ(t):*

$$\langle F(t)F(s)\rangle = \xi(t-s)$$

Query: can we choose a class of kernels which are sufficiently general

1. to capture wide range of observed scaling behavior?
2. to be amenable to Maximum Likelihood inference methods from path data?
3. to forecast passage time distributions vs. layer thickness?

Note: these kernels are due to the microscopic physics of the complex fluid & fluid-particle surface chemistry, which you will not understand anytime soon in mucus!

Exponential memory kernels

If we posit an exponential memory kernel,

$$\xi(t) = ce^{-t/\lambda}, c = \frac{6a\pi G}{m}$$

then **the colored noise *F(t)***, satisfying $\langle F(t)F(s)\rangle = k_B T\xi(t-s)$

is equivalent to an Ornstein-Uhlenbeck process

$$\frac{dF(t)}{dt} = -\frac{1}{\lambda}F(t) + \sqrt{\frac{2c}{\lambda}}W(t),$$

where *W(t)* is white noise. That's suggestive….in fact the next step was done by different folks for a single exponential over the years….**an old idea of Mori to embed non-Markovian processes into higher dimensional Markovian systems.**

Exploit convolution with exponentials to reformulate GLEs

By introducing $$Z(t) = \int_0^t e^{-(t-s)/\lambda} V(s) ds$$

The GLE can be recast as a **vector Ornstein-Uhlenbeck process**

$$\frac{d}{dt} Y(t) = AY + KW \qquad Y(t) = (X(t), V(t), Z(t), F(t))^T$$

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & -c & \sqrt{\frac{k_B T}{m}} \\ 0 & 1 & -\frac{1}{\lambda} & 0 \\ 0 & 0 & 0 & -\frac{1}{\lambda} \end{pmatrix}, \qquad K = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sqrt{\frac{2c}{\lambda}} \end{pmatrix}.$$

Now: all features of the viscous Langevin ODE extend to the GLE, for both Discrete auto-regressive processes (generation of paths) & Maximum Likelihood / Kalman filter (parameter estimation).

But…..complex fluids like mucus have a wide colored noise spectrum, not one color!

Memory Kernels as Prony series

SIAP 2009, Fricks, Yao, Elston, GF

There isn't one timescale of memory --- hundreds to thousands in mucus.

We approximate the memory function as an exponential series

$$\xi(t) = c_1 e^{-t/\lambda_1} + c_2 e^{-t/\lambda_2} + \cdots + c_N e^{-t/\lambda_N},$$

then define intermediate variables and forces

$$Z_j(t) = \int_0^t e^{-(t-s)/\lambda_j} V(s) ds$$

$$\frac{dF_j(t)}{dt} = -\frac{1}{\lambda_j} F_j(t) + \sqrt{\frac{2c_j}{\lambda_j}} W(t),$$

which yields an O(N) vector OU process representation for the GLE system:

$$\frac{d}{dt} Y(t) = AY + KW$$

Ito quadrature again yields discrete AutoRegressive Process for generating paths & MSD, modulo a linear algebra hurdle

The solution is again of the form

$$Y(t) = e^{At}Y(0) + \int_0^t e^{A(t-s)}KW(s)ds$$

So the exact discrete statistical representation is:

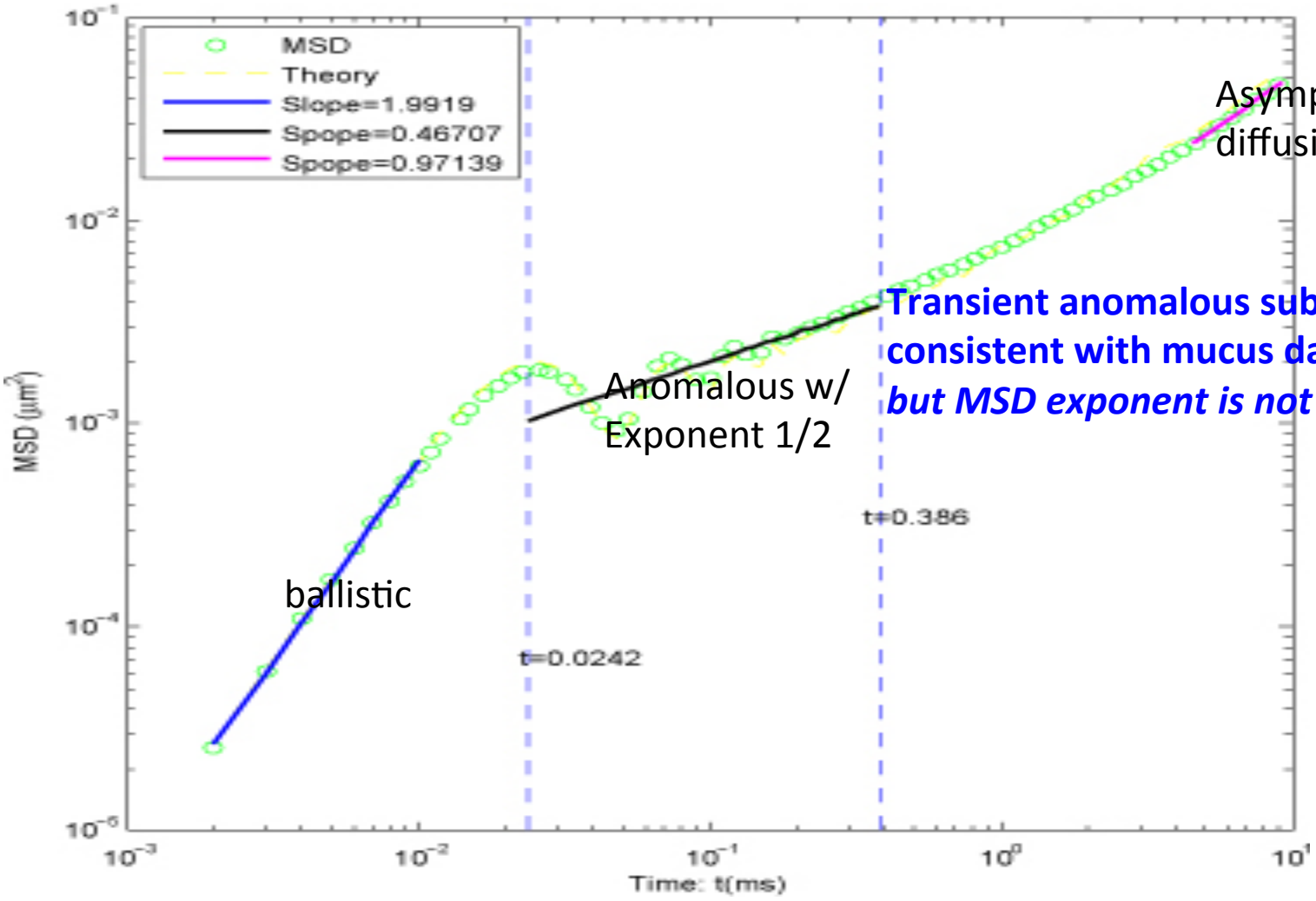
$$Y_n = e^{A\Delta}Y_{n-1} + \varepsilon_n, \text{ and } Y_n = Y(n\Delta)$$

with covariance matrix for ε_n defined as

$$S(\Delta) = e^{A\Delta}\left(\int_0^\Delta e^{-As}KK^T e^{-A^T s}ds\right)e^{A^T\Delta}$$

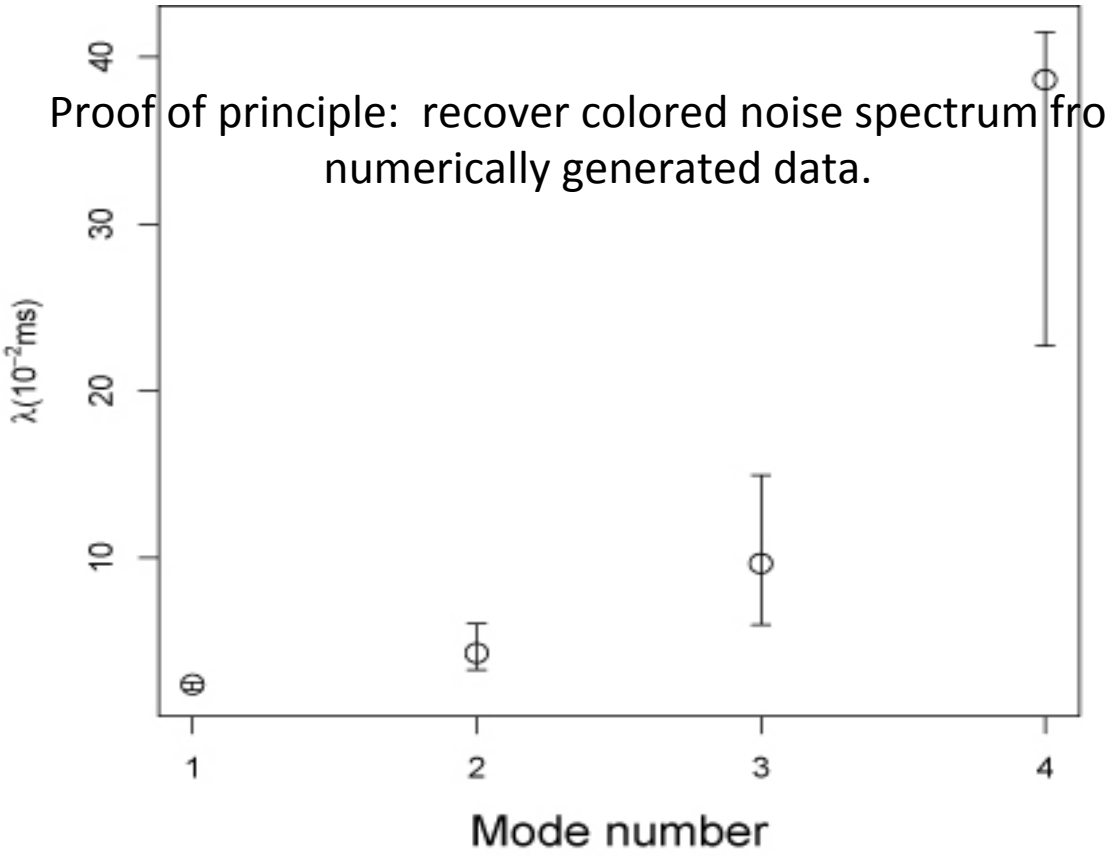
If you can handle the linear algebra, you have a method of generating paths and explicit formulas for statistical properties such as MSD. See SIAP 09 paper and Lingxing Yao's thesis. Direct paths & MSD generation is extremely accurate!

Rouse model: ½ power law exponent in MSD



Transient anomalous sub-diffusion consistent with mucus data,
but MSD exponent is not tunable

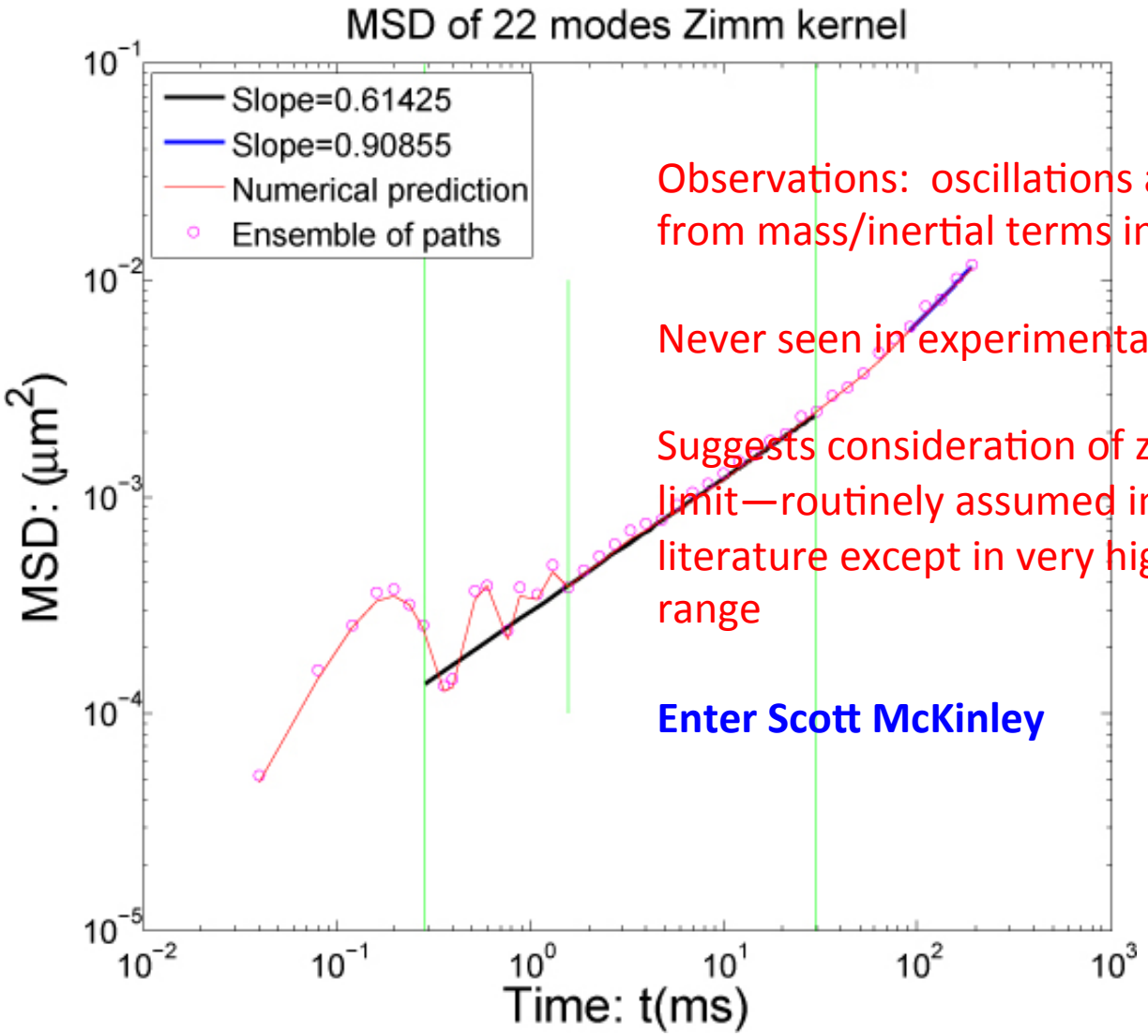
Maximum Likelihood results



These methods are:

- **not limited by the dimension of colored noise spectrum for direct paths & statistics**
- **definitely limited in # of colors of noise for inversion—need a new idea**
- **limited in the ability to fit or tune observed MSD power law behavior** aside from kernels of Rouse and Zimm that yield exactly/only two exponents in (0,1)

Zimm model: 2/3 exponent



Observations: oscillations are coming from mass/inertial terms in GLE

Never seen in experimental data

Suggests consideration of zero mass limit—routinely assumed in experimental literature except in very high frequency range

Enter Scott McKinley

A new family of stochastic processes with "tunable" MSD transient power law scaling, that are amenable to "inference" methods from noisy time series data, and that can be simulated accurately and fast. Two fundamental theorems.

McKinley, Yao, Forest J. Rheology, Nov/Dec 2009 issue

Each color of noise in the data, or memory timescale in the particle diffusion, obeys

$$dF_k(t) = -\frac{1}{\lambda_k} F_k(t)dt + \frac{\sqrt{6\pi a \eta_k k_B T}}{\lambda_k} dW_k(t)$$

upon transforming the GLE (~ represents the Laplace transform) we have a standard looking representation in Laplace space

$$\tilde{X}(z) = \frac{\sum_{k=1}^{N} \frac{\sqrt{6\pi a k_B T \eta_k}}{(z+\lambda_k^{-1})\lambda_k} \tilde{W}_k(z)}{mz + \sum_{k=1}^{N} \frac{6\pi a}{z+\lambda_k^{-1}} \frac{\eta_k}{\lambda_k}}$$

Exact (closed-form) solution of GLEs in the zero mass limit for arbitrary Prony series kernels.

Explicit particle path representation

Equation:

$$m\frac{dV(t)}{dt} = -\int_0^t \xi(t-\tau)V(t) + \sqrt{\frac{k_bT}{m}}F(t) \qquad \xi(t) = \beta\sum_{n=1}^{N} e^{-\alpha_k t}$$

Colored noise in the equation

$$p(x) = \prod_{k=1}^{N}(x+\alpha_k)$$

Interpolant of colored noise between GLE & paths

$$dZ_j(t) = -r_j Z(t)dt + \sqrt{2k_BT}\,dB_j(t)$$

$$x(t) = C_0 B_0(t) + \sum_{j=1}^{N-1} C_j Z_j \qquad C_0 = \sqrt{\frac{2k_bT}{\beta N\bar{\tau}}}$$

$$\{-r_j\}_{i=1}^{N-1}$$

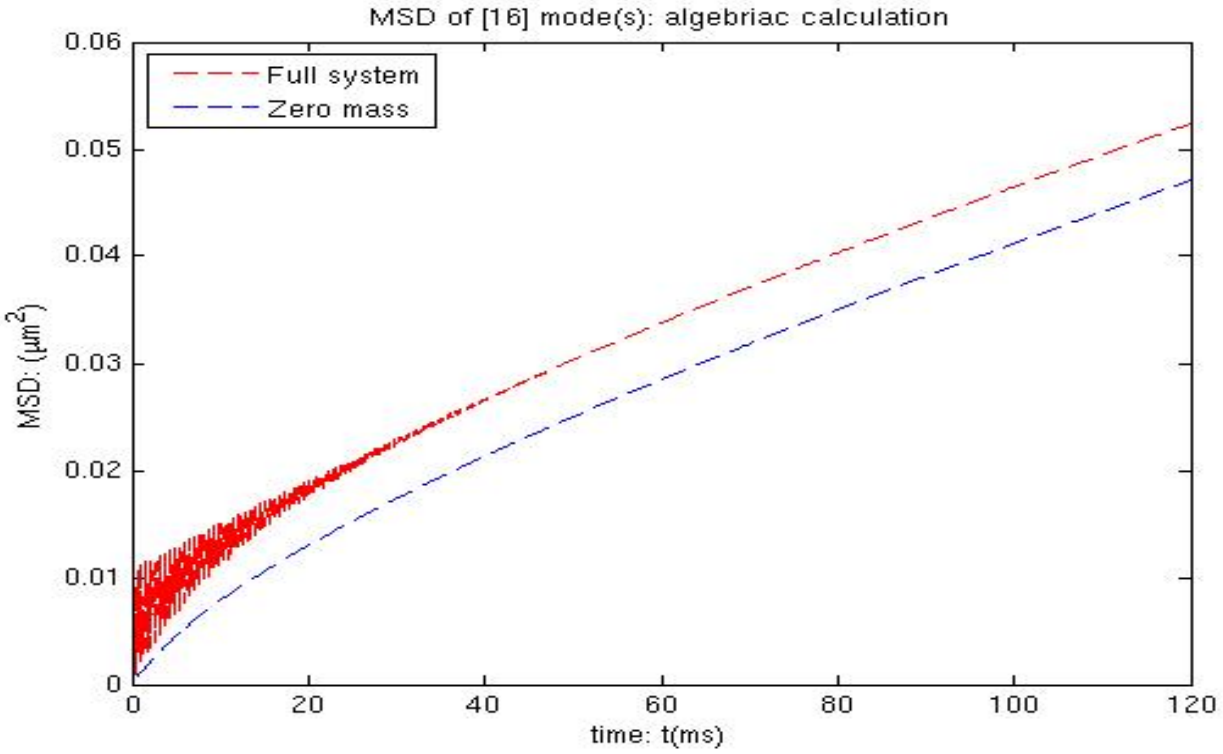
Are the roots of the polynomial p'(x)

Exact GLE solution: a Brownian motion + (N-1) Ornstein-Uhlenbeck processes w/ explicit color spectrum interpolated from the diffusive spectrum

BUT: STILL NO CONTROL OVER MSD SCALING

Upshot: the zero mass GLE is integrable for any Prony series kernel with uniform weights!

Even faster direct simulation of zero-mass paths. Check against inertial AR code for Rouse and Zimm that the MSD power law is right but …… the zero-mass approximation is clearly missing oscillations (knew that) + "shift factor" in MSD of direct GLE simulations ---bad news for FPT predictions!



Classical signature of a singular limit: ghost of mass survives in MSD even though it is apparently negligible. Singular correction provided in JOR paper.

A 3-parameter family of Prony series kernels yields: Transient anomalous diffusion process with tunable: 1. sub-diffusive exponent, 2. window of anomalous scaling and 3. with **longtime diffusive scaling**

Rouse spectrum
Kremer-Grest, Doi-Edwards

$$\alpha_{k,N} \sim \sin^2\left(\frac{k\pi}{2N}\right) \quad \langle X^2(t)\rangle \sim t^{\nu}, \quad \text{with } \nu = \begin{cases} 2 & t \le \tau_1 \\ 1/2 & \tau_1 \le t \le \tau_N \\ 1 & t \ge \tau_N \end{cases}$$

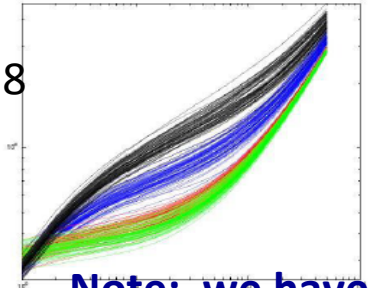
Generalization

$$\tau_k^{-1} = \alpha_k \sim \left(\frac{k}{N}\right)^p \tau_0^{-1} \quad \langle x^2(t)\rangle \sim t^{\nu}, \quad \text{with } \nu = \begin{cases} 1 & t \le \tau_1 \\ 1-\frac{1}{p} & \tau_1 \le t \le \tau_N \\ 1 & t \ge \tau_1 \end{cases}$$

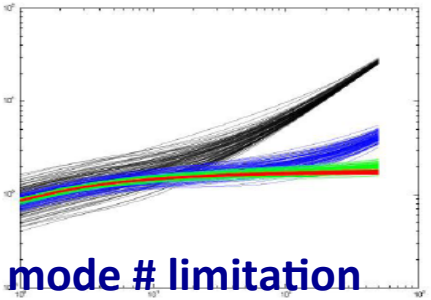
Upshot: scaling behavior of the diffusive spectrum prescribes anomalous exponent

Robustness (McKinley): anomalous diffusive scaling is robust to random perturbations of the uniform weight distribution; *as N >> 1, the weights essentially don't matter, and the window of anomalous scaling becomes arbitrarily long* (**MSD vs time on log-log plot**)

p=3/2, 2, 4, 8
N=200

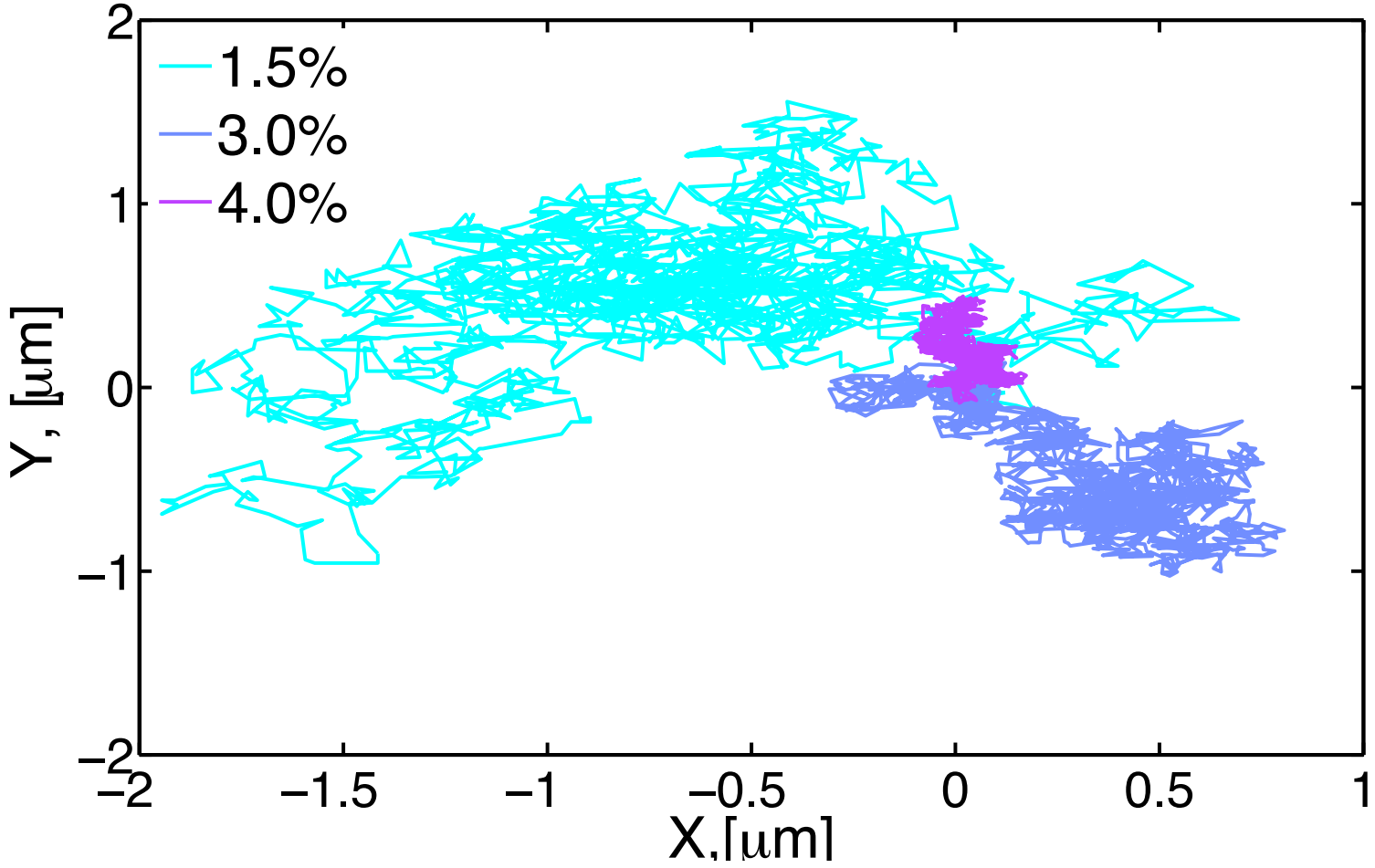


p=2
N=20, 200, 2000
20,000



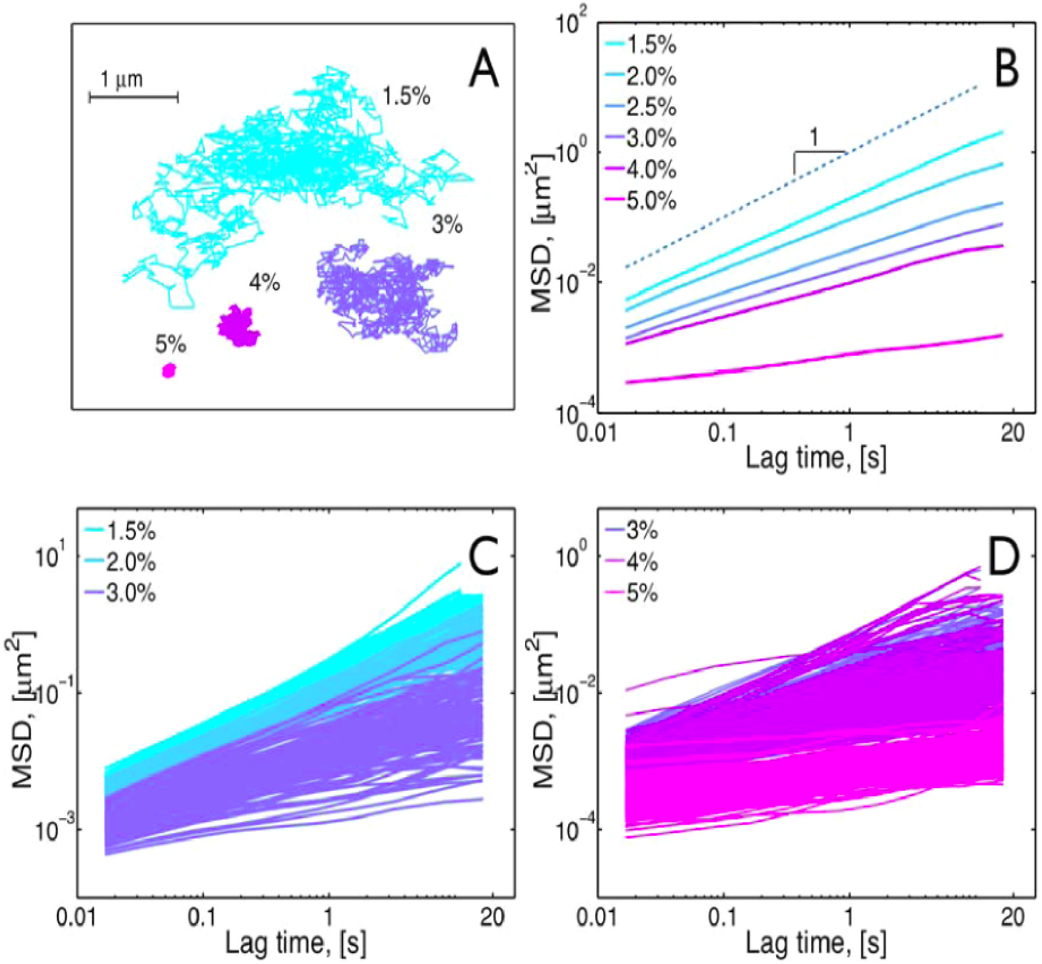
Note: we have thereby also lifted the mode # limitation

Armed with tunable GLEs, return to particle path data in HBE culture mucus vs concentration



Next: data analysis on many particle time series vs concentration

PLoS ONE, Feb 2014, by our mucus diffusion group – Scott et al.



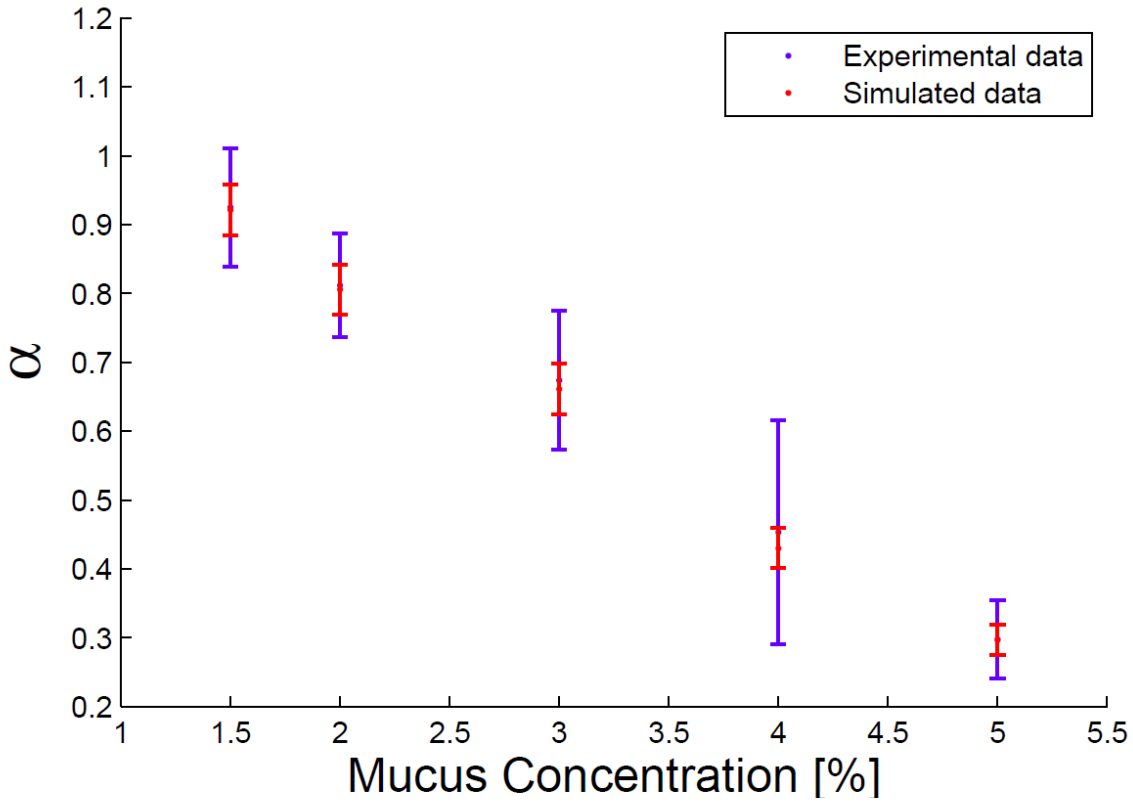
Upshot: over experimental timescales, MSD exhibits uniform sub-diffusive scaling with progressively lower exponents as wt % increases

Figure 2. Diffusivity properties of HBE mucus. A) Particle trajectories of 1 µm diameter particles for four concentrations over 30 s. **B**) Ensemble-averaged MSD versus lag time for different mucus solids concentrations. The dashed line represents a viscous fluid; any smaller slope indicates sub-diffusive scaling. **C**) Individual or path-wise MSD (iMSD) for particles embedded in mucus samples color-coded by solids concentration, for 1.5, 2.0, 3.0 *wt%*. **D**) iMSD for particles embedded in mucus samples color-coded by solids concentration, for 3.0, 4.0, 5.0 *wt%*. *Note the vertical scale disparity with Figure* **2C**.

doi:10.1371/journal.pone.0087681.g002

Remarkably: ***the power law exponent from fBm or GLEs linearly correlates with mucus wt%*** **---** Strong hint of what we were looking for viz a viz clinical objectives: a biomarker for disease assessment and progression PloS ONE Feb 2014

Subject of next lecture by Scott is the model selection and inference problem that must be solved if we are to make predictions beyond experimental timescales



Implications of this robust scaling of MSD with mucus wt%

- Scott's lecture based on work w/ Natesh Pillai, Martin Lysy, model testing and goodness of fit → **forecasting**
- Preliminary numerical results w/ J Mellnik, S McKinley for **scaling of passage times vs mucus layer thickness**
- Mason-Weitz-Crocker et al. protocols turn **MSD statistics → linear viscoelastic moduli** --**viscous and elastic moduli over physiological frequency range from tidal breathing to cilia.**
- **Yields a baseline for flow transport estimation, just like HBE culture mucus is a baseline for diffusive transport estimation**
- **Both estimates of transport timescales are required for judging the race between diffusion and flow.**
- **If we have some more time left, the next relevant issue is …..**

Next major hurdle is “predictive heterogeneity metrics” --- beyond “the ensemble step size distribution function fails a Gaussianity test”

For each path, compute SD of step size distribution in each coordinate, giving (s_x,s_y)_i for path I

Compute Euclidean distances between all “points” = paths

Cluster based on sliding resolution parameter, either coarse to fine or fine to coarse (shown here)

Fine resolution parameter all paths are their own cluster, then successively bin them as resolution coarsens: dendogram

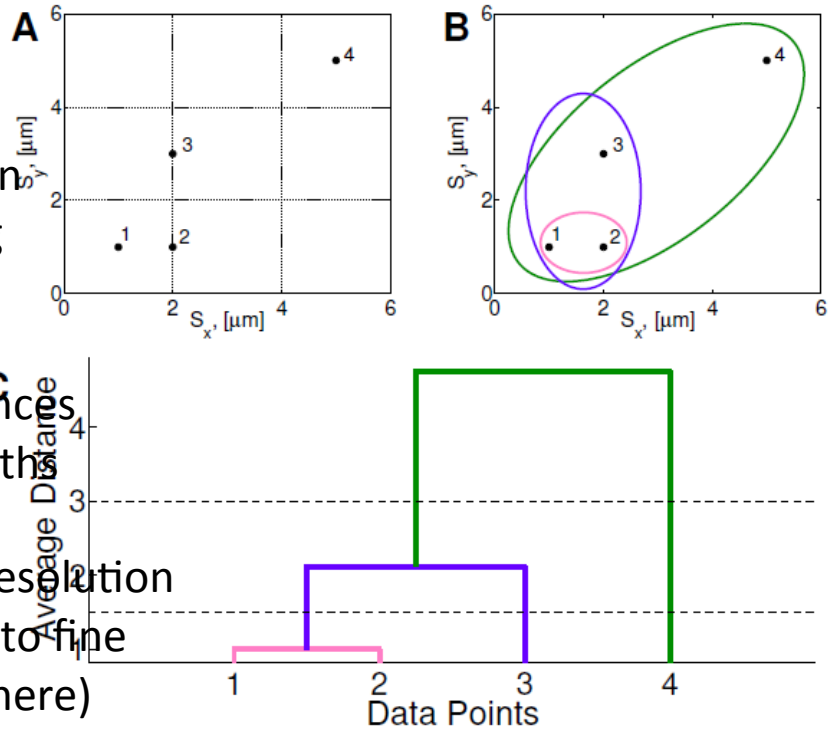


Fig. 1 Example of hierarchical clustering. A) The distribution of data points to be clustered. Each data point is assigned to a cluster containing only itself. The pairwise distances between all clusters are calculated and the closest two clusters are merged to form a new cluster. **B)** This process is repeated until all data points are in a single cluster. **C)** A dendrogram shows the distances between each cluster and the order in which they were merged. The solid lines at 3 and 1.5 show cutoff values that produce two and three clusters, respectively.

So, depending on the resolution you choose, 4, 3, 2, or 1 clusters.

Which # of clusters is more likely given the data?

After all the distances are calculated (Figure 1C), the *number of clusters*, K_τ, is determined by a cutoff value ζ that partitions the dendrogram at resolution ζ. For instance, if we choose any $\zeta < 1$ in Figure 1, all particles remain in their own cluster, and there are 4 clusters at this resolution. For any $1 < \zeta < 2.12$, say $\zeta = 1.5$ as in Figure 1C, the two points making up the pink cluster are now indistinguishable. Thus we declare 3 clusters for this range of ζ. Next, for $2.12 < \zeta < 4.75$, there are only 2 clusters, the blue cluster and point 4, as shown in the figure for $\zeta = 3$. Finally, for $\zeta > 4.75$, there is one cluster with that chosen degree of resolution, the green cluster containing all points. In this way, the parameter ζ solely determines the partitioning of the data, and as ζ varies from the smallest to largest values, the number of clusters K_τ spans 1 to N, where N is the number of observed particles. The next critical step is to select the degree of resolution, i.e. the value of ζ, and thus to determine the number of clusters K_τ that best delineates the ensemble of paths *at lag time* τ.

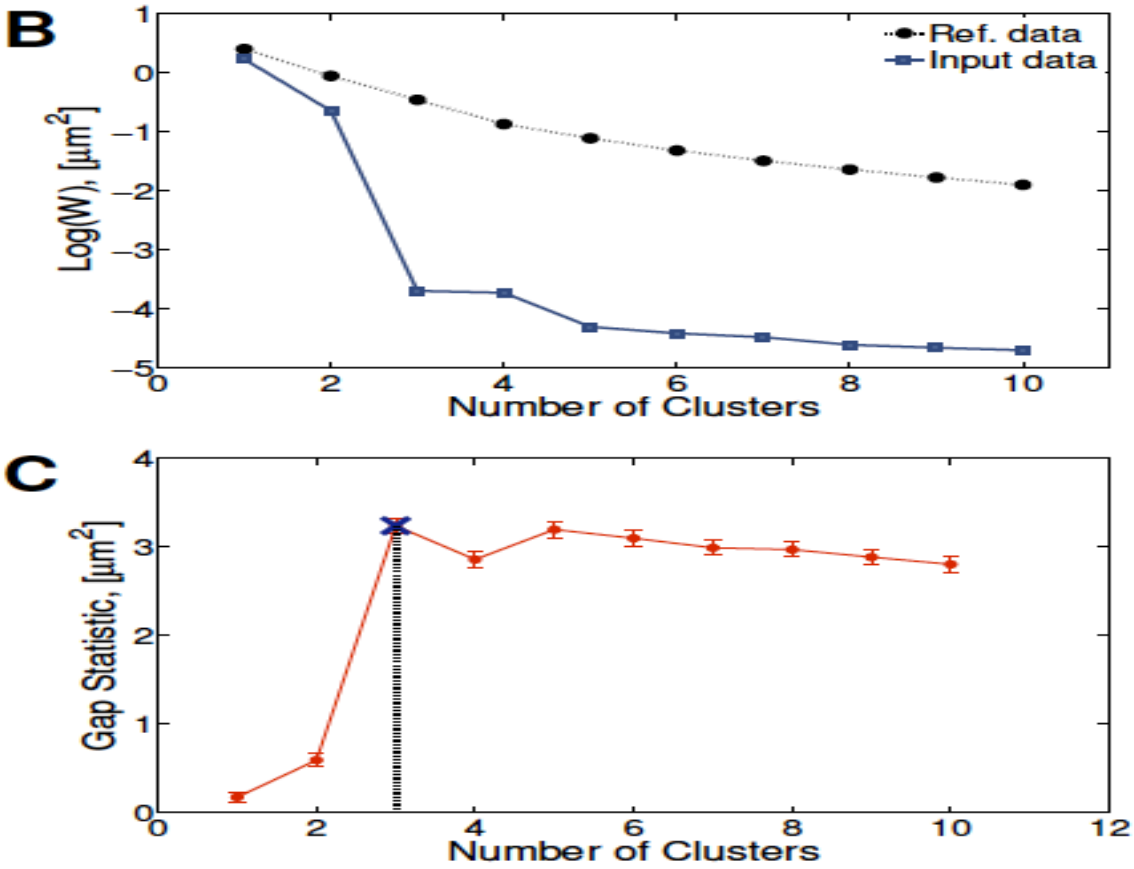
We use the gap statistic to determine -- which "assignment" of # of clusters and which paths belong to each cluster -- is most likely given the data

$$W_K(K_\tau) = \sum_{c=1}^{K_\tau} \frac{1}{2n_c} D_c$$

the sum of the "mean squared spread per cluster"
is computed for each level in the dendogram and then (next)
these data are compared to 100 realizations
of null reference data generated from a uniform distribution

where n_c is the number of elements in cluster c and D_c is the sum of the pairwise squared distances between all the elements of cluster c. As ζ decreases, the number of clusters, K_τ, increases causing W_K to decrease due to the increasing mean intra-cluster density. Next, we use these values of W_K to compare the distribution of standard deviations of the van Hove functions, $s_x(\tau)$ and $s_y(\tau)$, which may or may not contain statistically distinct clusters, to a null reference data set containing only one cluster and with uniform density. In order to ensure that the null reference data set only contains a single cluster with uniform density, it is generated from a uni-modal uniform distribution. To match the input data as closely as possible (apart from the number of clusters present), the reference data set is created such that its cardinality and domain is the same as the input data, i.e. the distribution of $(s_x^j(\tau), s_y^j(\tau))$. To remove the variability and arbitrariness associated with the comparison of the input data to a single reference data set, it is common practice to compare the input data to multiple of reference data sets. We have determined that 100 reference data sets suffices to consistently partition the data.

The gap statistic computes the trends in this comparison between log(W_k) of the data and log(W_k) of the uniform reference data versus the resolution parameter that chooses the number of clusters, starting from 1 cluster and low resolution and then increasing across the # of clusters. The "winner" is then chosen to be the degree of resolution / # of clusters with the largest jump from the lower resolution / lower # of clusters. Not perfect, but it is deterministic.



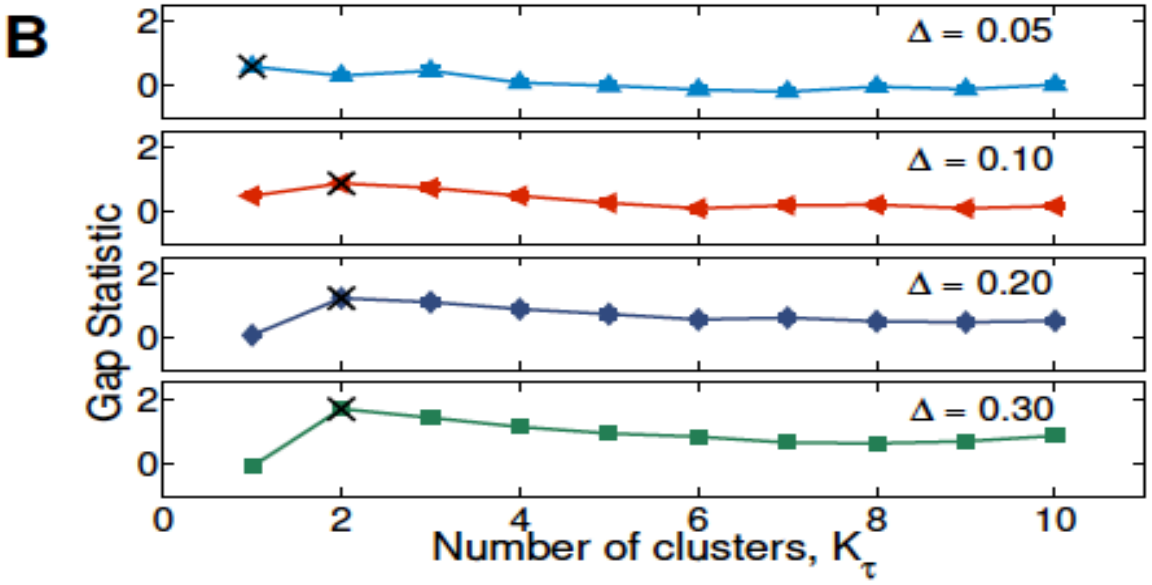


Fig. 3 Test of the gap statistic test: numerical data with controlled degrees of heterogeneity in the diffusion coefficients. Four heterogeneous Newtonian data sets are generated, where each data set consists of 100 paths of particles of diameter 1 μm. For each data set, the first 50 paths have diffusion coefficient $D_1 = 3.10\mu m^2/s$ while the next 50 paths have diffusion coefficient $D_\Delta = D_1(1+\Delta)$ for $\Delta = 0.05, 0.10, 0.20, 0.30$. **A)** As Δ increases, the 'bend' in the $\log(W)$ vs. K_τ plot at $K_\tau = 2$ becomes more pronounced. **B)** The gap statistic correctly indicates two clusters for $\Delta \geq 0.10$. The number of clusters selected by the gap statistic is indicated by a black $\times$.

In our submitted paper, we illustrate this algorithm on tunable experimental and numerical data for normal and sub-diffusive processes.

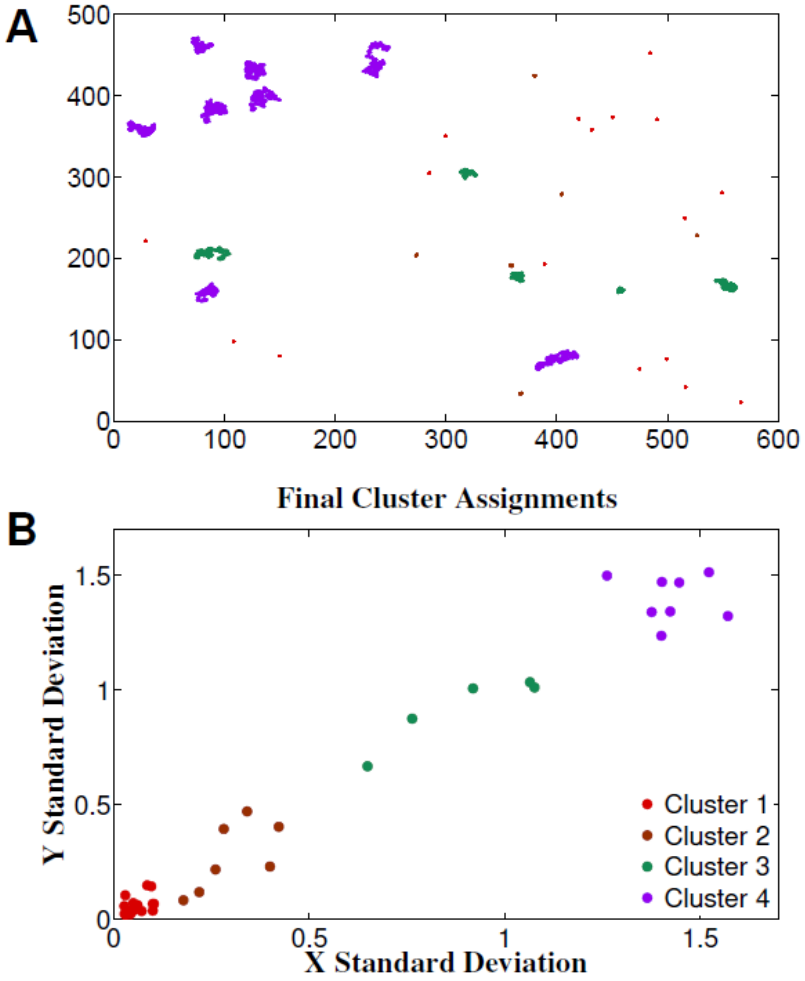
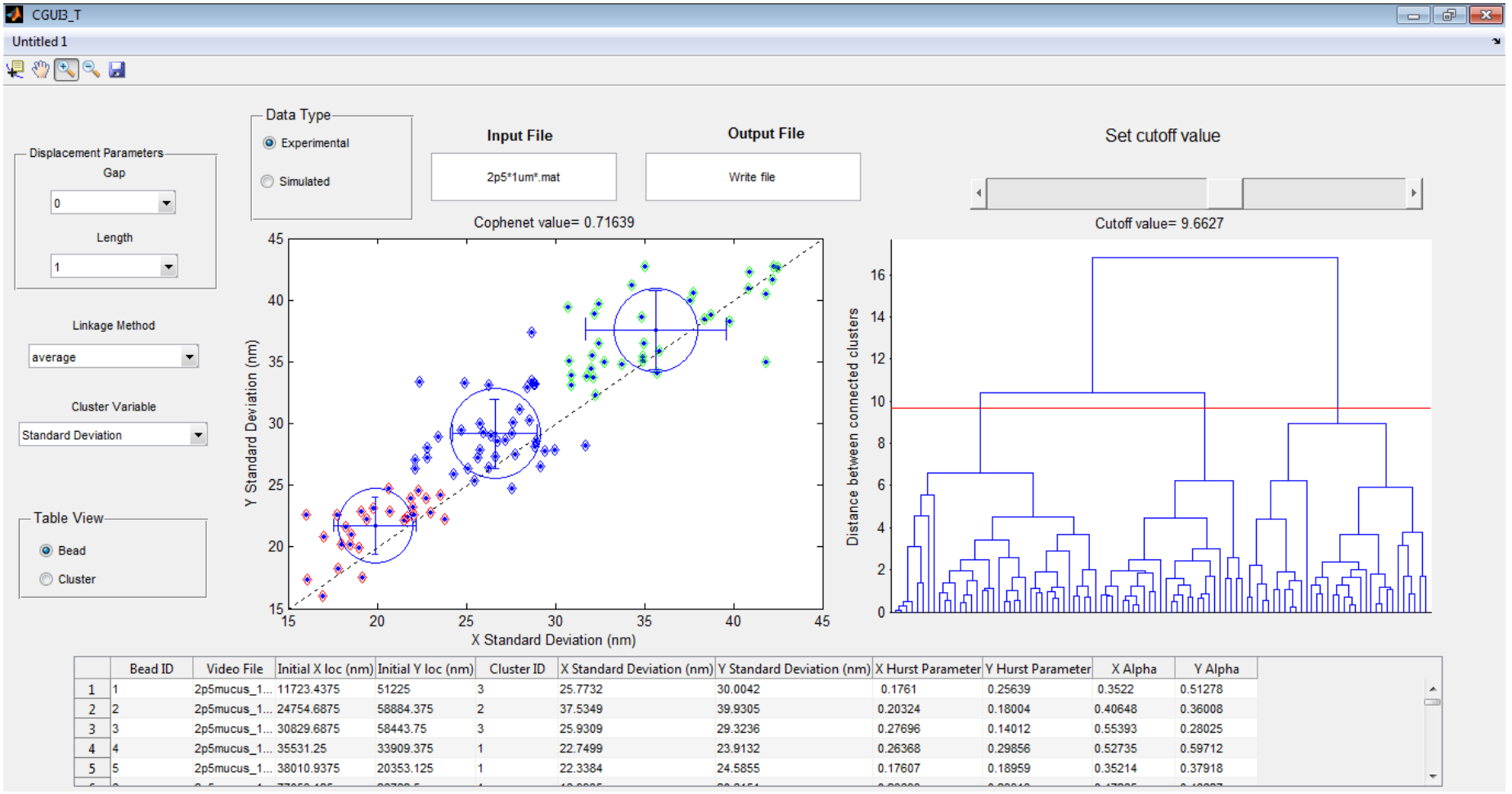


Fig. 14 Experimental Agarose data. 1μm beads diffusing in 0.2% w/w agarose. **A)** Particle paths representing the heterogeneous behavior within the sample. **B)** Results from our clustering algorithm included four clusters. Clusters are color coded in both figures.

Apply metrics to **detect heterogeneity in diffusive paths of 2.5 wt% mucus**
Keeping spatial position in sample gives insight into **lengthscales and degrees of diffusive and viscoelastic heterogeneity.**
Next step: map this spatial heterogeneity to PT distributions….



	Bead ID	Video File	Initial X loc (nm)	Initial Y loc (nm)	Cluster ID	X Standard Deviation (nm)	Y Standard Deviation (nm)	X Hurst Parameter	Y Hurst Parameter	X Alpha	Y Alpha
1	1	2p5mucus_1...	11723.4375	51225	3	25.7732	30.0042	0.1761	0.25639	0.3522	0.51278
2	2	2p5mucus_1...	24754.6875	58884.375	2	37.5349	39.9305	0.20324	0.18004	0.40648	0.36008
3	3	2p5mucus_1...	30829.6875	58443.75	3	25.9309	29.3236	0.27696	0.14012	0.55393	0.28025
4	4	2p5mucus_1...	35531.25	33909.375	1	22.7499	23.9132	0.26368	0.29856	0.52735	0.59712
5	5	2p5mucus_1...	38010.9375	20353.125	1	22.3384	24.5855	0.17607	0.18959	0.35214	0.37918

Potential impact of results to date?

- *Extend tools from culture mucus to clinical samples --- insights into behavioral effects beyond wt%*
- *Mucus wt% + "behavior" poses as a potential clinical marker for lung pathogenesis*
- *Predictive power:*

1. If we can show accurate fits to diffusive models, then we can *simulate passage times of specific agents in specific mucus samples.*
2. If we can show accurate fits to viscoelasticity models, then we can *simulate flow of mucus layers in physiological airways*

I am not sure which project will succeed first. Bets?

Thanks for listening.