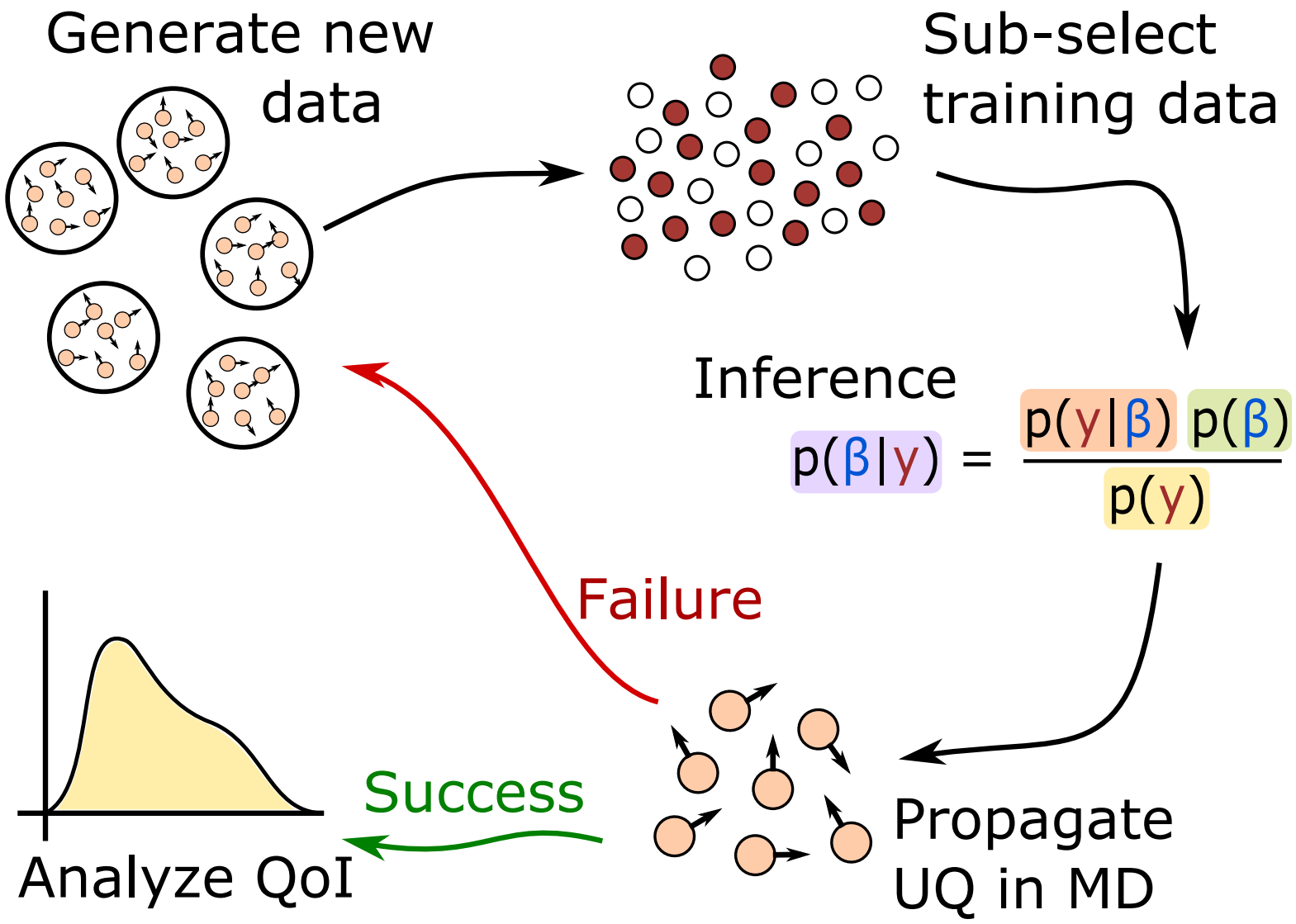
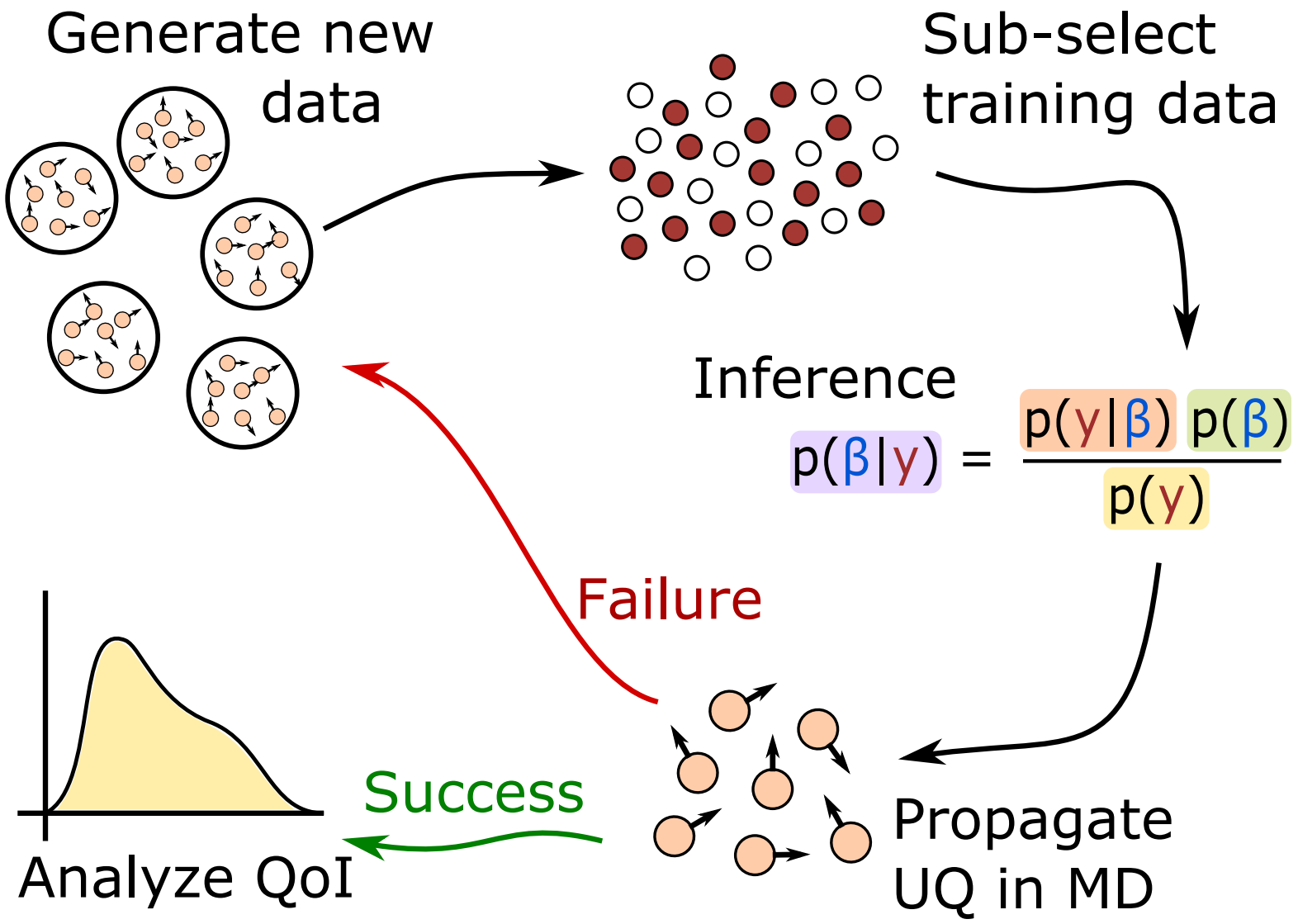
Nuances in Bayesian Modeling and Active Learning for Molecular Dynamics:



Dallas Foster,
Joanna Zhou
Jayanth Jagalur-Mohan,
Youssef Marzouk,
Center for the Exascale Simulation of Materials in Extreme Environments,
Massachusetts Institute of Technology

MITCESMIX

How I learned to stop learning and love uncertainties (to an extent)



Dallas Foster,
Joanna Zhou
Jayanth Jagalur-Mohan,
Youssef Marzouk,
Center for the Exascale Simulation of Materials in Extreme Environments,
Massachusetts Institute of Technology

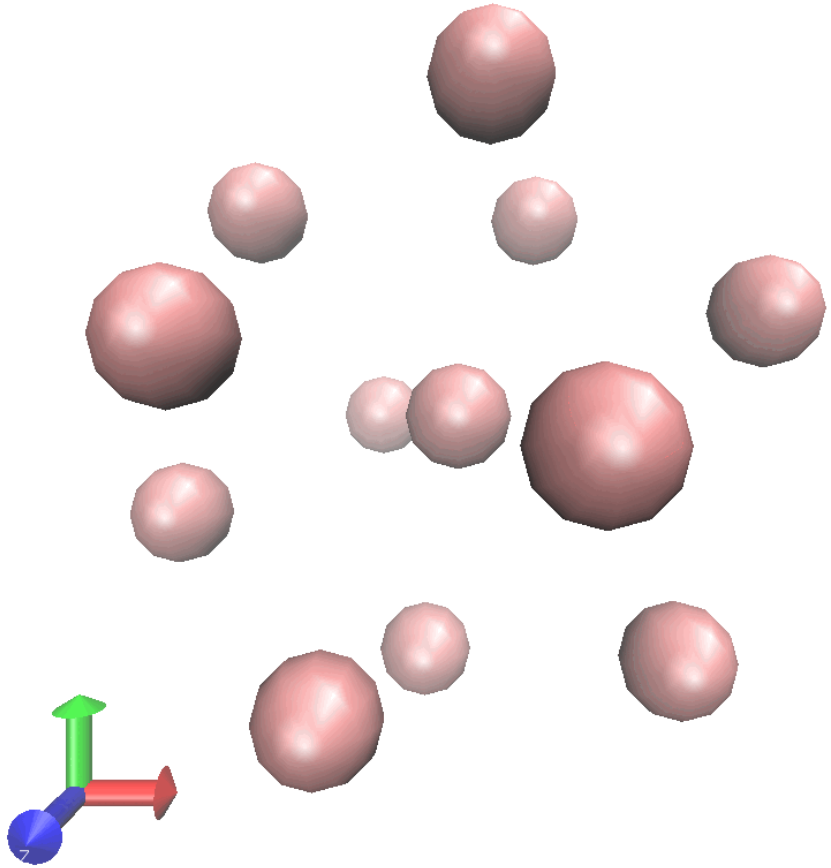
MITCESMIX

Test Application: Lennard Jones Clusters

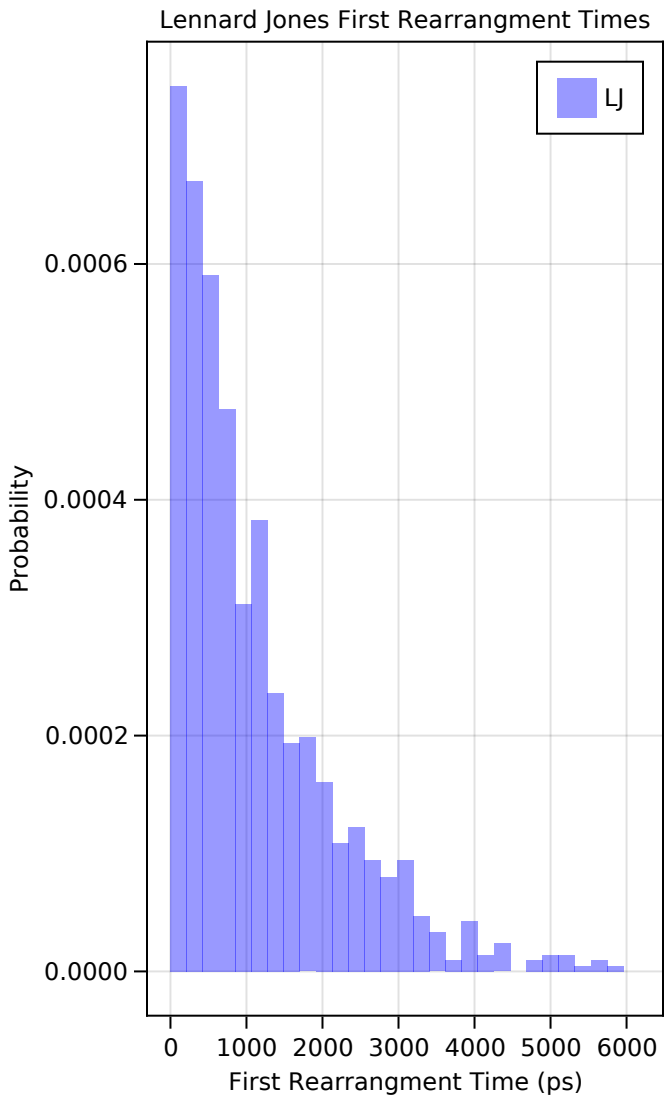
- Cluster of 13 atoms, evolve according to LJ potential.

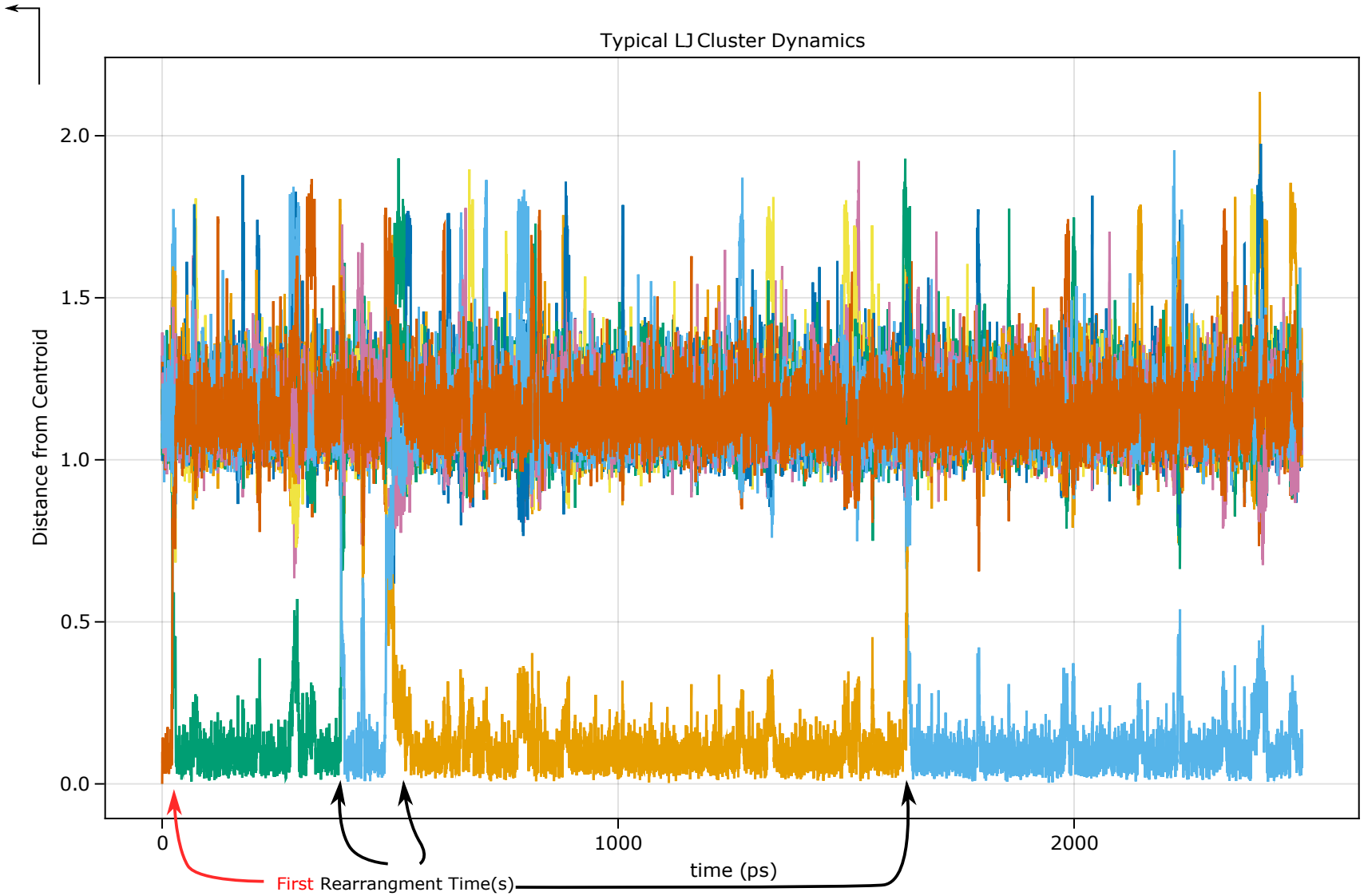
$$V(r) \propto \frac{1}{r^{12}} - \frac{1}{r^{6}}$$

- How sensitive is the rearrangement rate to errors in fitted potential?

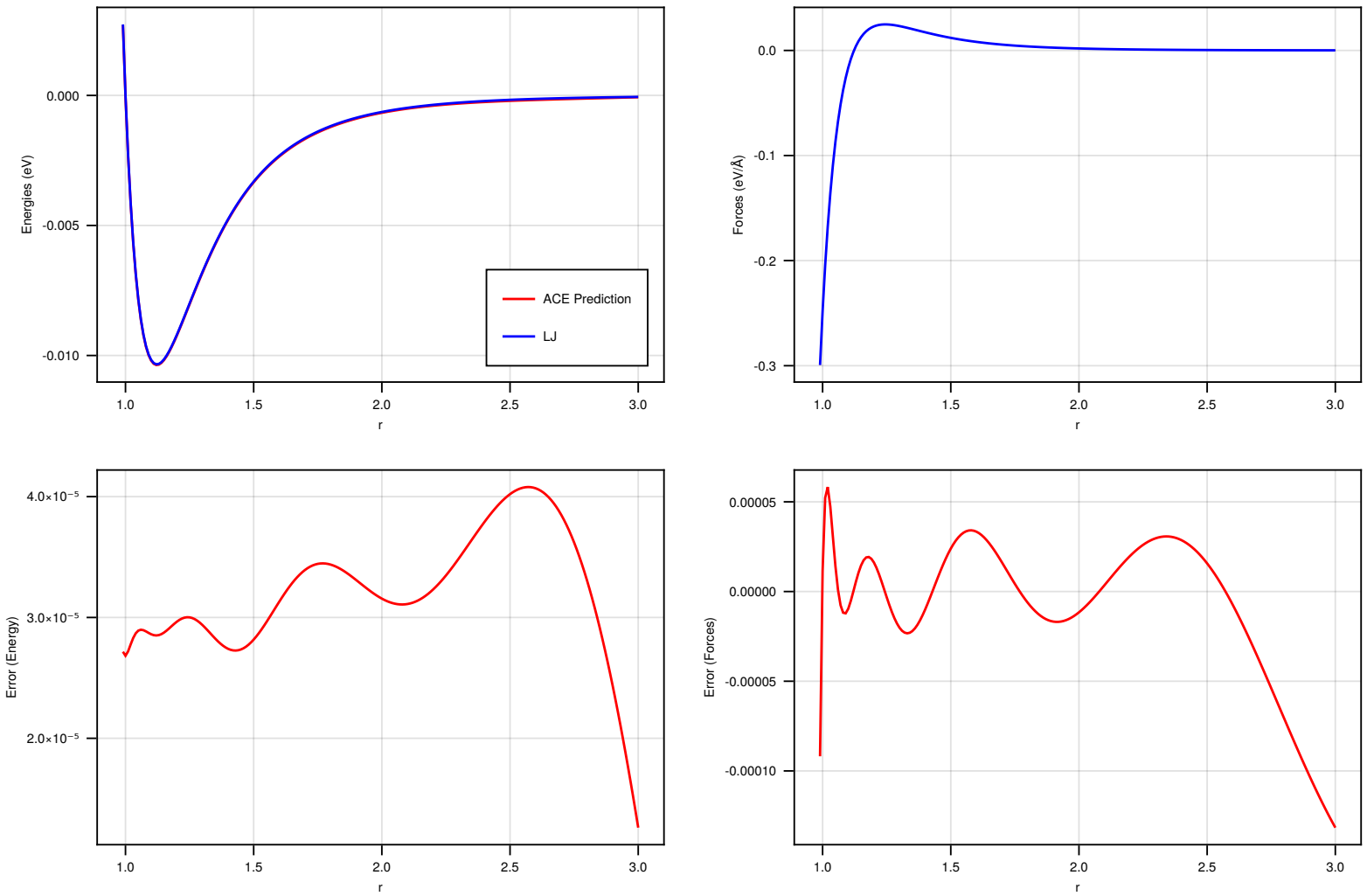


Application: Lennard Jones Cluster Rearrangements





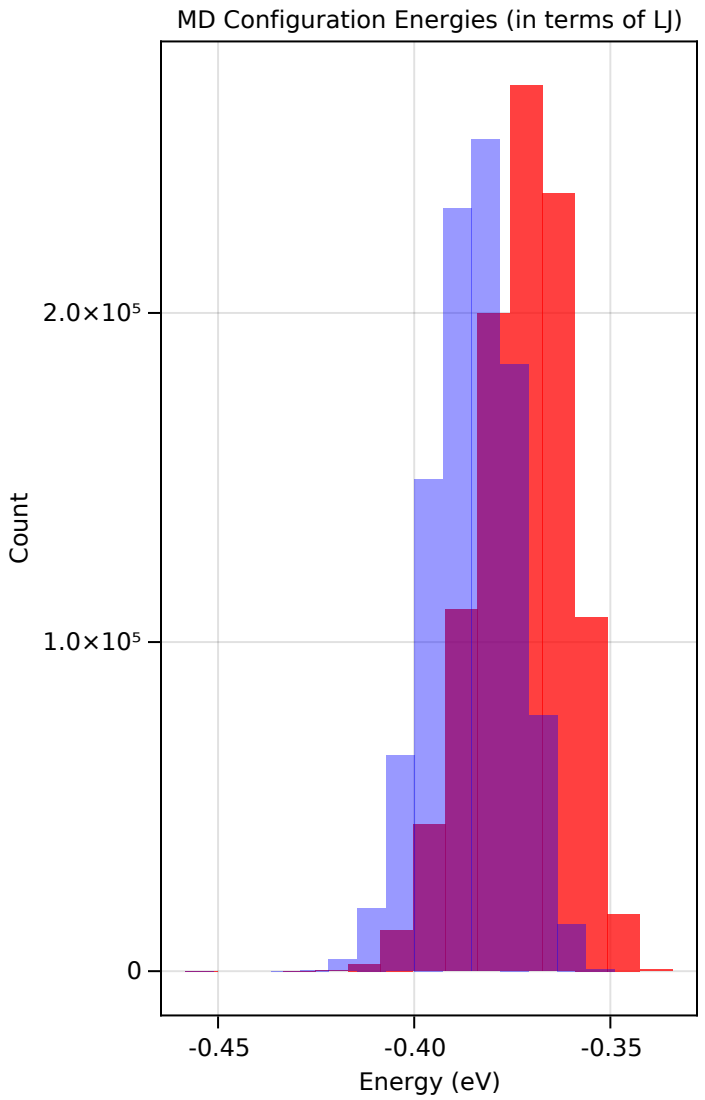
Naïve Fitting

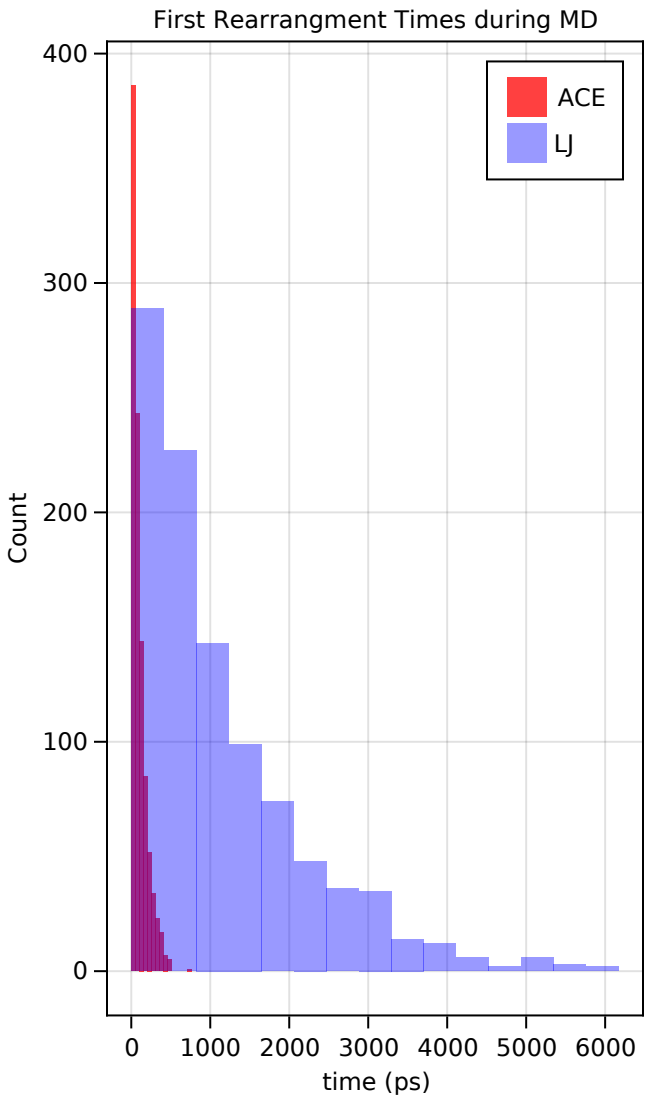


(Top) Predicted and true energy and force profiles (as a function of interparticle distances)
(Bottom) Error in predicted energy and force.

- Arbitrarily select large subset of data.
- Select radial basis set to approximate potential energy surface (e.g. using ACE)
- Fit to data using weighted least squares (i.e. maximum likelihood).

Estimating LJ First Rearrangement Times





(Left) Energy histogram over MD trajectories. (Right) FRT over 1000 random seeds, given single optimal estimate of IAP coefficients.

- Run MD using MLE estimate of basis coefficients with different random seeds and initial velocities.

Bayesian Methodology

Data Likelihood:
All models make approximations.
How do we quantify errors and account for discrepancies? Do we have to incorporate additional regularizations?

Prior Information:
Leverage expert knowledge.
Are there relationships between parameters and physical properties?

$$p(\beta|y) = \frac{p(y|\beta)\,p(\beta)}{p(y)}$$

Posterior:
How do we represent parameter uncertainty?
How do we propagate uncertainty through MD?

Model evidence:
All data is not created equal.
How do we select data for training or query new data?

Bayesian Methodology

$$p(\beta|y) = \frac{p(y|\beta)\,p(\beta)}{p(y)}$$

Likelihood:

- Gaussian likelihood: $\boldsymbol{y} \sim N(A\boldsymbol{\beta}, Q)$
- Sampling distributions
- "Simulation-based inference": $\mathrm{q} \sim N(g\,(\boldsymbol{\beta}), Q)$

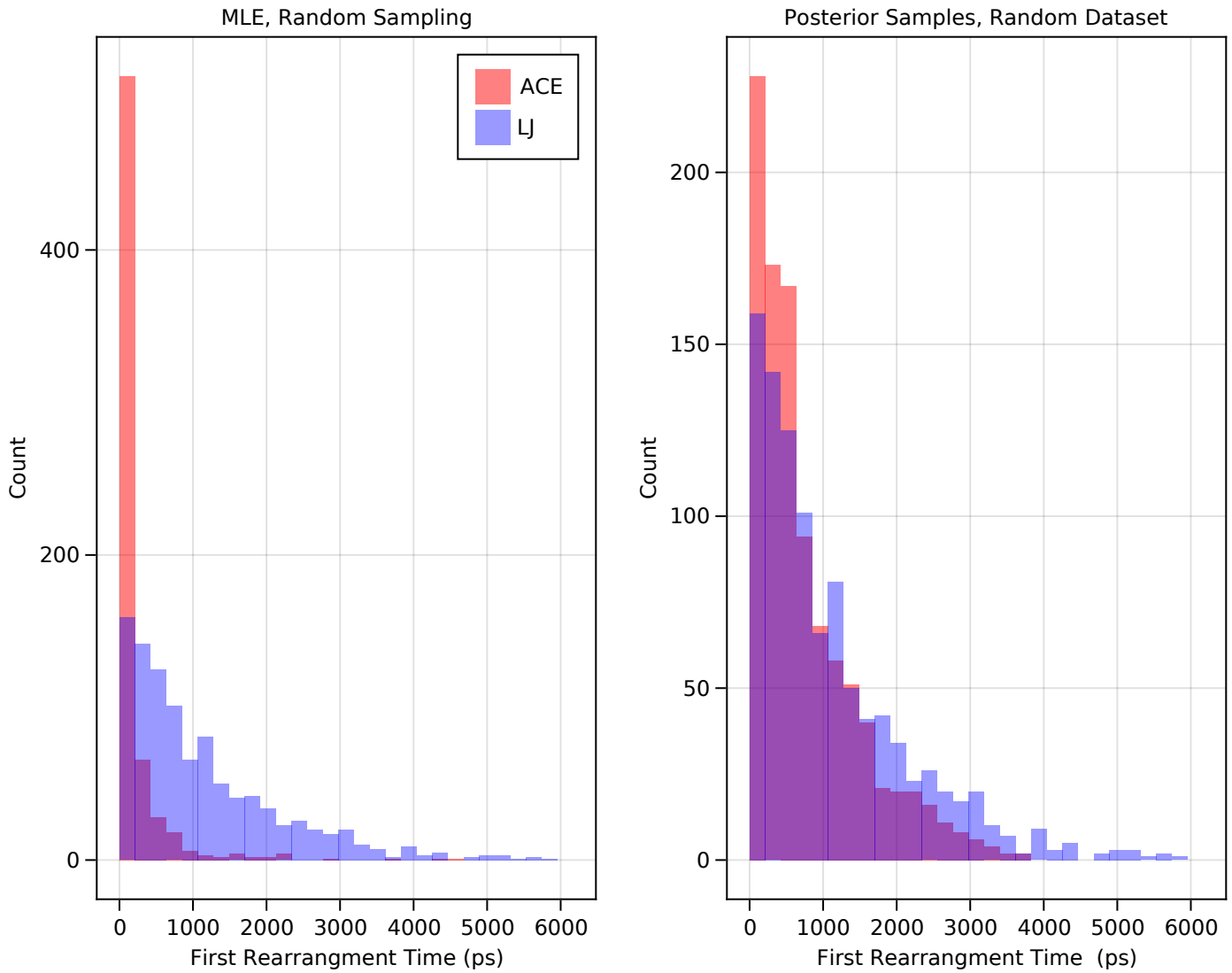
Prior:

- Gaussian prior: $\boldsymbol{\beta} \sim N(\mathrm{m}, \Sigma)$
- Sparsity inducing priors: Horseshoe, Laplace, …

Things to consider:

- Gaussian likelihoods can lead to vanishing variance, overconfident predictions with increasing data.
- Assumption of gaussian errors might not be appropriate and lead to model mispecification.
- How to incorporate uncertainty in hyperparameters?
- MAPs with Gaussian priors further shrink predictive covariance.
- None of these typical strategies focus on components of the potential energy surface that we are less certain about.

LJ First Rearrangement Times

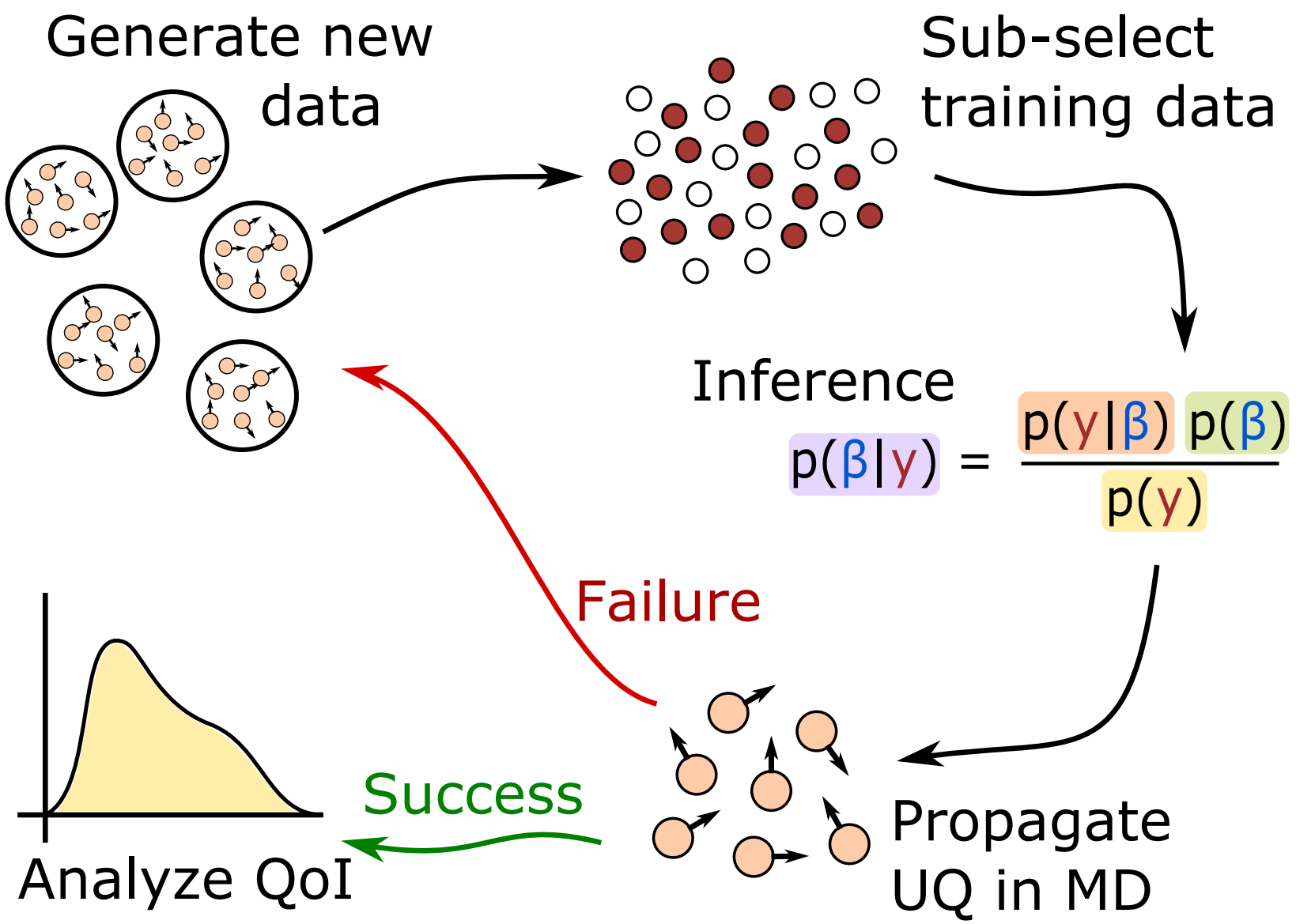


(Left) FRT histogram with IAP coefficients optimized over different randomly sampled data subsets. Seeds and coefficients are independently sampled. (Right) FRT over 1000 random seeds, sampled over the Bayesian posterior for a single dataset.

- Attempt #2:
  - Select multiple random training datasets.
  - Optimize parameters
  - Run MD with multiple seeds and coefficients
- Attempt #3
  - Select random training dataset
  - Optimize parameters
  - Run MD with samples from Bayesian posterior

Kernelized Active Learning

- Use similarity measures and kernels to drive active learning.
  - Determinantal Point Processes (DPPs) for Active Learning
  - Gaussian Processes for error tracking (or fitting, GAP)
  - Kernelized modified molecular dynamics for generating new configurations.



Active Learning: Determinantal Point Processes

- First identified as “fermion processes” by Macchi (1975), leading to the “anti-bunching” effect.

- Given a set A, a point process is simply a probability measure on the set of subsets of A.

- P is determinantal point process if, when Y is a random subset drawn according to P, we have

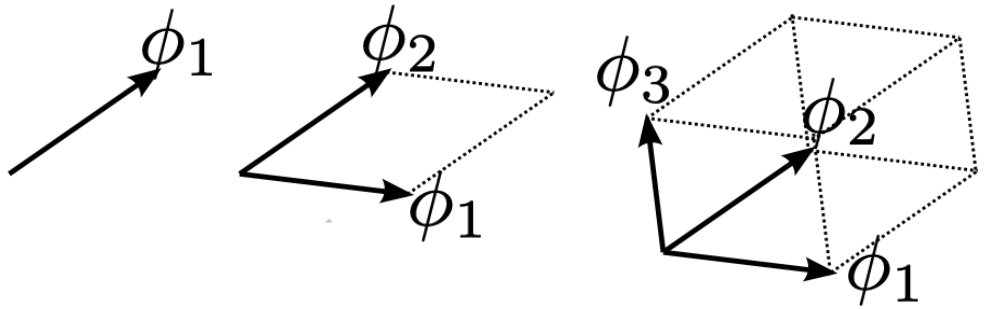
$$P_{L(Y=A)} \propto det(L_A)$$

- Here L is positive semi-definite, defines the DPP, and is labeled the L −ensemble or the DPP kernel.

Active Learning: DPP

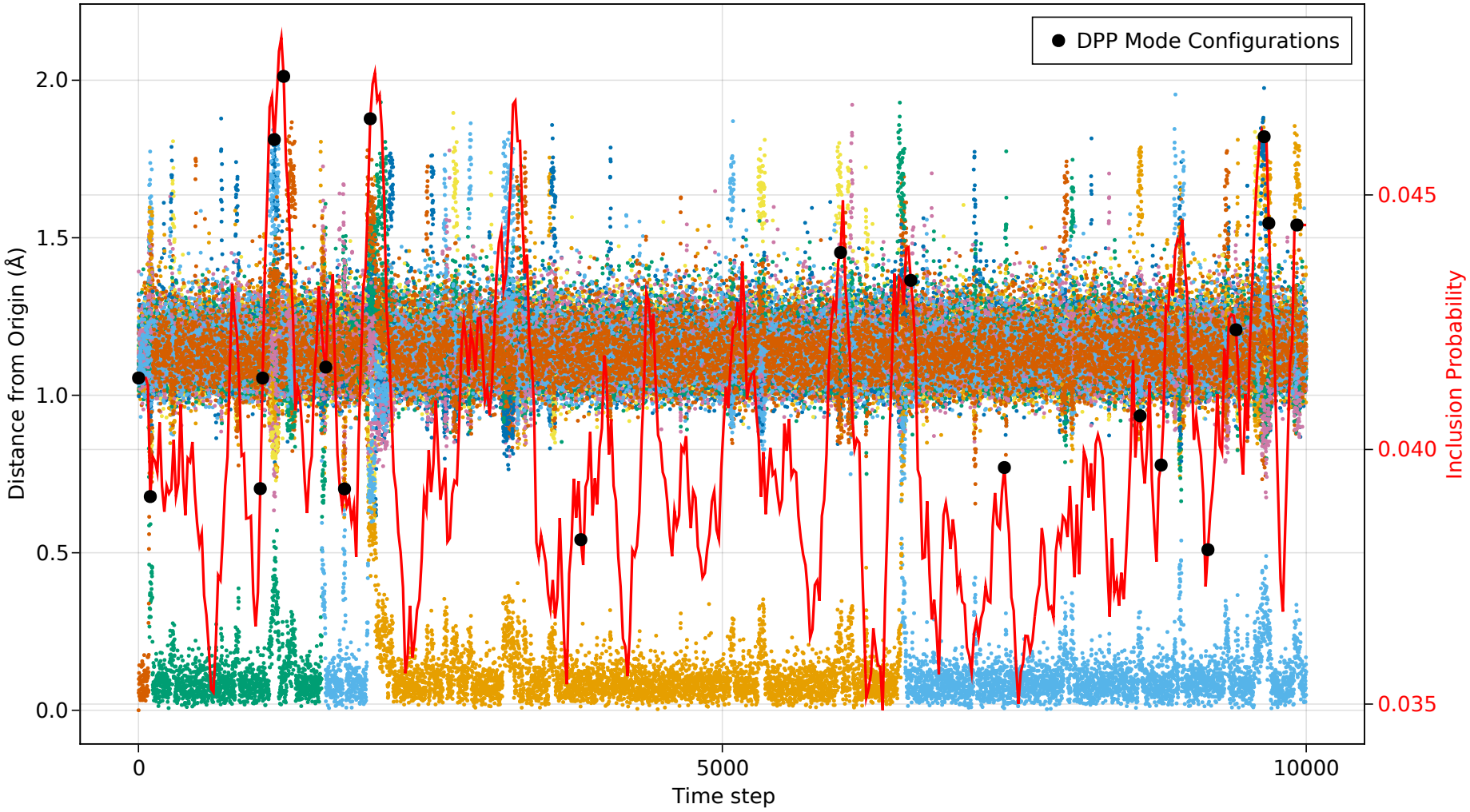
- The entries of the kernel can be built as $\mathrm{L_{ij}} = k(\phi_i, \phi_j) q_i q_j$, where $\mathrm{q_i} \in \mathrm{R}^+$ is the quality term, and $k(\cdot,\cdot)$ is a function on the set of feature vectors, $\phi_\mathrm{i} \in \mathrm{R}^\mathrm{d}$, that models diversity between features.

$$\mathcal{P}_{L(Y)} \propto \prod_{I \in Y} q_i^2 \det(S_Y), \quad \text{where} \quad S_{ij} = \phi_i^\top \phi_j$$



- In the context of training Inter-atomic potentials, ϕ_i can be a set of descriptors associated with each configuration, and the quality $\mathrm{q_i}$ can be some measure of entropy.

DPP Application: Fitting Lennard Jones Clusters

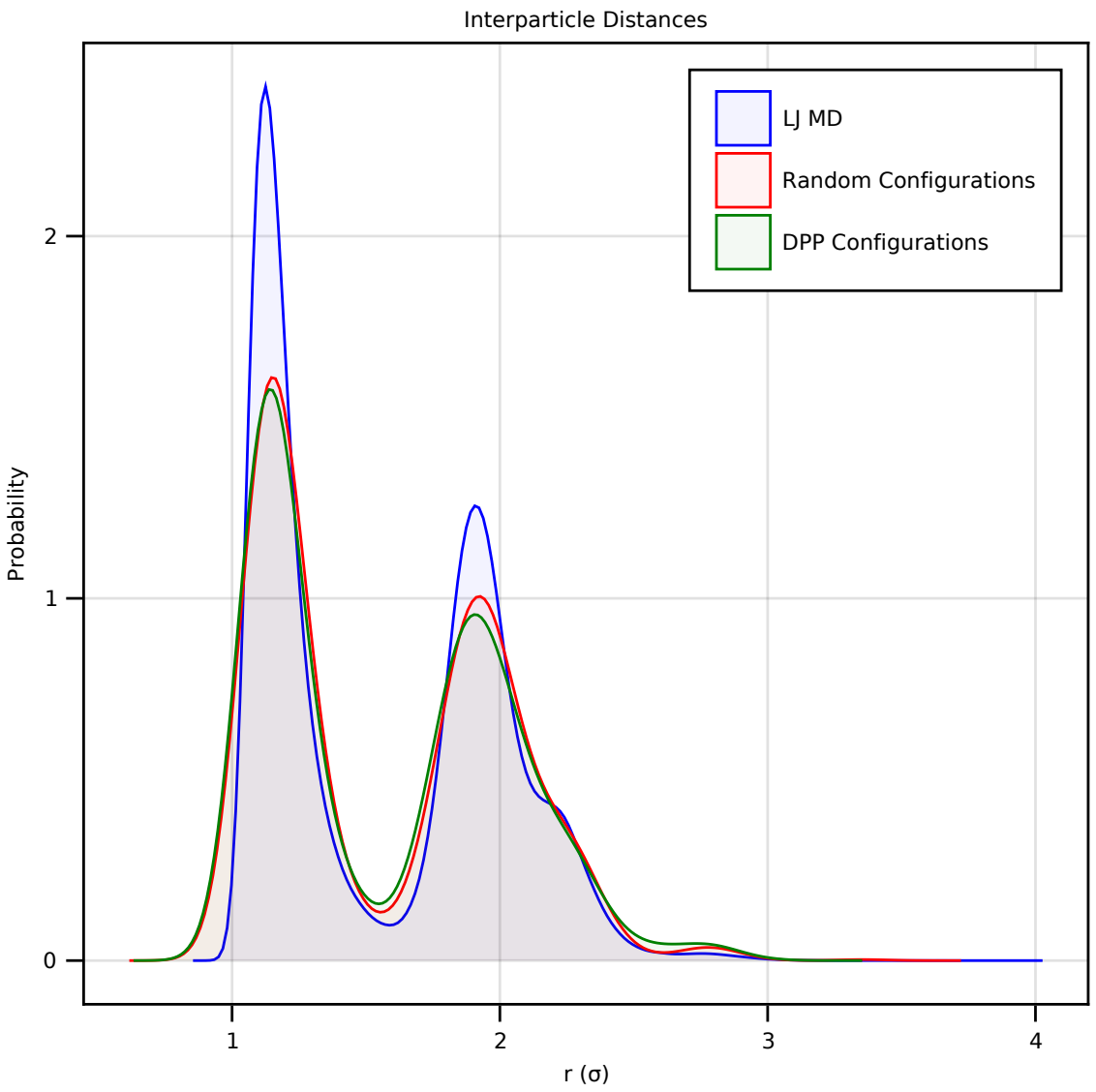


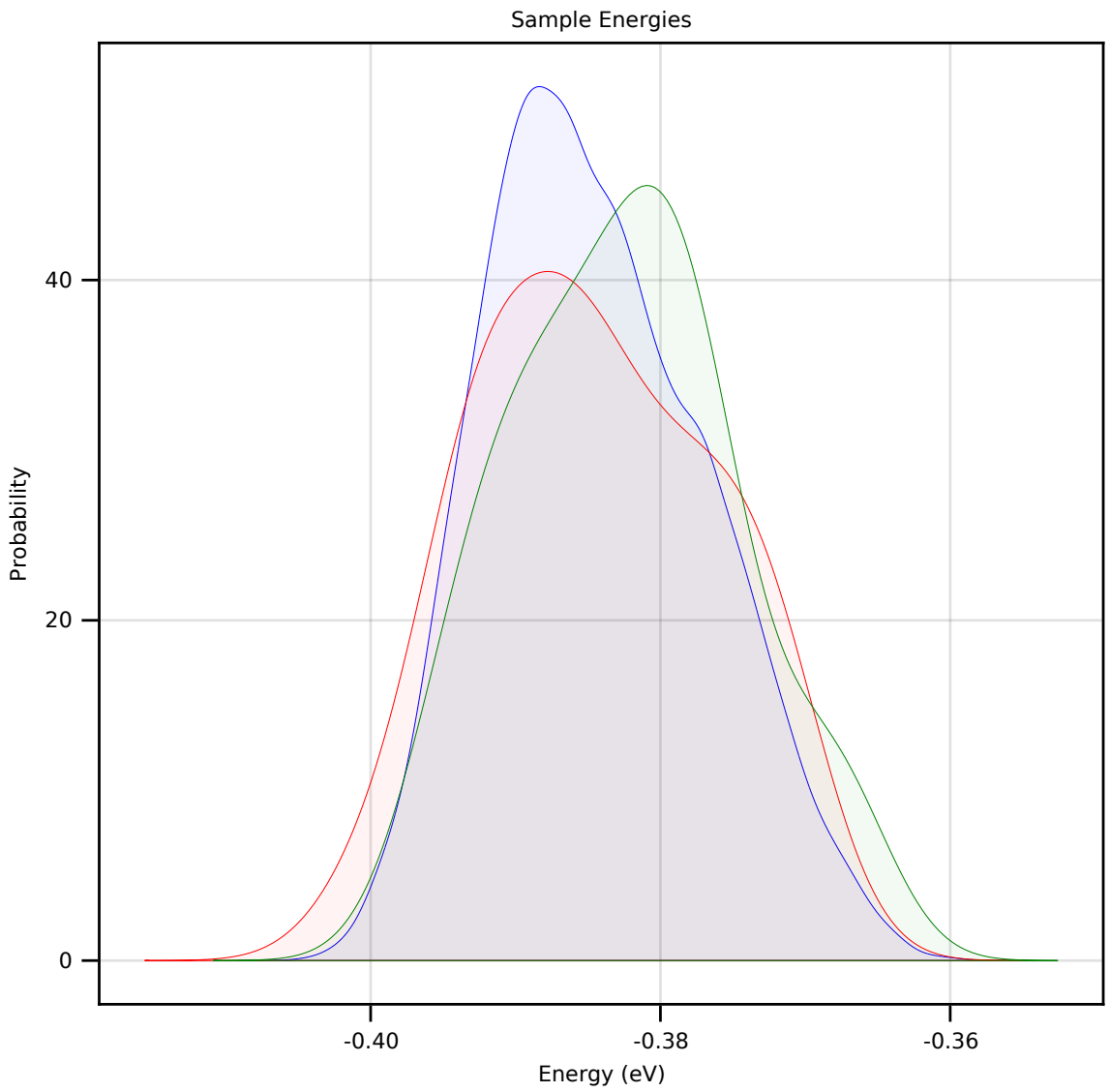
Inclusion Probabilities (notion of uniqueness of certain configurations) as produced from DPP kernel.

Two portions of time series were purposely omitted to reduce dimensionality of data.

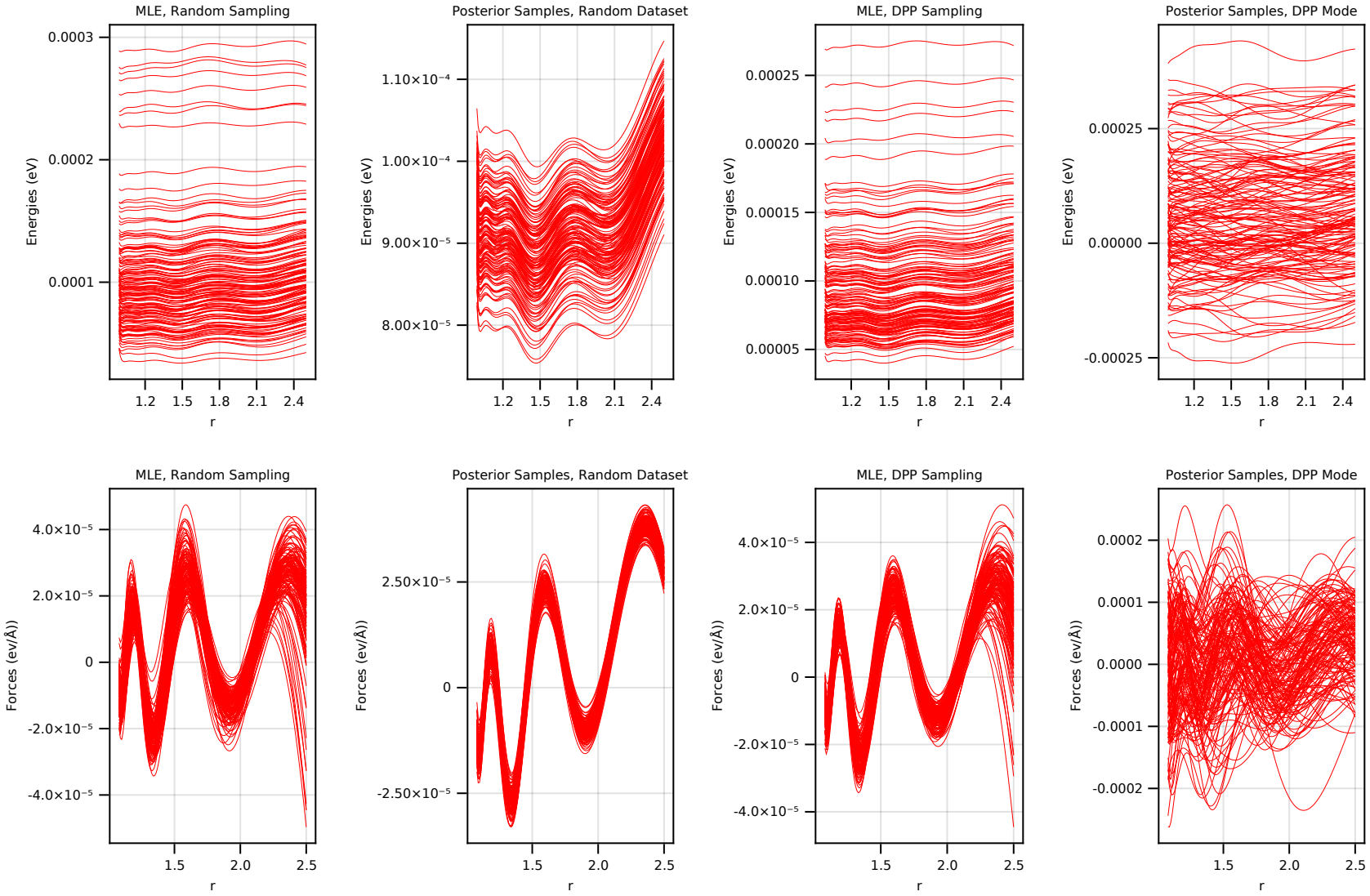
Qualitative results are not overly sensitive to (appropriate) choice of kernel.

DPP Application: Fitting Lennard Jones Clusters





Variance and Bias trade-off



(Top) Normed errors between Lennard Jones energy, and predicted energies.

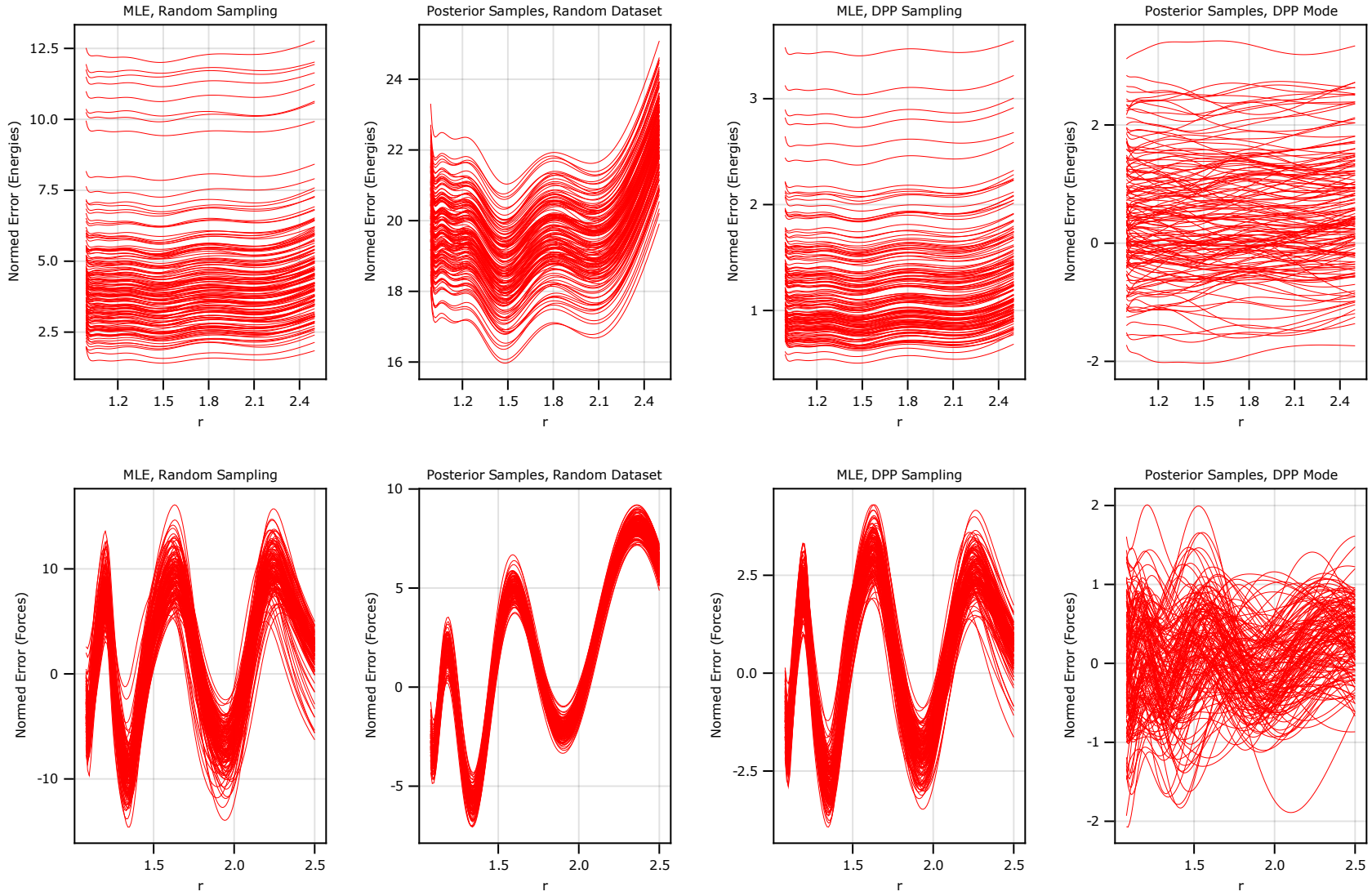
Individual lines correspond to different sets of IAP coefficients.

(Bottom) Same as above, for forces.

Energy errors are defined as

$$\mathbf{e_{true}} - \mathbf{e_{pred}}$$

Effect of variance of predictions



(Top) Normed errors between Lennard Jones energy, and predicted energies.

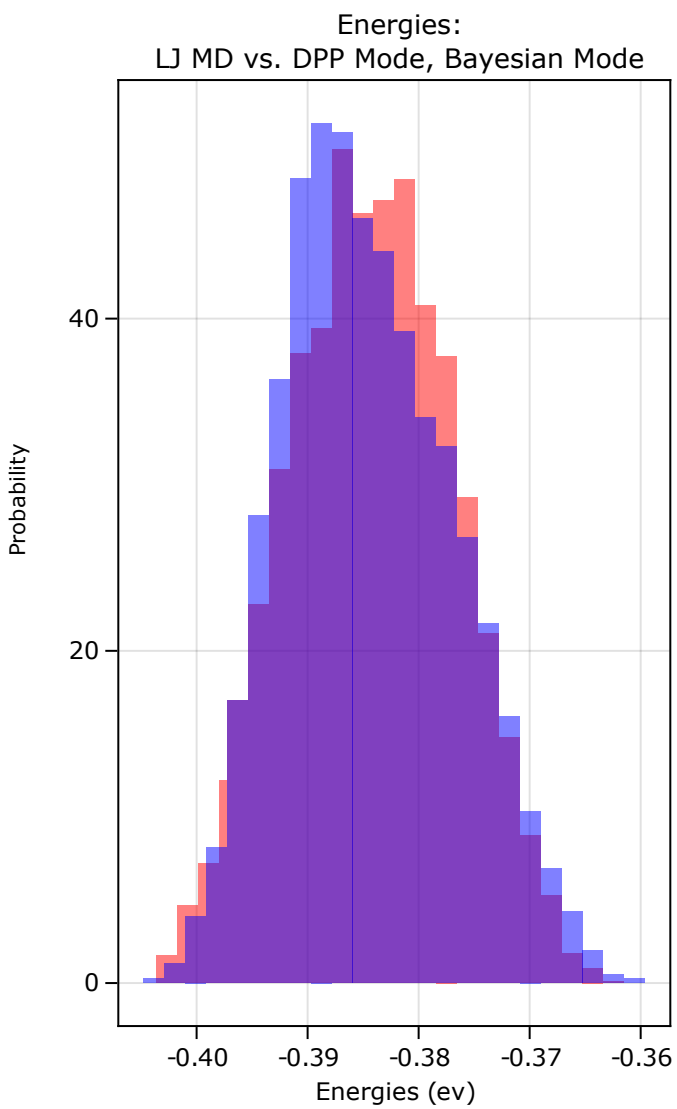
Individual lines correspond to different sets of IAP coefficients.

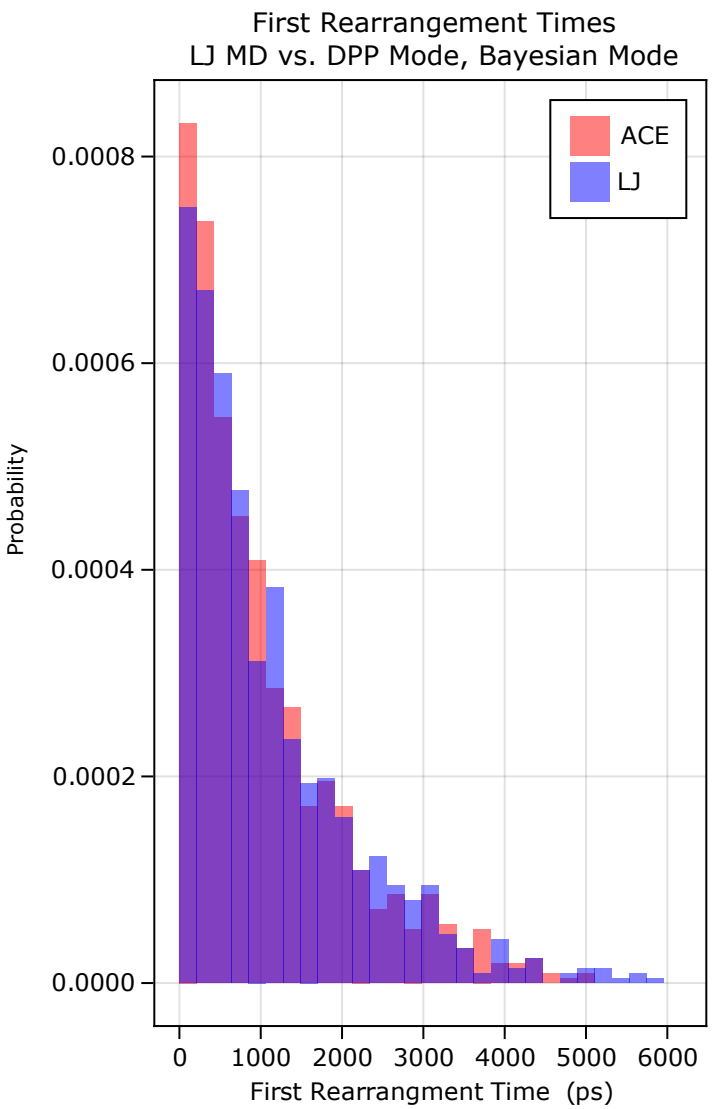
(Bottom) Same as above, for forces.

Normed (energy) errors are defined as

$$\frac{e_{true} - e_{pred}}{std(e_{pred})}$$

LJ First Rearrangement Times



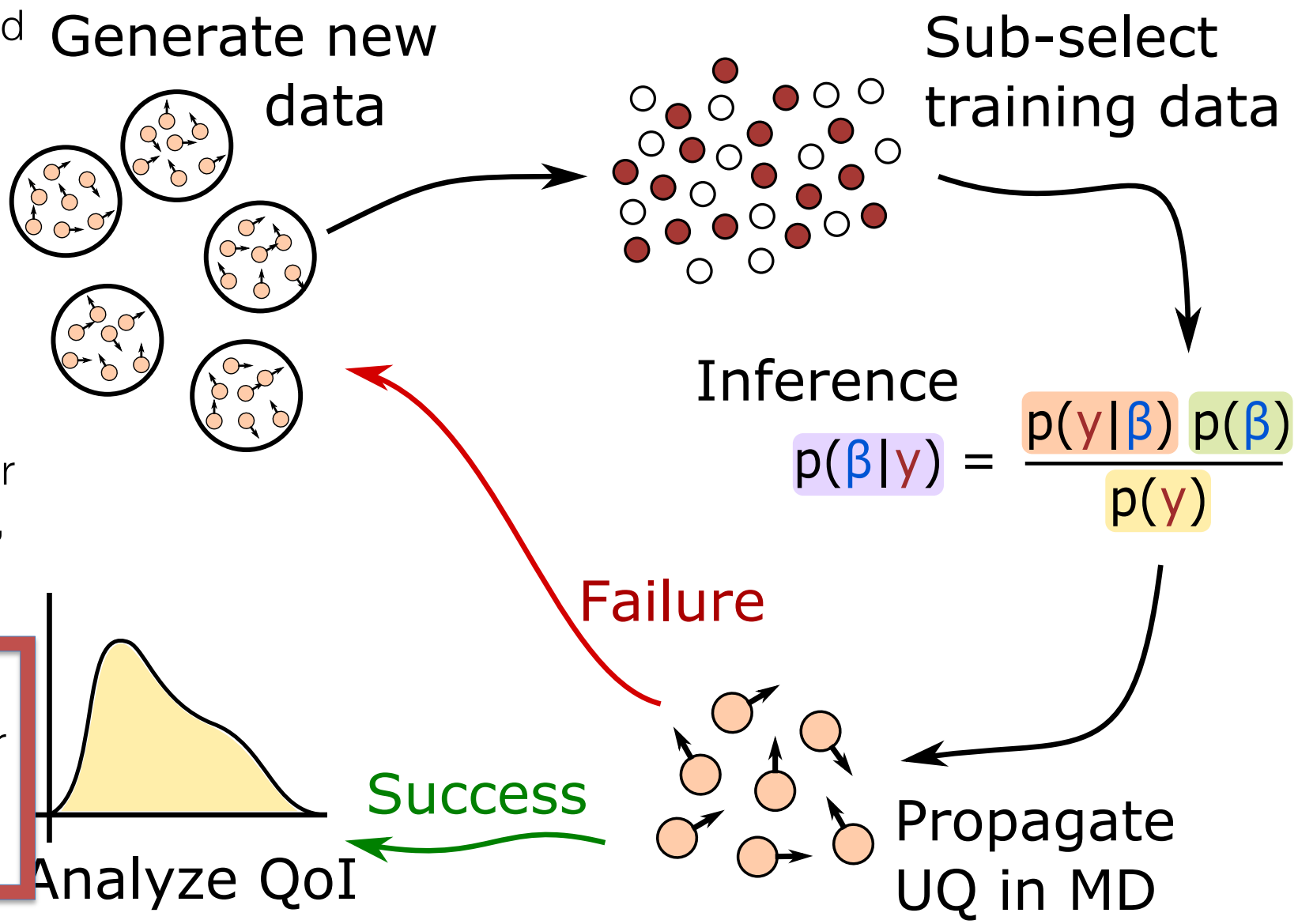


(Left) Energy histogram over MD trajectories. (Right) FRT over 1000 random seeds, given single optimal estimate of IAP coefficients.

- Using DPP with Bayesian Inference:
  - Approximate DPP Mode dataset.
  - Infer weights and parameterize jointly.
  - Run MD with Bayesian MAP.
  - seeds

Kernelized Active Learning

- Use similarity measures and kernels to drive active learning.
  - Determinantal Point Processes (DPPs) for Active Learning
  - Gaussian Processes for error tracking (or fitting, GAP)
  - Kernelized modified molecular dynamics for generating new configurations.



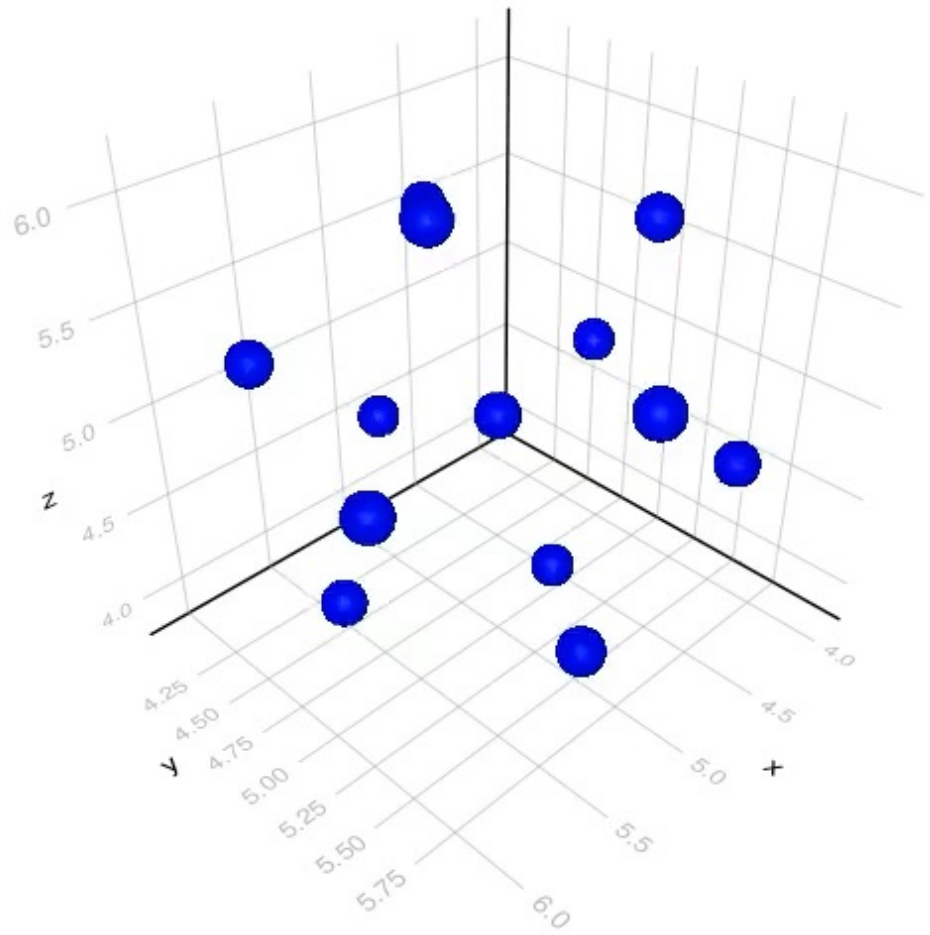
Kernalized Modified Molecular Dynamics

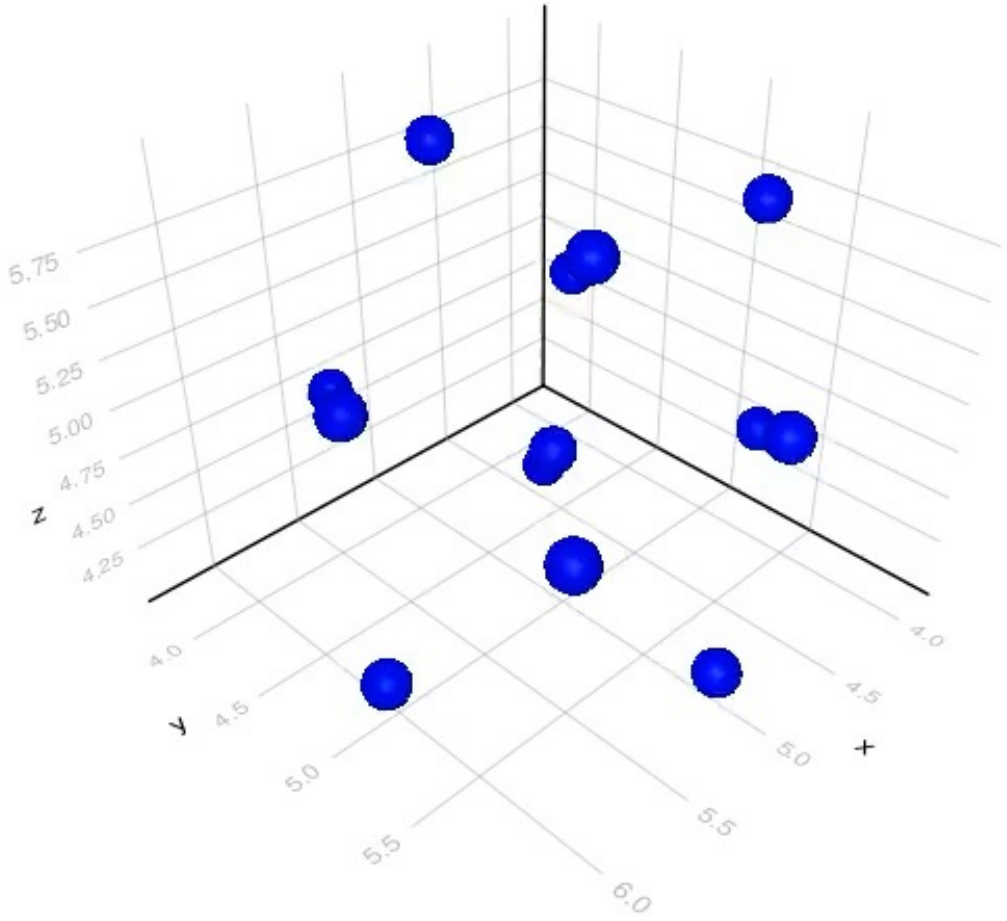
- Select $i = 1, \ldots, N$ initial configurations, $\{x\}_0^i$
- Compute similarity L_{ij} between pairs of configuration.
- Run modified molecular dynamics,

$$dx_t^i = \frac{1}{N} \sum_{j=1}^{N} \left[\underbrace{L_{ij} \nabla V(x_t^j)}_{\text{Computes smoothed gradient of “nearby” configurations.}} + \underbrace{\nabla_{x^j} L_{ij}}_{\text{Repulsive term}} \right] + \sigma dW_t^i$$

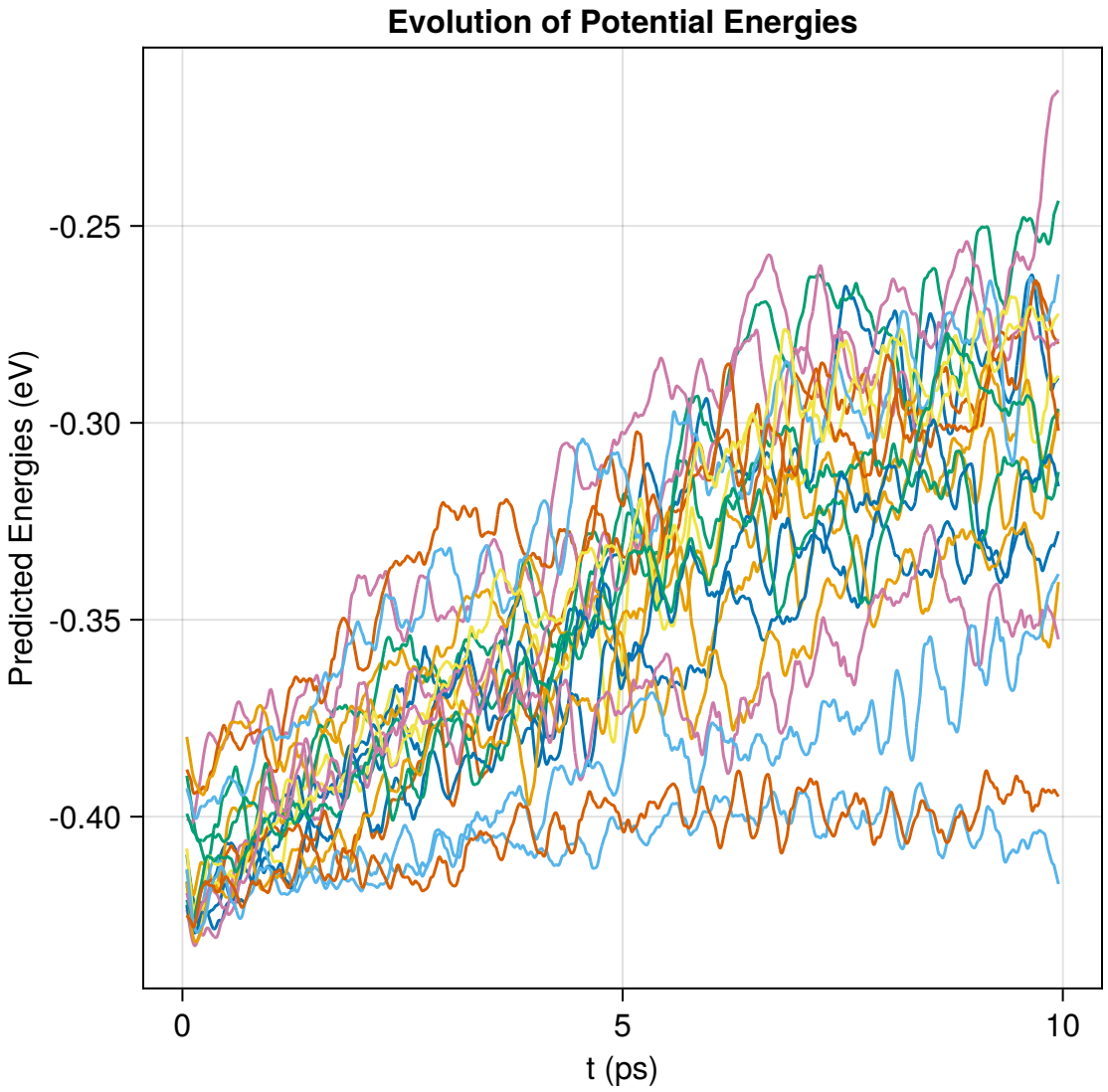
- Take advantage of theory from RKHS, Interacting Particle Systems
  - Induces “quadrature points” for the potential energy surface.

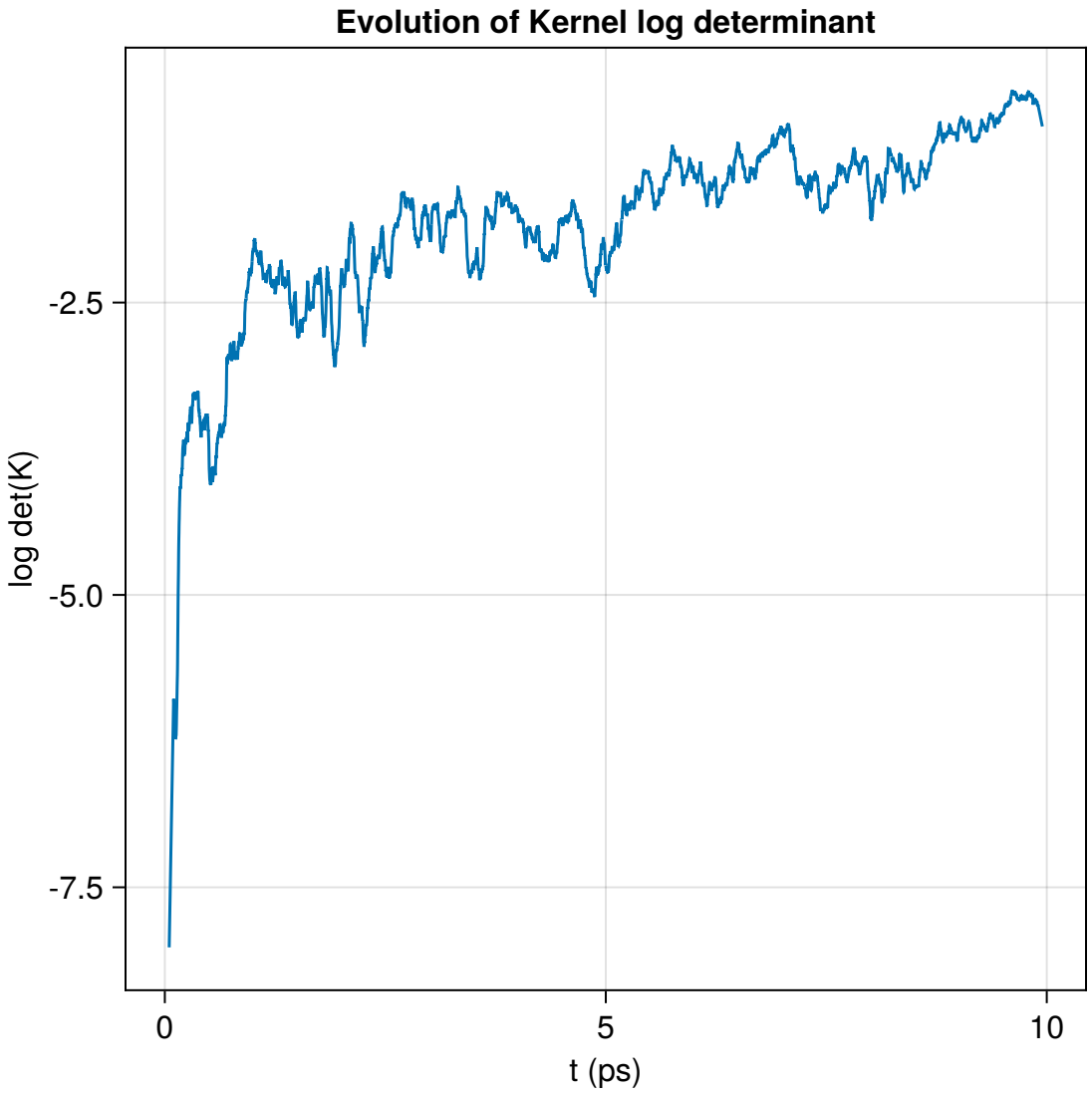
Active Learning: Kernel Modified Molecular Dynamics



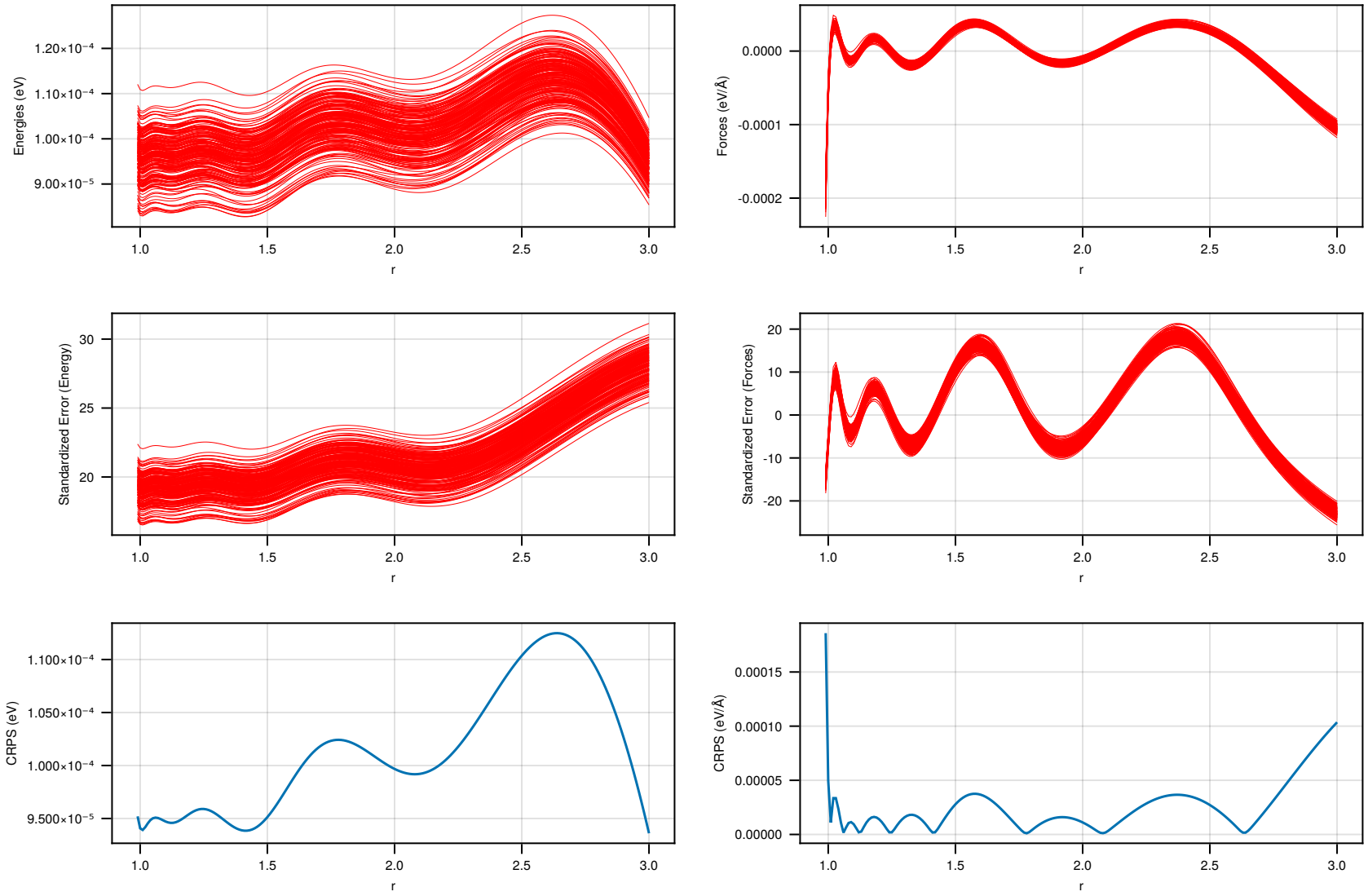


Active Learning: Kernel Modified Molecular Dynamics



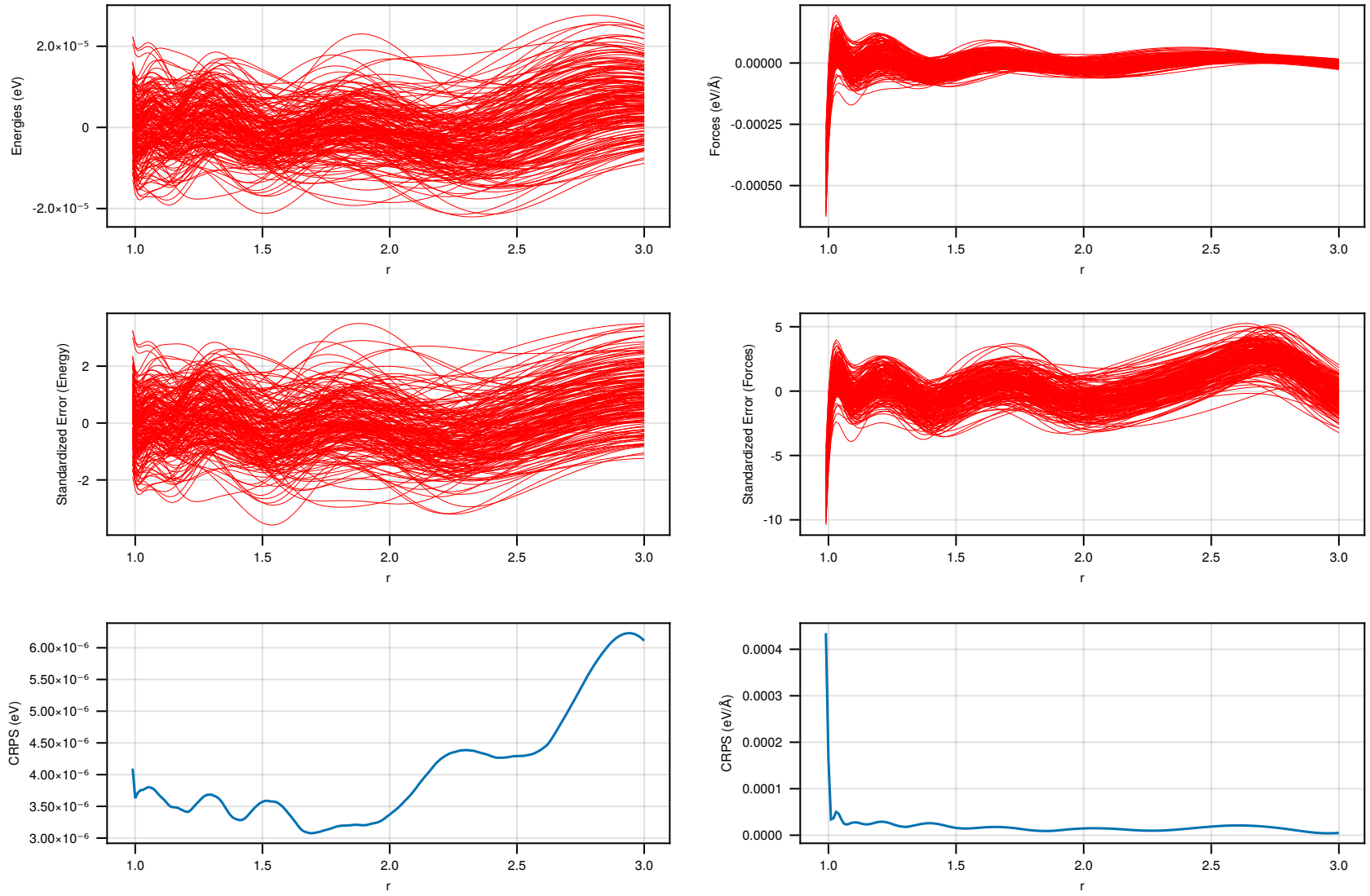


Active Learning: Kernel Modified Molecular Dynamics



Fits with initial sample of configuration leads to biased predictions with small variance.

Active Learning: Kernel Modified Molecular Dynamics



Fits with samples taken from modified molecular dynamics have less bias, smaller CRPS.

Conclusion

- Variability of dataset, representation of errors have large impact on data-driven IAP performance.
  - Critical to have performant, efficient method to sub-select data based on diversity, quality.
  - Accurate estimation of error in training data, and error in fitted potentials ensure calibrated probability estimates.
    - Mileage may vary depending on QoI.
- Many areas of continued work
  - Can we use QoI's to inform regularization, dimension reduction techniques.
  - Understand theory on metrics for comparing descriptors of atomic environments.