Approximation Theory and Special Functions
ATSF 2026 Conference - 9th Series
September 2-5, 2026, Ankara - Türkiye

Enhancing Feature Selection in Molecular Dynamics Trajectory Data for Hidden Markov Models

Elif Naz Fidancı[(1)] and Vilda Purutçuoğlu[(1,2)]

[(1)] *Middle East Technical University, Department of Biomedical Engineering, Ankara, Türkiye*
e-mail: elif.fidanci@metu.edu.tr

[(2)] *Middle East Technical University, Department of Statistics, Ankara, Türkiye*
e-mail: vpurutcu@metu.edu.tr

Abstract

Hidden Markov Models (HMMs) are well-established probabilistic models that have been successfully used in genomic and protein analyses to identify meaningful sequence patterns, functional regions, and binding sites associated with molecular recognition. However, their integration into molecular dynamics (MD) simulation analysis, particularly with systematic feature selection, represents a relatively novel application. This study integrates HMMs with MD trajectory data to identify hidden kinetic states and conformational transitions.

Although HMMs are effective for analysing complex, high-dimensional time-series data, their accuracy and interpretability may be limited by the initial selection of input features. Conventionally, all available features may be included without a specific feature-selection procedure. However, irrelevant, redundant, or highly correlated features can introduce noise and obscure physically meaningful molecular states. Therefore, this study aims to identify and retain only the most informative features before constructing the HMM.

A feature-selection pipeline combining LightGBM with SHAP is proposed. LightGBM models the relationships within the trajectory data, while SHAP evaluates the relative importance of individual features. The VAMP score is also employed as a complementary, kinetically informed feature-selection method. The selected features are subsequently used to construct the HMM.

The framework is evaluated using MD trajectory data from the WLALL pentapeptide, consisting of tryptophan, leucine, alanine, leucine, and leucine residues. Its performance is also compared with Time-Lagged Independent Component Analysis (TICA), a conventional dimensionality-reduction method.

The results show that the proposed framework is promising for MD trajectory analysis. The LightGBM–SHAP and VAMP approaches successfully identify important features while reducing redundant and noise-generating variables. This enables the HMM to detect physically meaningful kinetic states more clearly. Overall, integrating feature selection with HMM analysis provides a more interpretable and effective approach for understanding hidden conformational states and molecular transitions in MD simulation data.

Keywords: Molecular dynamics simulations, hidden Markov model, explainable feature selection.