MIT-IBM Watson AI Lab

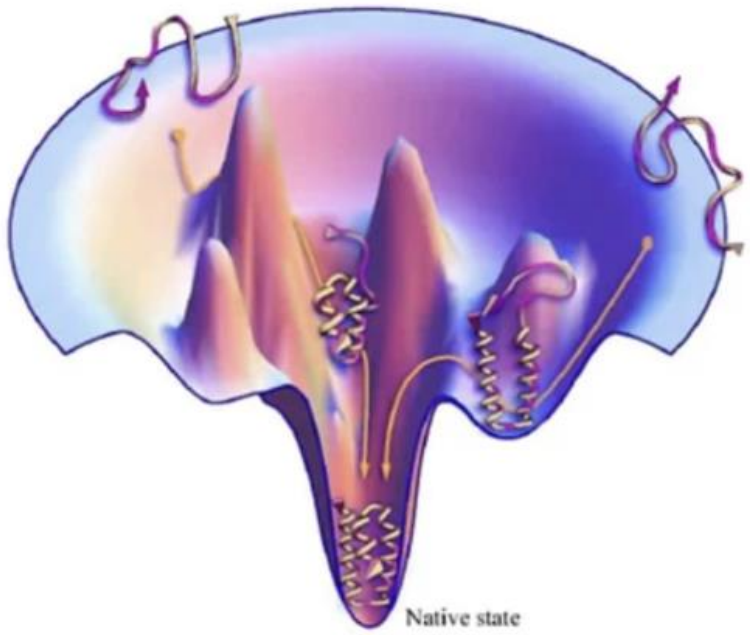
Symmetry-Driven Discovery of Dynamical Variables in Molecular Simulations

Jeet Mohapatra[1], Nima Dehmamy[2], Csaba Both[2,3],
Subhro Das[2], Tommi Jaakkola[1]

MIT CSAIL

Challenge: Boltzmann Sampling of Molecular Configurations

- Molecular systems reside in **high-dimensional space** with **rugged energy landscape** making direct importance sampling infeasible.
- **Metastability Problem :** MD Simulations used to sample from the space
  - Simulations get stuck in **local energy minima (metastable states)**
  - Standard MD samples well **within** a metastable basin, but **global transitions** are difficult.
  - Transitions between metastable states are **rare but significant events** (e.g., folding, binding)



K. A. Dill and J. L. MacCallum, Science 338, 1042 (2012)

How do we overcome the Metastability Problem?

The Metastability Problem

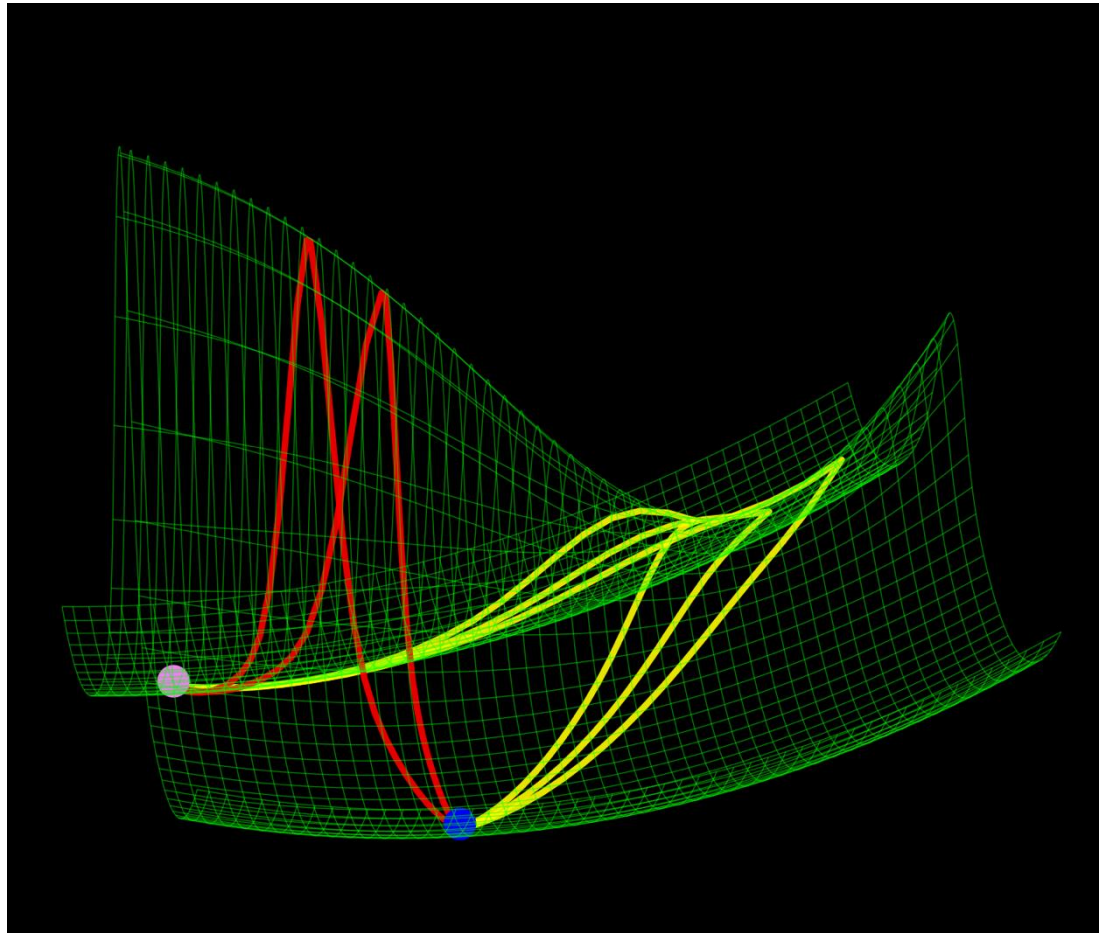
Trapped in Energy Wells

- Solid balls - Metastable states are traps
- Red Paths - Infeasible due to high energy barrier
- Yellow Paths - Feasible but slow dynamics to get to valley

Key Insight: Standard MD spends most time oscillating in wells, rarely crossing barriers.

Goal: Speed up the slow dynamics.

Informally, we need to make yellow paths faster.



The Metastability Problem

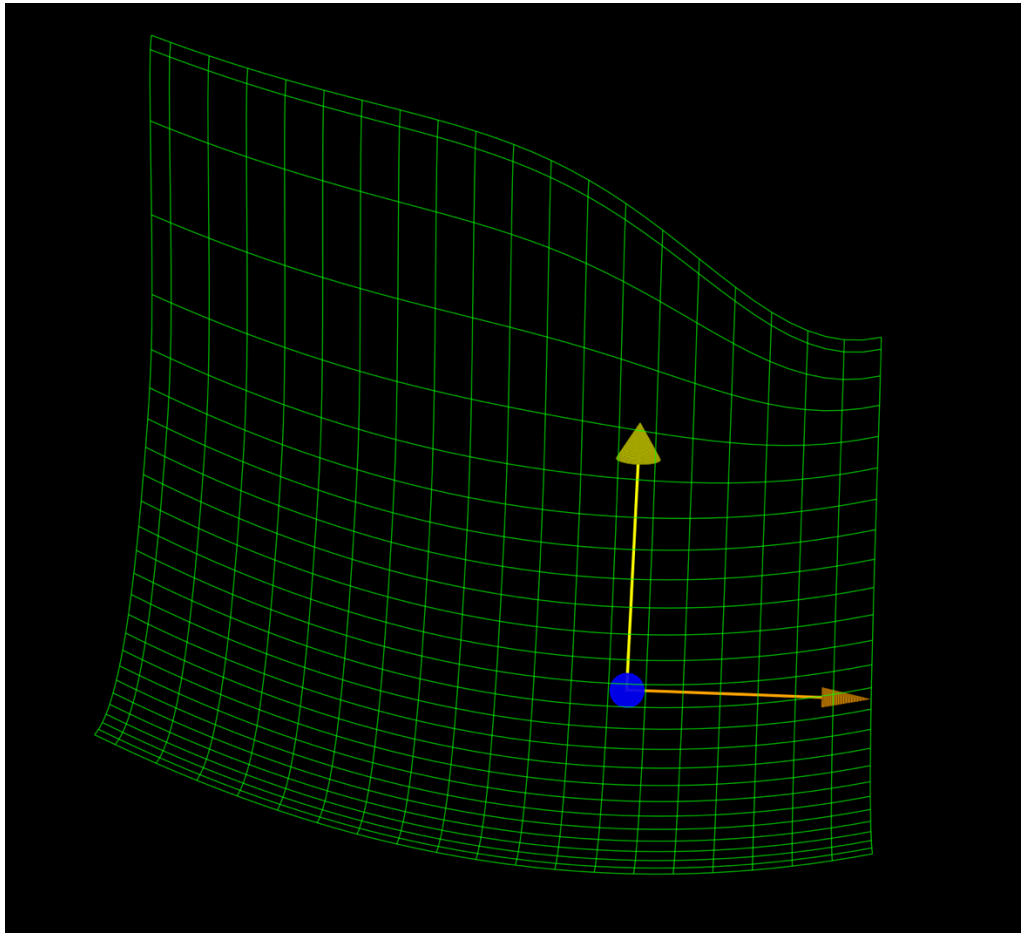
Symmetries Reveal Escape Directions

Core Idea : Near energy minima, some directions preserve energy better than others.

Mathematical Insight : Directions where $\delta E \approx 0$ correspond to approximate symmetries.

"Orange transform better symmetry compared to yellow."

Why this matters: The approximate symmetry also correlate with the slow dynamics. Motion along approximate symmetry largely due to random thermal fluctuations.



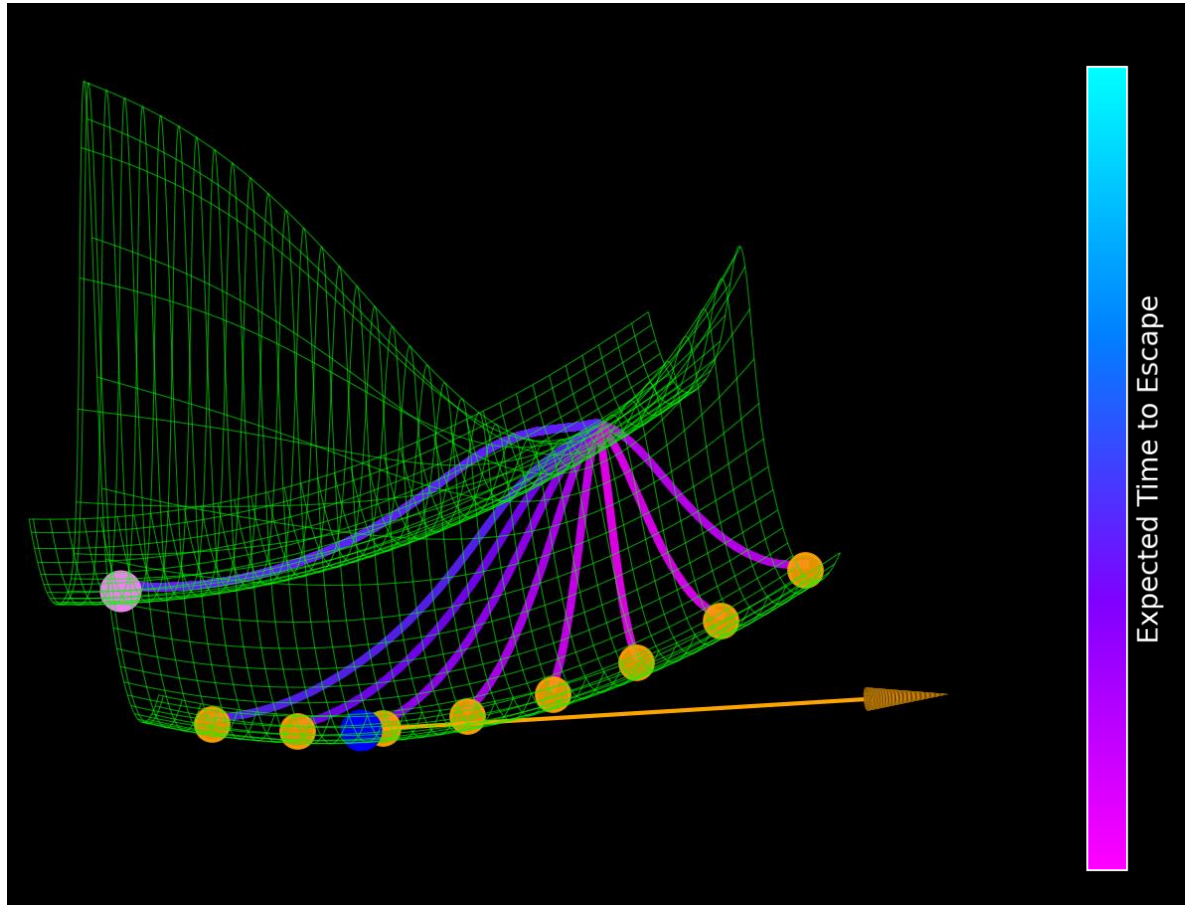
The Metastability Problem

Symmetry based Starting Points for Parallel Sampling

Acceleration Strategy : Multiple parallel simulations with starting points given by perturbing the local minima's configuration along discovered symmetries.

"Orange balls are starting points for simulation, perturbed blue along the orange arrow"

Key advantage: Instead of waiting for rare transitions, we sample from multiple strategic locations simultaneously. Speeds up slow mixing along orange arrow. Simultaneous simulations are "*embarrassingly parallel*".



Method Overview : Symmetry Discovery

Formalizing our intuition we define the following optimization problem for symmetry discovery. Using Lie Algebra direction L to define the transformations, we define symmetry loss at configuration x.

$$\textbf{Symmetry loss:} \quad \mathcal{L}(\boldsymbol{L}, x) = \left(\nabla E(x) \cdot \boldsymbol{L}x\right)^2$$

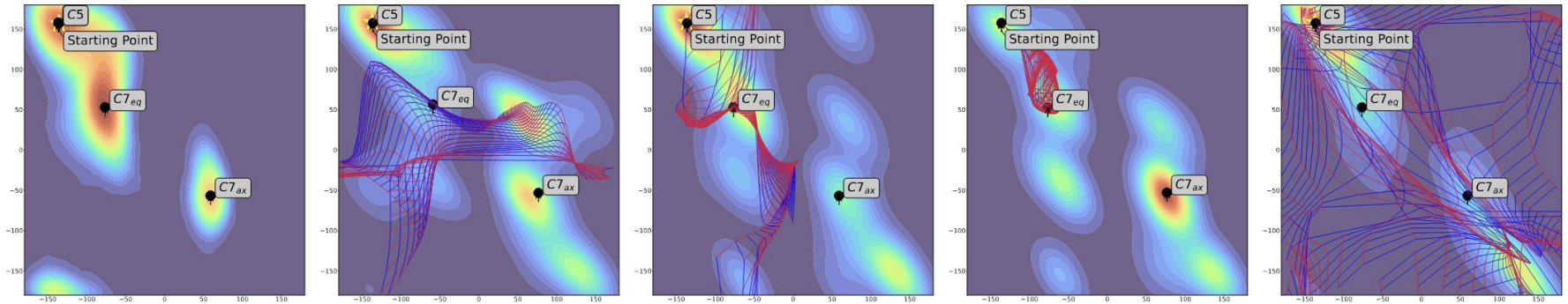
To avoid learning global symmetries, we constrain L to satisfy some invariances, yielding :

$$SE(3)\textbf{-invariant loss:} \quad \mathcal{L}(\boldsymbol{L}, x) = \left(\sum_{i,j,\mu} \frac{\partial E}{\partial x_j^\mu} \boldsymbol{L}_j^i x_i^\mu\right)^2 = \left(\mathrm{Tr}\left[(\nabla E)^\top \boldsymbol{L}x\right]\right)^2$$

Furthermore, using the Gaussian formulation of thermal noise, we optimize for L over gaussian perturbations of the minima x*. We consider two formulations 1) Direct Empirical Optimization and 2) Hessian Based Functional Optimization

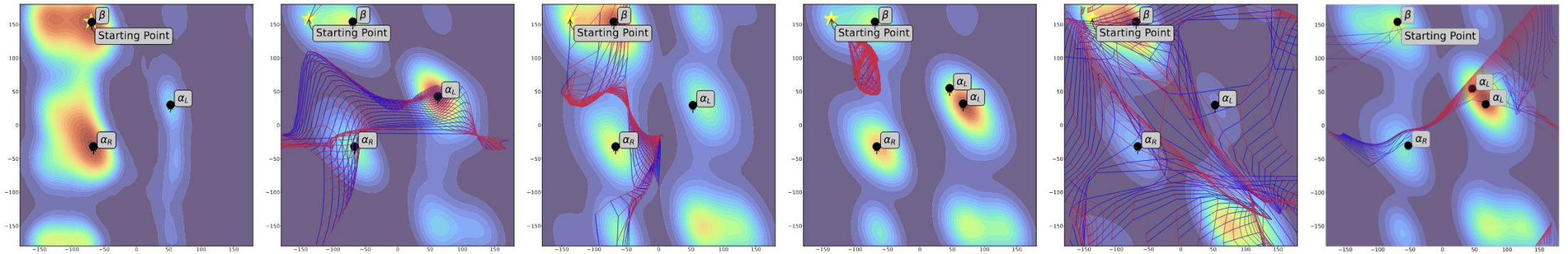
Experiment : Alanine Dipeptide

In Vacuum



(a) Long Simulation (b) Direct Optim (c) Full Hessian (d) Slow Hessian (e) Degen Hessian

In Solvent



(a) Long Simulation (b) Direct Optim (c) Full Hessian (d) Slow Hessian (e) Degen Hessian (f) Optim w Solvent

Sampling Efficiency Gain

Alanine Dipeptide (2 fs step size, 300 K, 1 ps^{-1} friction coefficient)

- Long simulation
  - 25 ns simulations needed for full coverage.
  - No tunable parameters.
- Replica Exchange Meta Dynamics (8 replicas, exchange attempted every 100 fs)
  - 1 ns / replica. Total effective 8 ns.
  - Diminishing returns with more replicas. Per replica time does not get much lower.
- Proposed Method (No optimization of grid size, 2d grid with 31 equidistant points)
  - All conformers found within <= 2ps simulation of starting points.
  - Total 961 simulations (total effective < 2 ns simulation time but embarrassingly parallel).