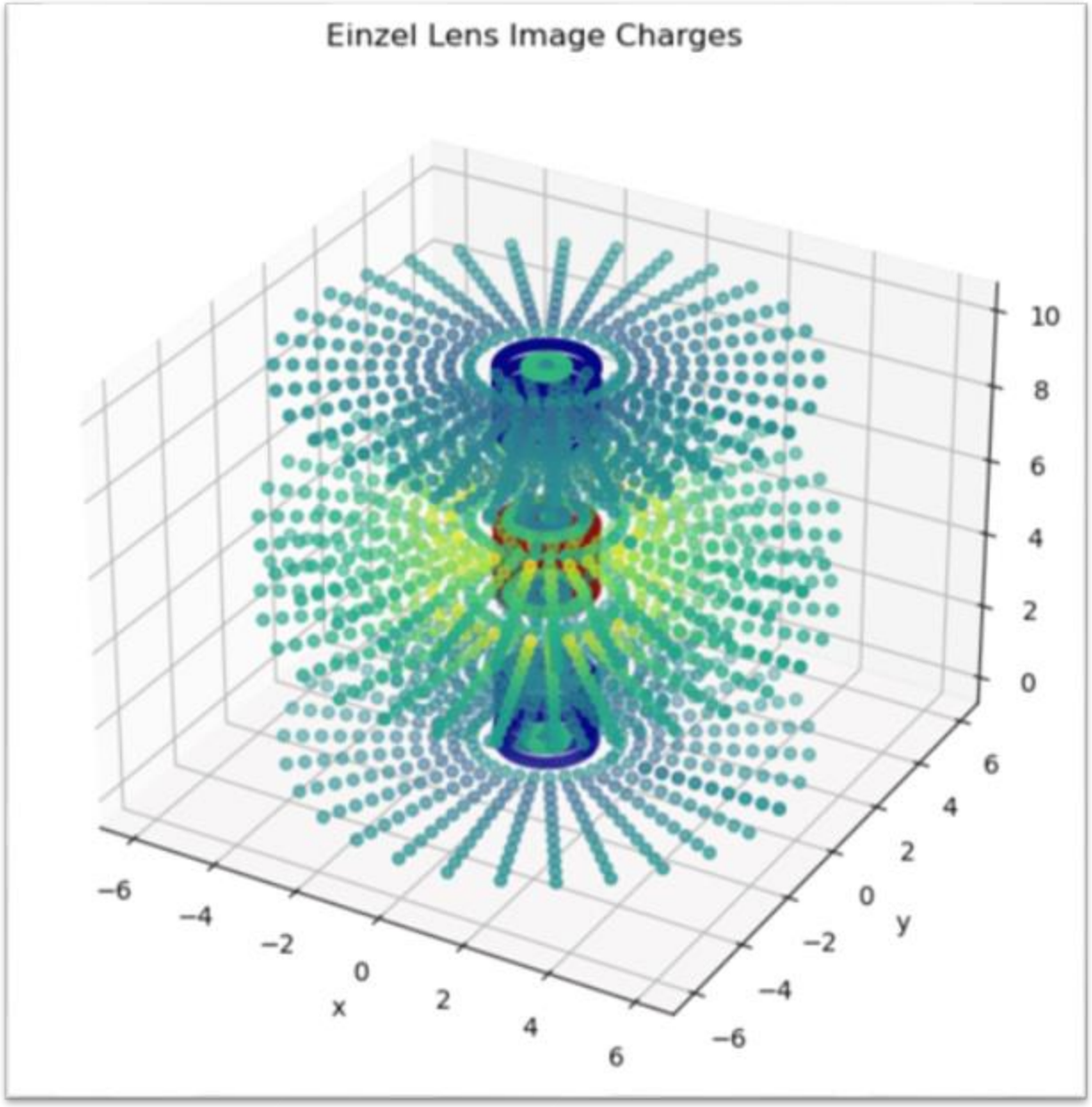
Vast Technical Presentation:
Langevin Models of DNA

Jeremy Welsh-Kavan

Dec. 6th, 2023

My Background: Scientific Computing

- Guenza Lab (University of Oregon)
  - Built models for macromolecules
  - Used High Performance Computing for large simulations
  - Developed data analysis software
- Thermo Fisher Scientific
  - Optimized Focused Ion Beam column designs
  - Used High Performance Computing for many simulations
  - Developed data analysis software
  - Performed experiments on SEM-FIB Dual Beam microscope



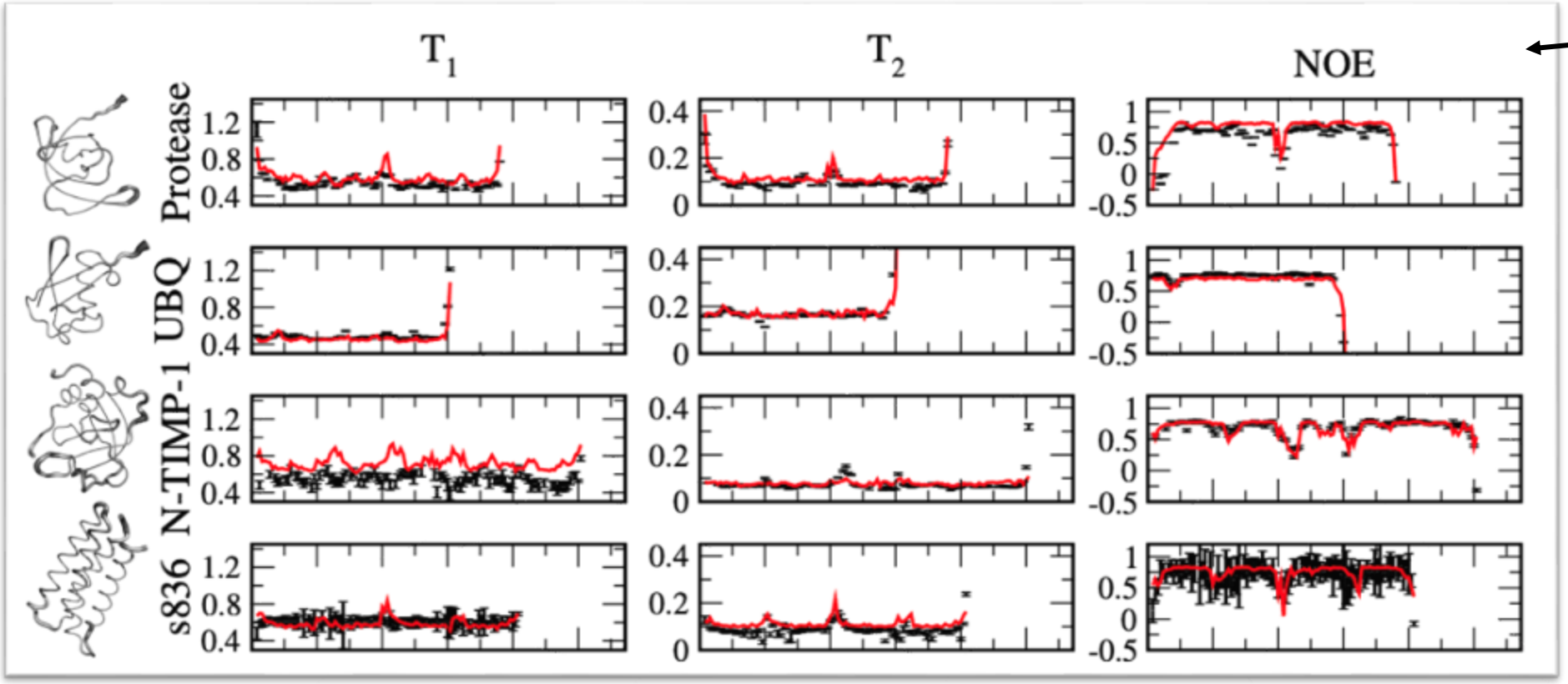
Why Vast?

- Why I'm interested
  - Interesting work – develop artificial gravity for spacecraft
  - High stakes – things have to work in space
  - Startup – things have to get done, lots of opportunities for personal professional growth.
- What I can bring to Vast
  - Problem solving – success in physics requires strong problem solving skills
  - Curiosity – I'm passionate about aerospace and software
  - Programming – I can write good code to perform complex tasks, such as machine control

SR-71 Blackbird. Image from Google

Langevin Models of DNA

- Goal: develop accurate and interpretable descriptions of macromolecules.
- Prior work on proteins supports Langevin model.
- Ongoing!

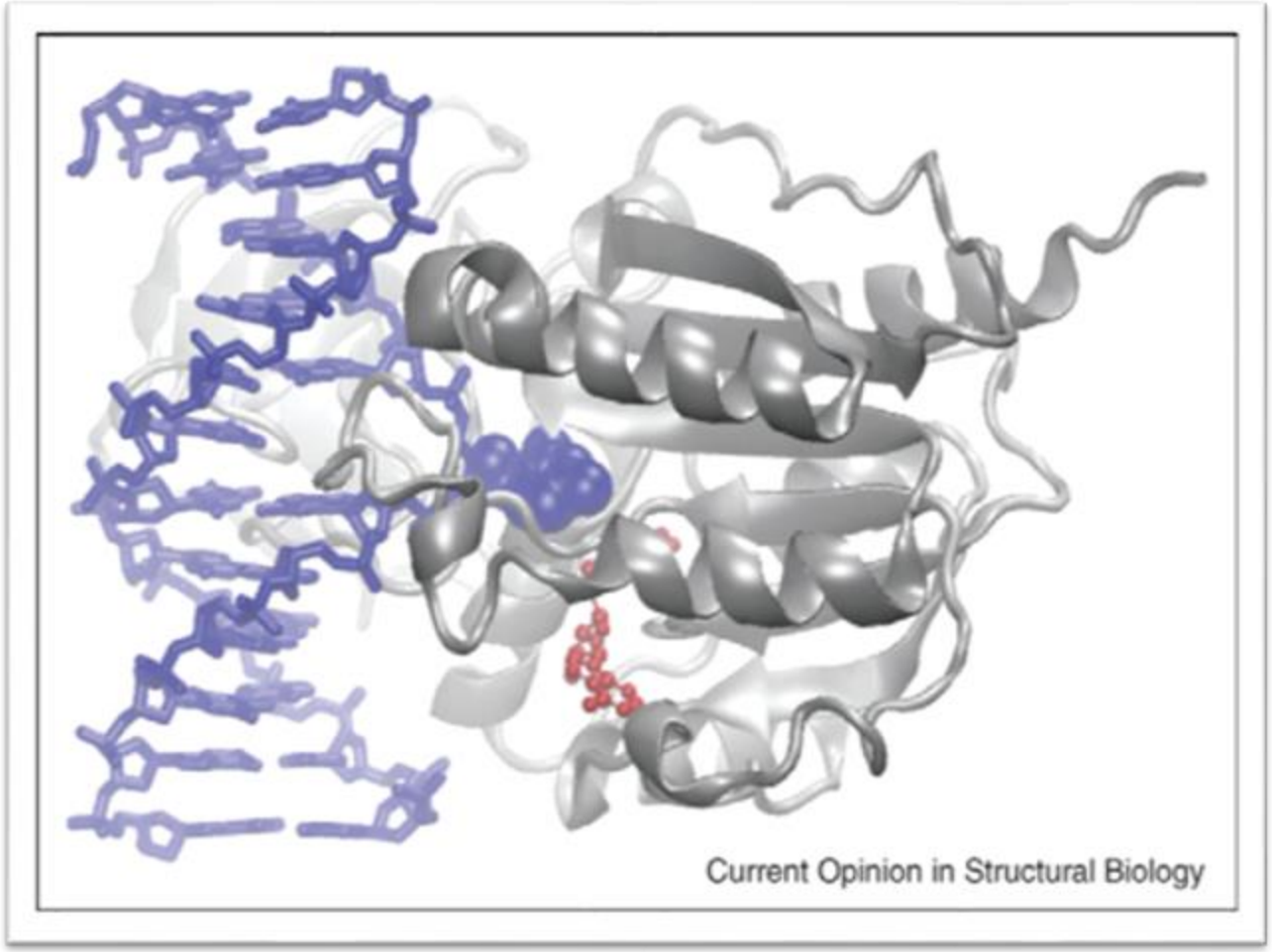


Successful science

Image from *Predicting protein dynamics from structural ensembles,* J. Copperman, M.G. Guenza

Why Study DNA Physics?

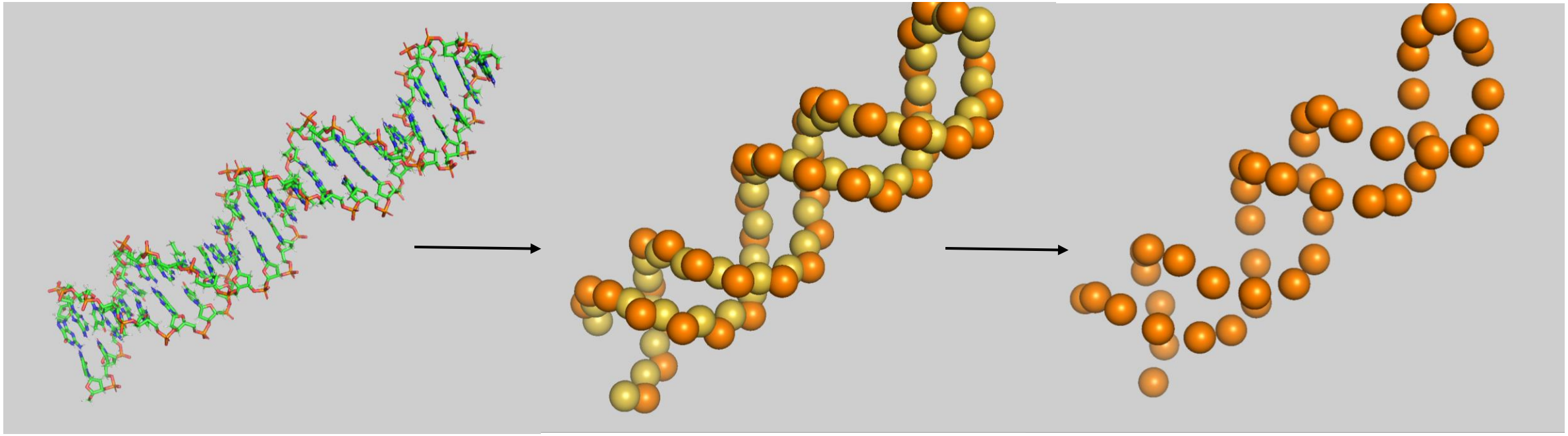
- We know that...
  - Genetic information is encoded in DNA base pairs.
  - DNA must expose its interior to proteins.
  - DNA must undergo conformational changes (e.g. "breathing fluctuations").
- We would like to know...
  - What parts of the molecule are involved
  - How frequently fluctuations occur
  - How rapidly fluctuations occur
- One way to explore DNA physics is with a coarse-grained model.



MacKerell, Jr. and Nilsson, *Current Opinion in Structural Biology,* 2008, **18**:194-199.

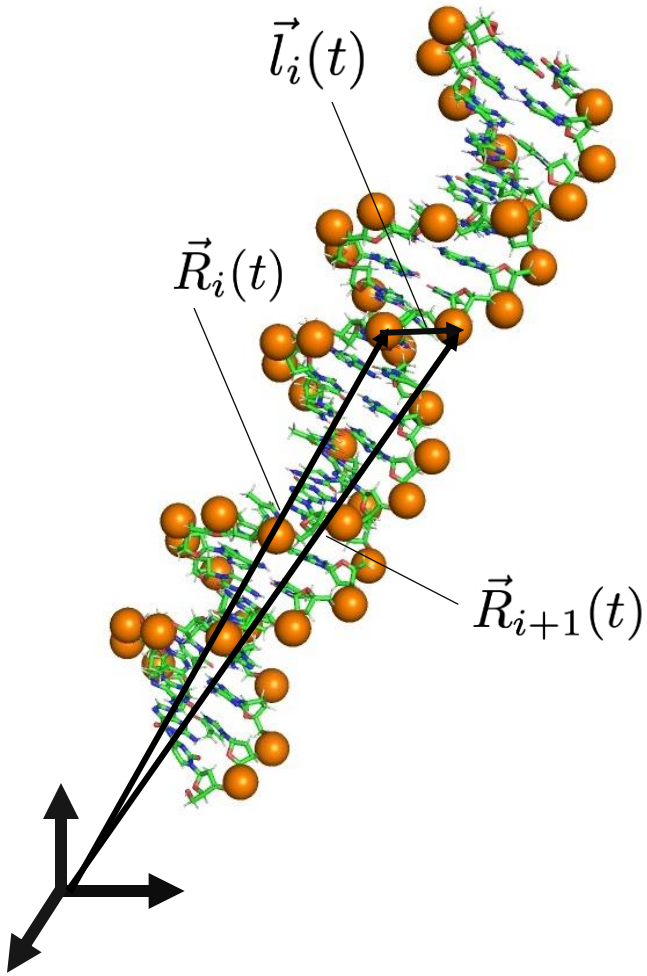
Coarse Graining

- Large molecules can be modeled using molecular dynamics (MD) simulations, but MD simulations
  - are computationally expensive
  - are short (~1 us)
  - Contain extraneous detail
- Large molecules can also be modeled using a coarse-grained (CG) model which
  - ignores unnecessary degrees of freedom
  - Allows longer simulations
  - Directly answers the questions we care about
- One possible model is the Langevin Equation for Protein Dynamics (LE4PD).



Langevin Equation for Protein Dynamics

- The LE4PD model is a network of coarse-grained sites connected by springs.
- The model accounts for:
  - Forces between coarse-grained sites due to chemical bonds
  - Forces between coarse grained sites due to interaction through a solvent
  - Frictional forces internal and external to the molecule
- Solutions to this equation can be used to construct experimentally verifiable quantities.



$$\zeta \frac{\partial \vec{R}_i(t)}{\partial t} = -\frac{3k_B T}{l^2} \sum_{j,k} H_{ij} A_{jk} \vec{R}_k(t) + \vec{F}_i^{\text{random}}(t)$$

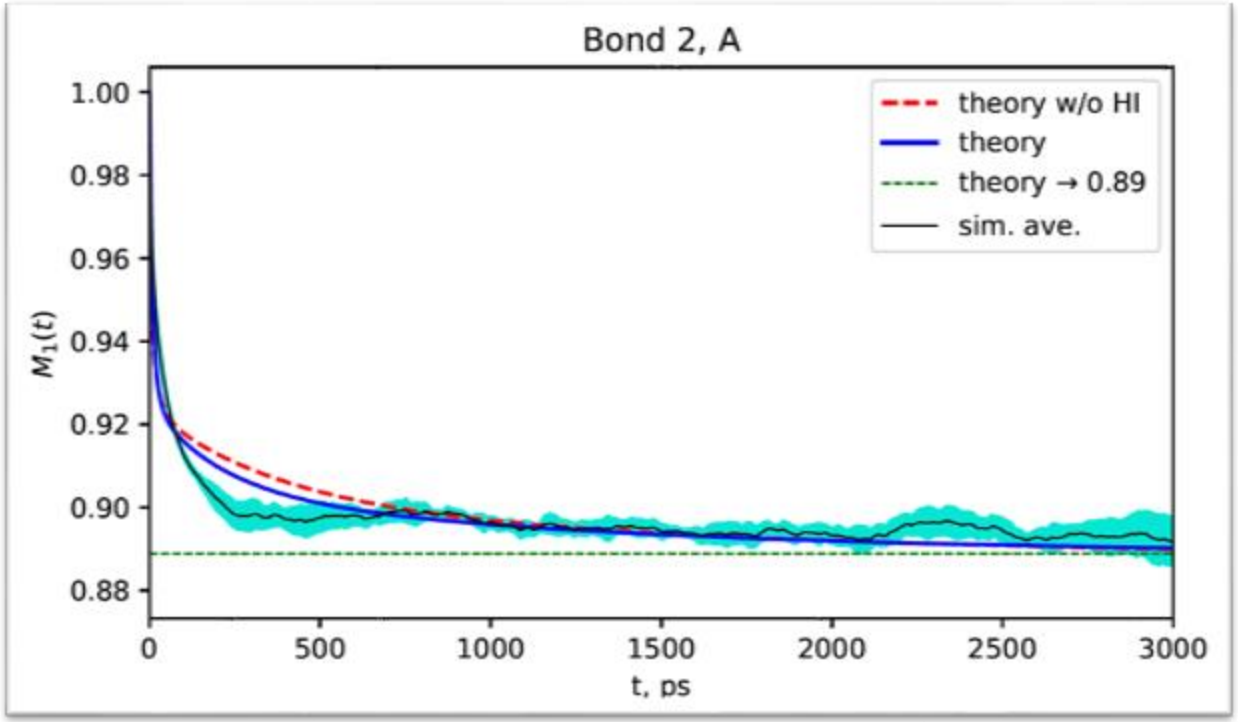
Workflow

1. Find suitable model molecule in the Protein Data Bank (https://www.rcsb.org/).
2. Run molecular dynamics simulations of molecule
3. Use data analysis tools to process simulation trajectory:
   a. Re-format simulation output
   b. Compute friction coefficients
   c. Project trajectory coordinates onto the center-of-mass of each CG site
   d. Compute correlation and hydrodynamic interaction matrices
4. Use Fortran data analysis tools to compute experimental observables (time correlation functions)

```
#volume of sugar+base
Srad_dict = { "DA5" : np.cbrt((3*243.7) / (4*pi)),
              "DA3" : np.cbrt((3*243.7) / (4*pi)),
              "DA"  : np.cbrt((3*243.7) / (4*pi)),
              "DC5" : np.cbrt((3*220.8) / (4*pi)),
              "DC3" : np.cbrt((3*220.8) / (4*pi)),
              "DC"  : np.cbrt((3*220.8) / (4*pi)),
              "DG5" : np.cbrt((3*251.4) / (4*pi)),
              "DG3" : np.cbrt((3*251.4) / (4*pi)),
              "DG"  : np.cbrt((3*251.4) / (4*pi)),
              "DT5" : np.cbrt((3*240.2) / (4*pi)),
              "DT3" : np.cbrt((3*240.2) / (4*pi)),
              "DT"  : np.cbrt((3*240.2) / (4*pi))}

#Calculate the Miller radius per bead
mradlist = [] #radii list
with open(TOP) as f: # PDB file
    for line in f:
        # Skip header
        if line[0:4] != 'ATOM':
            pass
        elif line[0:4] == 'ATOM':
            dummy = line.split()
            if dummy[2] == "P":
                mradlist.append(Prad_dict["P"])
            elif dummy[2] == "S":
                mradlist.append(Srad_dict[dummy[3]])

mradlist = np.array(mradlist)

#file should look like: "./SP_resarea.xvg"
if os.path.exists(path_to_resarea + cgmodel + "_resarea.xvg"):
    pass
else:
    raise FileNotFoundError('''I can't find the resarea.xvg file containing the so
                            Please either run the process.sh file, if you have not
                            file into the current working directory.''')

# To compute the solvent exposed surface area, run the following GROMACS command:
# gmx_mpi sasa -f after_rot_sim0.xtc -s dsAXA.pdb -n SP_index.ndx -or SP_resarea.x
# This assumes that SP_index.ndx contains the atom indices for the sugar+base and
# 2 and 3, respectively.
```

Experimental Validation

- LE4PD solutions yield time correlation functions (TCFs) of bond vectors.
- TCFs
  - characterize the flexibility of regions within the molecule.
  - predict observables from NMR experiments (T1, T2, and NOE relaxation times).

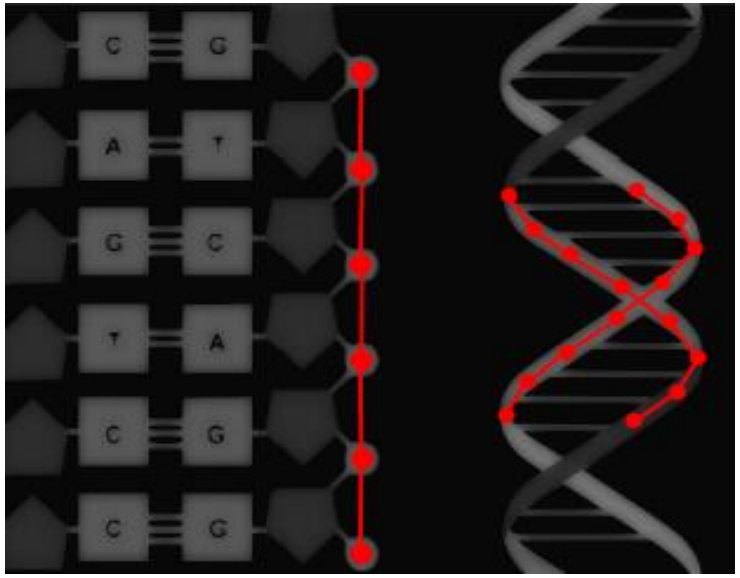


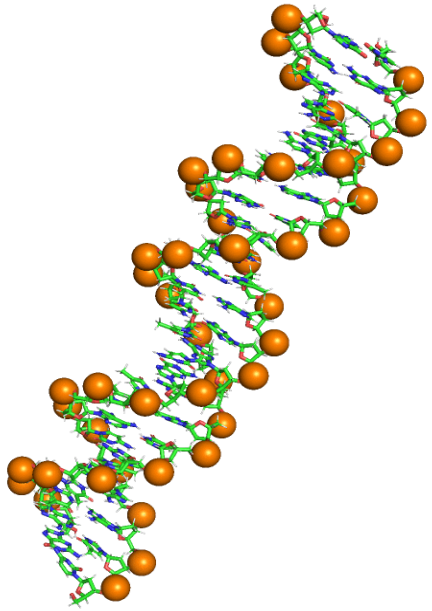
$$M_{1,i}(t) = \frac{\langle \vec{l}_i(t) \cdot \vec{l}_i(0) \rangle}{\langle l_i^2 \rangle}$$

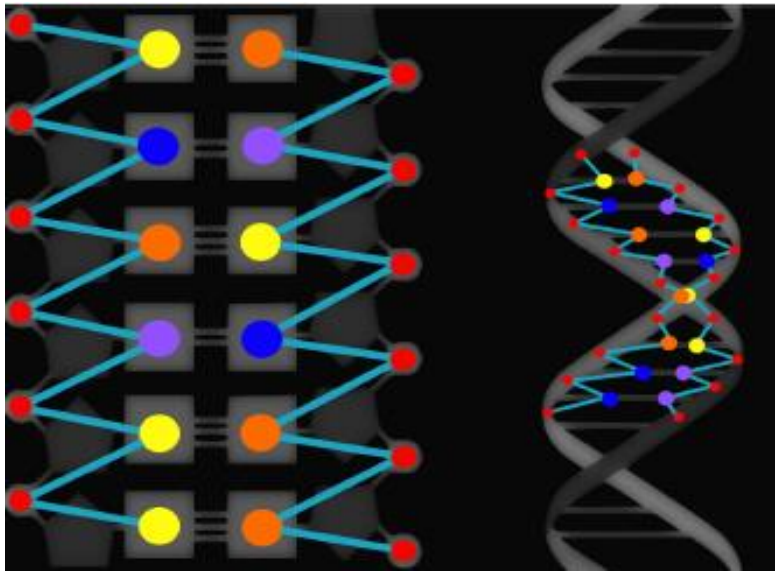
$$M_{1,i}(t) = \sum_{a=1}^{N-1} \frac{Q_{ia}^2}{\langle l_i^2 \rangle} \langle \vec{\xi}_a(t) \cdot \vec{\xi}_a(0) \rangle = \sum_{a=1}^{N-1} A_a^i \exp[-\sigma \lambda_a t]$$

Summary

- Advantages of the LE4PD model:
  - Simple: molecules are networks of masses on springs.
  - Solvable with elementary techniques.
  - Easily compared to statistical models (e.g. PCA).
  - Accurately predicts experimental observables.
  - Admits computationally efficient simulations.







Thank you!

Ideas for Vast

- Use micro-gravity for crop farming
  - Large collections of modular micro-gravity environments could be used for crop farming
  - Micro-gravity could have benefits in crop yield for certain plants
  - Avoids deforestation on Earth
- Artificial gravity to assist in mining the cosmos
  - Artificial gravity environments could be used as centrifuges to collect heavy metals from ore
- Collaborate with physicists to use funding for particle accelerators to create a dual-purpose Ring World + Particle Accelerator.

Supplemental Slides

The LE4PD Model

$$\bar{\zeta}\frac{\partial \vec{R}_i(t)}{\partial t} = -\frac{3k_B T}{l^2}\sum_{j,k} H_{ij}A_{jk}\vec{R}_k(t) + \vec{F}_i^{\text{random}}(t)$$

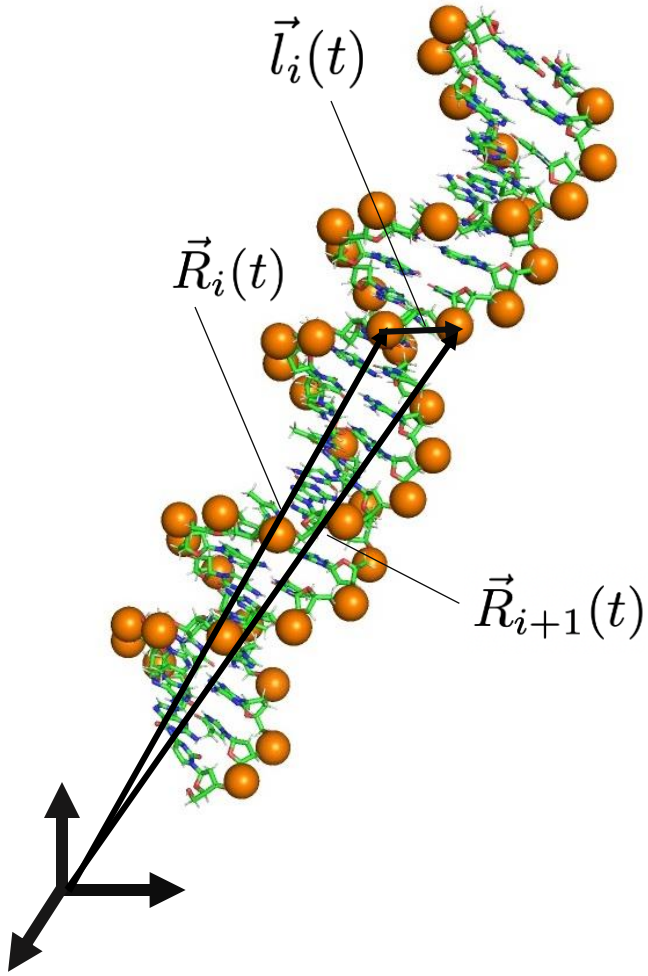
$$\longrightarrow \vec{l}_i = \sum_j M_{ij}\vec{R}_j$$

$$\frac{\partial \vec{l}_i(t)}{\partial t} = -\sigma\sum_{j,k} L_{ij}U_{jk}\vec{l}_k(t) + \vec{v}_i^{\text{random}}(t)$$

$$\mathbf{U}^{-1} = \frac{\left\langle \vec{l}_i \cdot \vec{l}_j \right\rangle}{l^2} \qquad H_{ij} = \frac{\bar{\zeta}}{\zeta_i}\delta_{ij} + (1-\delta_{ij})\bar{r}^{\,w}\left\langle \frac{1}{r_{ij}} \right\rangle$$

$$\mathbf{A} = \mathbf{M}^T\begin{pmatrix} 0 & 0 \\ 0 & \mathbf{U} \end{pmatrix}\mathbf{M} \qquad \zeta_i = 6\pi(\eta_w r_i^w + \eta_p r_i^p)$$

$$\langle \vec{F}_i(t)\cdot\vec{F}_j(t+\Delta t)\rangle = 6k_B T\bar{\zeta}\delta_{ij}\delta(\Delta t)$$



Easy to Solve

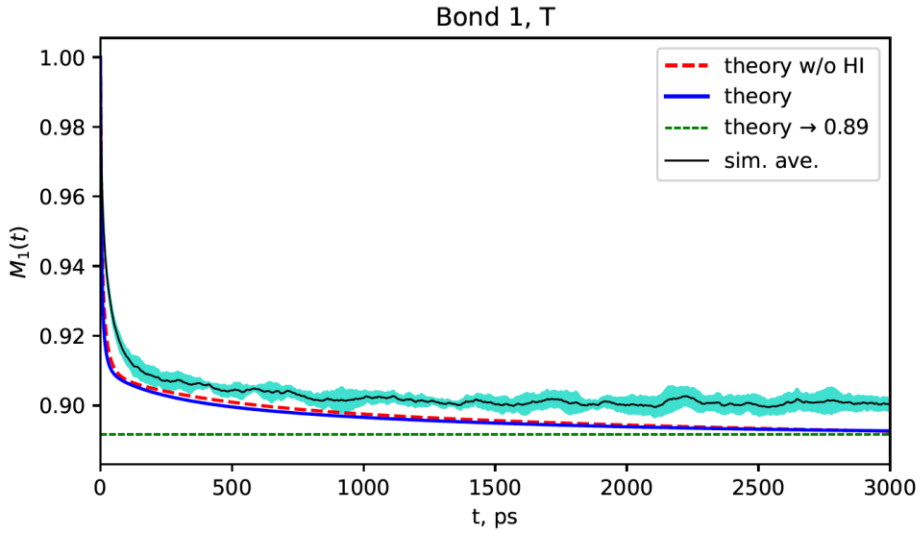
- To solve the LE4PD, we can decouple it by diagonalizing the LU matrix.
- The resulting basis is a set of normal modes.

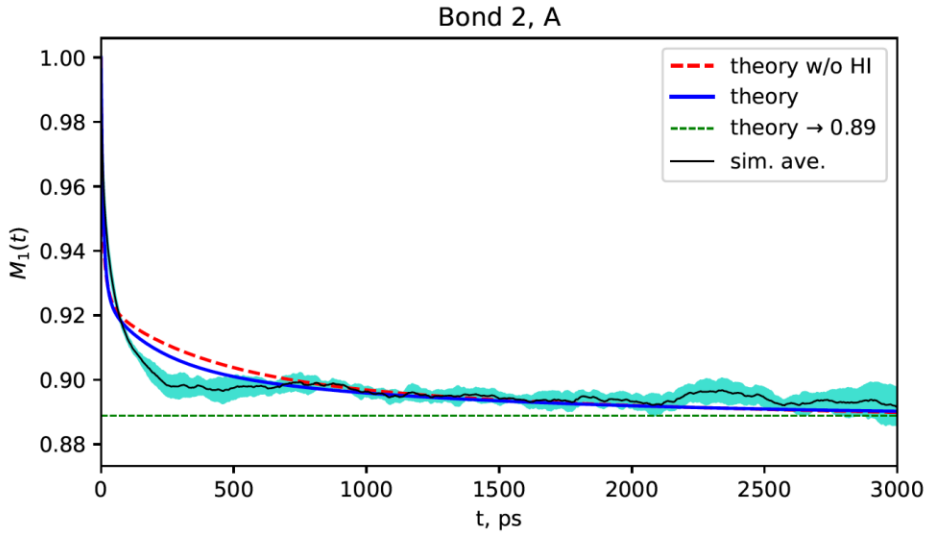
$$\frac{\partial \vec{l}_i(t)}{\partial t} = -\sigma \sum_{j,k} L_{ij} U_{jk} \vec{l}_k(t) + \vec{v}_i^{\,\text{random}}(t) \quad \xrightarrow{\mathbf{Q}^{-1}} \quad \frac{\partial \vec{\xi}_a(t)}{\partial t} = -\sigma \lambda_a \vec{\xi}_a(t) + \vec{v}'^{\,\text{random}}_a(t)$$

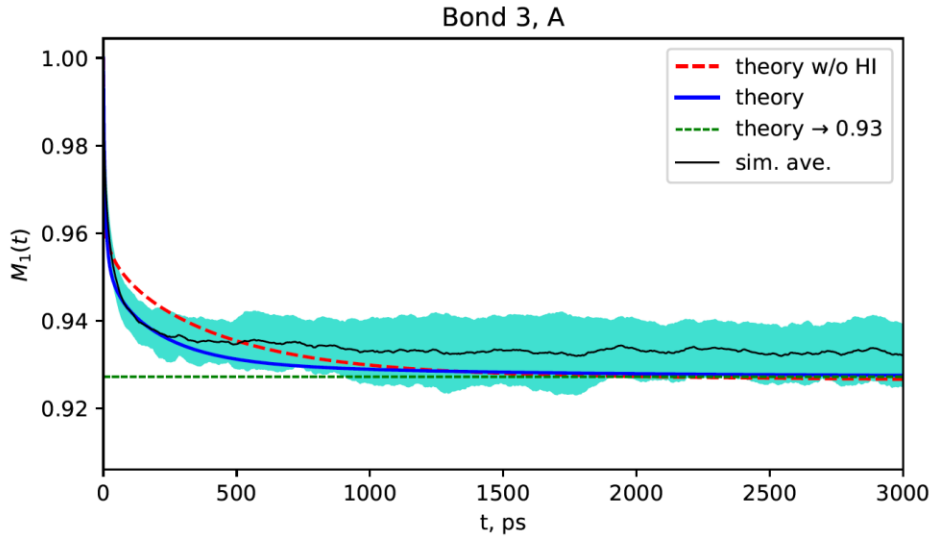
$$\vec{\xi}_a(t) = \sum_i Q^{-1}_{ai} \vec{l}_i(t)$$

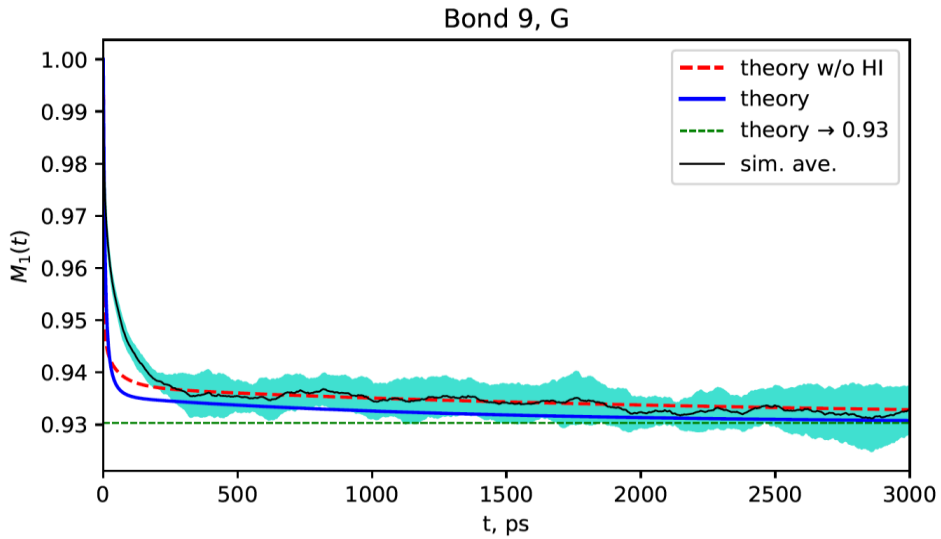
$$\delta_{ab} \lambda_a = \sum_{i,j,k} Q^{-1}_{ai} L_{ij} U_{jk} Q_{kb}$$

More bond TCFs





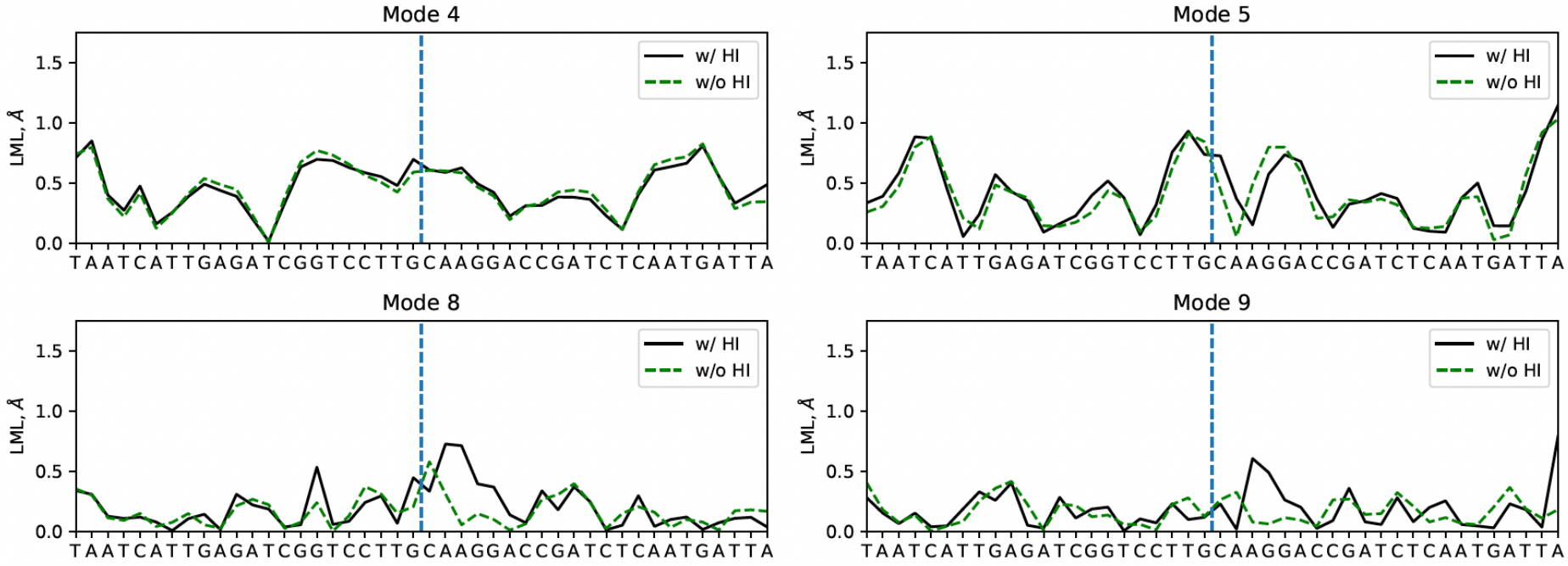




Local Mode Length Scale

- The Local Mode Length scale measures the contribution of bond i to the fluctuations in mode a.

$$L_{i,a}^2 = \frac{Q_{i,a}^2 l^2}{\mu_a}.$$



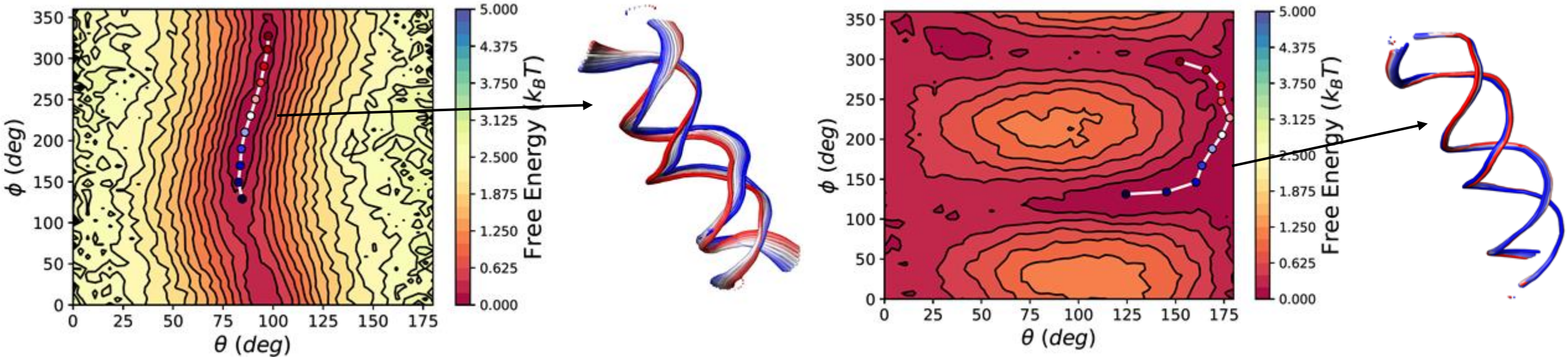
Figures by Eric Beyerle

Analysis: Free Energy Surfaces

- LE4PD normal modes can be used to construct free energy surfaces
- Free energy surfaces reveal slow and independent fluctuations

$$\theta_a(t) = \arccos\left(\xi_{a,z}(t)/|\vec{\xi}_a(t)|\right)$$
$$\phi_a(t) = \arctan\left(\xi_{a,y}(t)/\xi_{a,x}(t)\right)$$

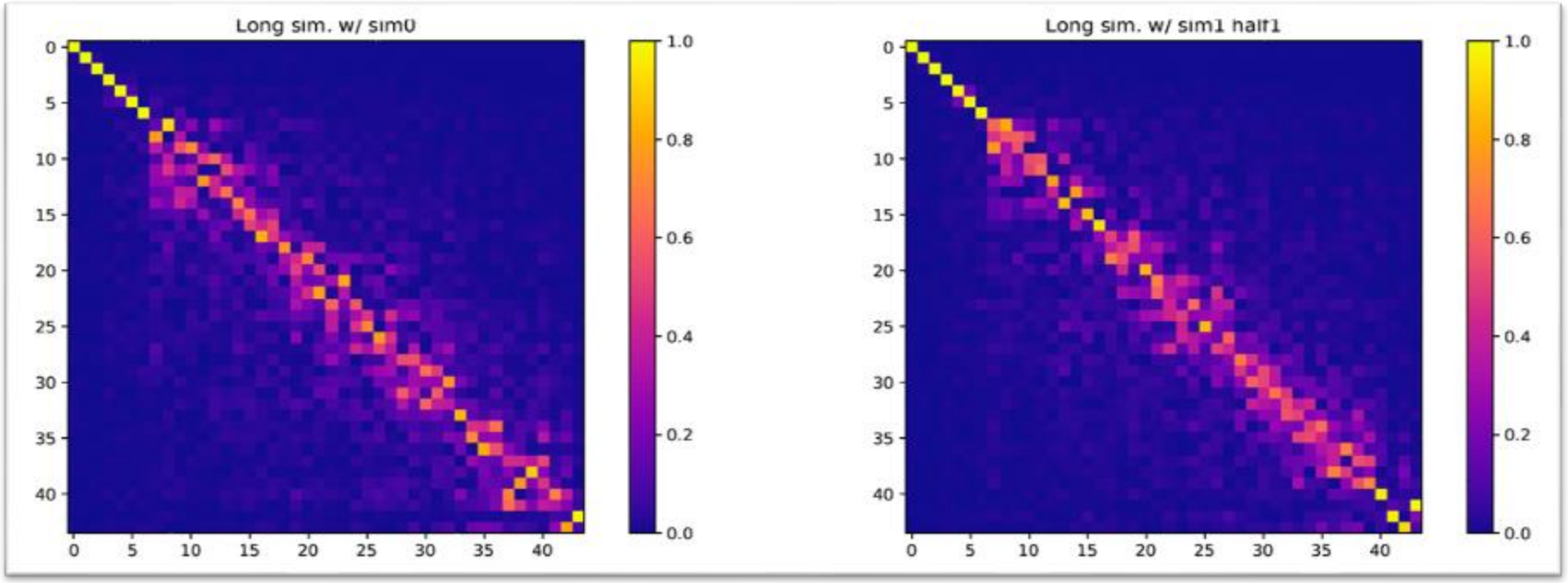
$$E(\theta_a, \phi_a) = -k_B T \ln(P(\theta_a, \phi_a))$$



Analysis: Convergence

- Determining convergence of simulation statistics can be done by comparing the diagonalizing matrix of a simulation with shorter simulations
- The degree to which this matrix can be used to invert the diagonalizing matrix of a shorter simulation is a measurement of the statistical convergence of a simulation

$$(O_{\text{overlap}})_{ij} = |\left(Q^{-1}_{\text{conc.}} \cdot Q_{\text{sim.}\alpha}\right)_{ij}|$$



Acknowledgements

- This presentation was made possible by
  - Marina Guenza
  - Eric Beyerle
  - Jesse Hall
  - Jeremy Copperman
  - David Grych