Detecting Rare Dynamical Events in Molecular Simulations Using Kinetic Modeling and Ensemble Anomaly Detection

Siya Jethliya, Department of Computer Science, Krea University | Supervisor: Prof. Chiranjib Sur

MOTIVATION

Rare dynamical events drive protein function but are extremely difficult to detect in large molecular simulation datasets.

- Molecular dynamics simulations produce high-dimensional time series describing protein motion, often containing millions of frames.
- Within these trajectories, biologically important conformational transitions occur rarely and transiently, making them difficult to identify.
- Traditional analysis methods focus on average structural fluctuations, which often obscure the dynamical signatures of rare events.
- As a result, extracting meaningful insights from large molecular trajectories requires computational approaches capable of learning system dynamics and detecting anomalous behaviour automatically.

COMPUTATIONAL CHALLENGE

Rare event detection in molecular simulations is fundamentally a machine learning problem.
Molecular dynamics trajectories form high-dimensional stochastic time series, where each frame encodes thousands of atomic degrees of freedom evolving over time.
Detecting rare dynamical events within these trajectories presents several computational challenges:
• High-dimensional configuration spaces with strongly correlated variables
• Long-range temporal dependencies between simulation frames
• Extreme class imbalance, where rare transitions occupy only a small fraction of the trajectory
• Absence of labeled ground truth for supervised learning
Consequently, identifying rare dynamical events requires unsupervised methods capable of modeling system dynamics and detecting deviations from typical behaviour.

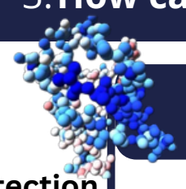
RESEARCH OBJECTIVE

Develop a computational framework for detecting rare dynamical events in molecular simulations.
The proposed framework combines machine learning and kinetic modeling to analyze high-dimensional molecular trajectories. It integrates:
• time-lagged dimensionality reduction to learn slow collective motions
• Markov State Models (MSMs) to represent conformational dynamics as a stochastic process
• ensemble anomaly detection using multiple dynamical signals
• interactive visualization to interpret anomaly signals in structural context

Together, these components transform large molecular simulation datasets into interpretable maps of rare dynamical behaviour and structural hotspots.

Research Questions

1. **How can rare dynamical events be detected in large molecular simulation trajectories?**
2. **Can kinetic modeling (tICA + MSM) reveal slow conformational dynamics and rare states?**
3. **How can anomaly signals be mapped onto molecular structures for interpretable analysis?**



Method

Machine Learning Pipeline for Rare Event Detection

Trajectory Representation

$$X = \{x_1, x_2, ..., x_T\}, \quad x_t \in \mathbb{R}^d$$

A molecular dynamics simulation produces a sequence of protein configurations over time.
Each frame represents a high-dimensional structural state describing atomic coordinates of the system.

2. Feature Representation

$$Y = \phi(X) \in \mathbb{R}^{T \times F}$$

Structural descriptors such as dihedral angles, residue contacts, and global shape metrics transform raw coordinates into a feature representation suitable for machine learning analysis.

3. Slow Dynamical Learning (tICA)

$$C_\tau w = \lambda C_0 w$$

Time-lagged Independent Component Analysis identifies slow collective motions by maximizing time-lagged autocorrelation.
The resulting components capture the dominant kinetic modes governing conformational transitions.

4.Kinetic Modeling (MSM)

$$P_{ij} = P(s_{t+\tau} = j \mid s_t = i)$$

The reduced trajectory is discretized into metastable states and modeled using a Markov State Model that estimates transition probabilities between conformations.

5.Ensemble Anomaly Detection

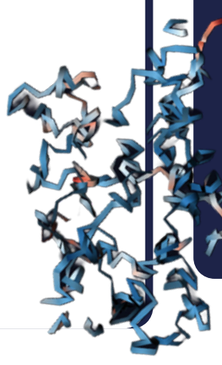
Rare dynamical events are identified by combining multiple anomaly signals derived from the kinetic model.
Frame-level anomaly score: $A_t = \sum_k w_k s_k(t)$
where each signal captures a different aspect of unusual dynamical behavior.

Signal Components

Rarity $R_i = 1 - \pi_i$

Transition surprise $S_{ij} = -\log P_{ij}$

Isolation — distance in tICA space
Flexibility — RMSF
Signals are normalized and combined to compute the anomaly score A_t.



Result & Discussion

Slow Dynamical Landscape

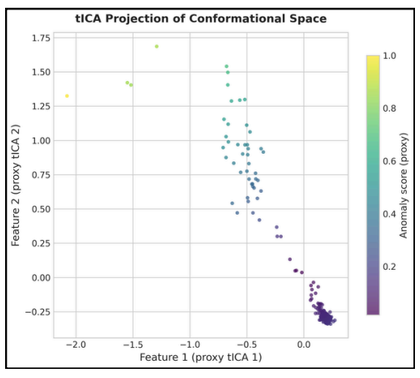
- The tICA embedding reveals the dominant conformational modes explored during the simulation.
- Dense clusters correspond to frequently visited metastable configurations.
- Sparse regions indicate transient conformations associated with rare structural transitions.

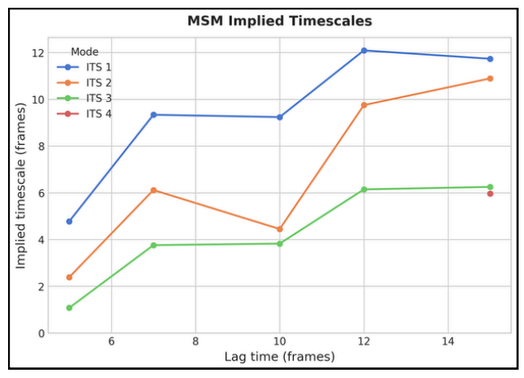
Kinetic State Structure

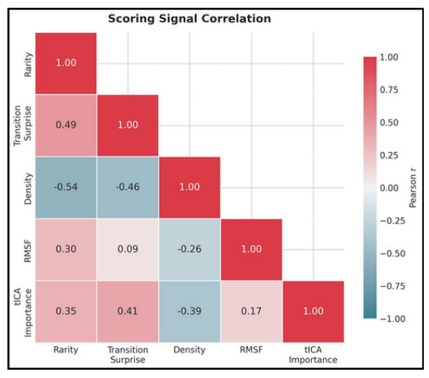
- Implied timescale analysis evaluates the stability of the kinetic model across lag times.
- The dominant relaxation timescales converge as lag time increases.
- This stabilization indicates that the learned state decomposition captures consistent long-timescale dynamics.

Dynamic Hotspot Detection

- The correlation matrix compares multiple anomaly signals derived from system dynamics.
- Moderate correlations indicate that each signal captures distinct dynamical properties.
- Combining these signals provides a more robust indicator of rare dynamical activity.







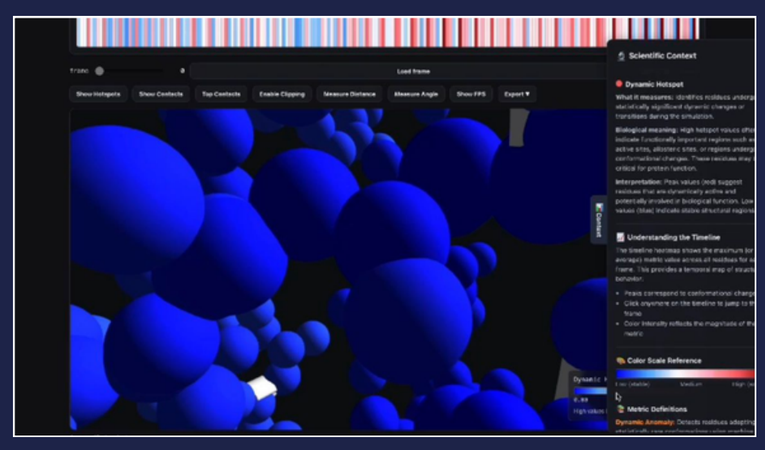
Interactive Visualization

An interactive molecular visualization tool was developed to explore simulation trajectories and interpret machine learning anomaly signals in structural context.

System Components
• Three.js GPU rendering for real-time molecular graphics
• Trame backend interface for synchronized data interaction
• Frame-by-frame trajectory navigation
• Residue-level anomaly coloring to highlight dynamic hotspots

Visualization Modes
• Point cloud — lightweight rendering of large trajectories
• Ribbon representation — global protein structure
• Ball-and-stick — detailed residue interactions

Interactive Analysis Tools
• anomaly metric selection (hotspot / anomaly score)
• timeline visualization of frame-level dynamics
• trajectory playback and frame navigation
• residue highlighting and structural contact analysis



CONTRIBUTIONS

• A computational pipeline for analyzing high-dimensional molecular dynamics trajectories

• A kinetic modeling framework combining tICA dimensionality reduction and Markov State Models

• An ensemble anomaly detection approach for identifying rare conformational transitions

• Residue-level hotspot mapping linking dynamical signals to protein structural regions

• An interactive molecular visualization system for exploring anomaly signals in simulation trajectories

CONCLUSION

This work introduces a computational framework for detecting rare dynamical events in molecular simulations.

By combining dimensionality reduction, kinetic modeling, and ensemble anomaly detection, the proposed pipeline identifies unusual dynamical behavior in high-dimensional trajectory data.

Integrating machine learning analysis with interactive molecular visualization enables structural interpretation of detected events, revealing residues and conformational transitions associated with rare dynamical activity.

The framework provides a scalable approach for exploring large molecular dynamics datasets and identifying functionally relevant structural fluctuations.