PeerJ Physical Chemistry

Laplacian score and genetic algorithm based automatic feature selection for Markov State Models in adaptive sampling based molecular dynamics

Anu George, Madhura Purnaprajna and Prashanth Athri

Department of Computer Science & Engineering, Amrita School of Engineering, Bengaluru, Amrita Vishwa Vidyapeetham, India

ABSTRACT

Adaptive sampling molecular dynamics based on Markov State Models use short parallel MD simulations to accelerate simulations, and are proven to identify hidden conformers. The accuracy of the predictions provided by it depends on the features extracted from the simulated data that is used to construct it. The identification of the most important features in the trajectories of the simulated system has a considerable effect on the results.

Methods. In this study, we use a combination of Laplacian scoring and genetic algorithms to obtain an optimized feature subset for the construction of the MSM. The approach is validated on simulations of three protein folding complexes, and two protein ligand binding complexes.

Results. Our experiments show that this approach produces better results when the number of samples is significantly lesser than the number of features extracted. We also observed that this method mitigates over fitting that occurs due to high dimensionality of large biosystems with shorter simulation times.

Subjects Theoretical and Computational Chemistry, Biophysical Chemistry

Keywords Molecular Dynamics, Markov State Models, Laplacian scores, Computational Chemistry, Accelerated Molecular Dynamics

Submitted 23 December 2019
Accepted 27 May 2020
Published 22 June 2020

Corresponding author
Prashanth Athri,
prashanthathri@gmail.com

Academic editor
Ho Leung Ng

Additional Information and Declarations can be found on page 17

DOI 10.7717/peerj-pchem.9

INTRODUCTION

The study of biomolecular dynamics and interactions play a crucial role in understanding cell functions, and development of novel drugs (*De Vivo et al., 2016*; *Śledź & Caflisch, 2018*) Molecular dynamics (MD) simulations have the unique capability to track conformational changes, which in turn allows the establishment of target-specific and conformation-specific drug discovery research (*Lecina, Gilabert & Guallar, 2017*). The cost of bringing a drug to the market is still estimated to be 2.6 billion dollars (USD) (*DiMasi, Grabowski & Hansen, 2016*). Techniques like MD, and other computer-aided drug discovery (CADD) tools, reduce the burden of exorbitant drug discovery and development costs to healthcare. Various experimental methodologies such as NMR, fluorescence, X-ray crystallography and high-throughput screening are routinely used in drug discovery and development. While they are still the core to conducting drug discovery research, CADD tools can contribute to making the process rationally driven, reduce trial and error costs and allow accurate high

How to cite this article George A, Purnaprajna M, Athri P. 2020. Laplacian score and genetic algorithm based automatic feature selection for Markov State Models in adaptive sampling based molecular dynamics. *PeerJ Physical Chemistry* 2:e9 http://doi.org/10.7717/peerj-pchem.9

throughput explorations that are otherwise too expensive to conduct. CADD tools have proven to complement experimental techniques by making the discovery process more systematic and rational, and even opening up new search spaces previously unexplored. The past decade has even seen CADD tools contribute to upstream discovery research that resulted in full life cycle drug development programs, effectively translating research from the bench to the bed-side (*Leelananda & Lindert, 2016*).

MD is a low-throughput CADD technology that has vast applications (*Hollingsworth & Dror, 2018*), and central to that is the ability of MDs to provide time-continuous structural variation data of residues, co-factors, small molecules, etc., alongside the ability to clearly understand the modalities of interaction with the biomolecule in question in 3D. Experimental structures are rigid representations of the structure. MDs use Newtonian physics, chemical property parametrizations (*Alder & Wainwright, 1957*) to transform this static information to dynamic trajectories, which in turn provide advanced analytic and predictive capabilities to researchers. Many applications of MD simulations exist. For example, studies that explore conformational changes (*Flocco & Mowbray, 1995*; *Grant, Gorfe & McCammon, 2010*; *Lindorff-Larsen et al., 2011*; *Schwantes, Shukla & Pande, 2016*), binding-unbinding (*Fabritiis et al., 2008*; *Buch, Giorgino & De Fabritiis, 2011*; *Kohlhoff et al., 2014*; *Meyer et al., 2014*), and others that influence fundamental drug design in conjunction with experimental work (*Martinez et al., 2003*; *Namboori et al, 2010*; *Mohan et al., 2010*; *De Vivo et al., 2016*; *Childers & Daggett, 2017*). In this study, we explore the usage of Laplacian scoring based enhanced feature selection that can be employed in adaptive sampling MD. Below, we introduce each of the components briefly before we move to a detailed Methods section.

MDs provide a time-continuous trajectory, which comprises movement of atoms in microscopic, spatial and temporal detail. Frequencies of oscillation of atoms are in the range of femtoseconds, which is a limitation when the biological process being investigated require longer timescales. This is especially true in the context of all-atom calculations of MD (*Vanatta et al., 2015*; *Kohlhoff et al., 2014*). In recent times, all-atom, single, long simulations of few micro- to milliseconds are able to run on highly specialized hardware (*Shaw et al., 2014*), and supercomputers (*Stone et al., 2007*; *Dakka et al., 2018*). Microsecond range calculations are now possible even on workstations or server machines via GPU computing (*Salomon-Ferrer et al., 2013*) at costs that are orders of magnitude lower, as compared to specialized hardware or large compute clusters. However, the timescales of biological processes for molecules with millions of atoms, such as RNA Polymerase, are still out of reach (*Da et al., 2016*; *Wang et al., 2018*).

Apart from the above hardware-assisted acceleration, enhanced sampling techniques such as Replica-Exchange (*Sugita & Okamoto, 1999*; *Zhang et al., 2016*), Meta-Dynamics (*Zheng & Pfaendtner, 2015*), Transition path sampling (*Bolhuis et al., 2002*), Adaptive sampling MD (*Noé et al., 2009*; *Plattner & Noé, 2015*; *Pande, 2014*), high-throughput simulations (*Buch et al., 2010*; *Harvey & De Fabritiis, 2012*) and other methods (*Laio & Parrinello, 2002*; *Namboori et al, 2010*; *Vargiu et al., 2008*; *Tiwary & van de Walle, 2016*) have also proven to accelerate MD calculations. Adaptive sampling MD initiates many short parallel simulations from different starting conformers of the biomolecule, and repeats this in sequential iterations. Each iteration of the set of simulations that run in tandem, constitute an epoch. The starting structures in successive epochs are determined based on the analysis of MSM constructed, which is in turn calculated using features obtained from the previous epochs (*Doerr et al., 2016*). The final integrated trajectory is an aggregation of many epochs. Adaptive sampling MD, based on Markov State Models (MSM), is shown to explore under-sampled regions of the conformational space. This allows it to effectively overcome energetic barriers, and eventually sample optimal conformations (*Husic & Pande, 2018*; *Mittal & Shukla, 2017*) through MSM analysis. Identification of meta-stable states in the bio-dynamics and acceleration of simulation, depends on the efficiency of the MSM constructed. It is characteristic of MSMs based on optimal features to better identify hidden conformers (*Chen et al., 2018*). At the end of all the epochs, the MSM constructed is used to predict the characteristics of interactions. These may include protein-ligand binding free energy (*Noé et al., 2009*), transition rate between different protein conformations with possible binding sites (*Plattner & Noé, 2015*), hidden conformer states in protein folding (*Pande, 2014*), etc.

Examples of statistical learning techniques include Laplacian eigenmaps, spectral clustering, and Laplacian scores. In the context of applications to MD trajectory analysis, they have been successfully used to select features that can identify hidden conformers (*Sgourakis et al., 2011*). Also, Genetic algorithms (GAs), a class of meta-heuristics, are a good fit to automate the selection of optimal set of features to construct efficient MSMs (*Chen et al., 2018*; *Mittal & Shukla, 2017*). In our study, we explore a method for improving MSM efficiency, as measured by the established Generalized Matrix Rayleigh Quotient (GMRQ) scoring system (*Noé & Nuske, 2013*), by combining both Laplacian scoring (to provide a heuristic initialization) and GAs. As a result of this study, we identify features using this technique with a goal to construct better MSMs, as quantified by GMRQ scores. Various features can be used to discriminate between conformers, and features we have collected are detailed later. Time-structure Independent Component Analysis (TICA) (*Molgedey & Schuster, 1994*) is a dimensionality reduction method that is used in the MSM construction. The use of lesser number of features as compared to data samples, assures that the co-variance matrix used by TICA is a positive definite. This can be achieved by feature selection. This in turn avoids redundancy, and ensures a more accurate representation of the features by the TICA components. In summary, we propose the use of Laplacian scores in conjunction with GA for feature selection to produce improved MSMs. The goal of this study was to verify if the use of these two selection criteria, in conjunction, can lead to an efficient MSM. We worked on the hypothesis that this would lead to a more efficient selection of meta-stable states or hidden conformers for the next epoch of the adaptive sampling MD, and result in the acceleration of the overall simulation since low energy conformers are found rapidly.

Feature selection is an important step that avoids increased sampling and over-fitting. Automatic feature selection is being widely researched to construct efficient MSMs (*Tiwary & Berne, 2016*; *Shamsi, Cheng & Shukla, 2018*). Feature selection techniques, towards better MSMs, endeavor to select optimal sets of Collective Variables (CVs) (*Noé & Clementi, 2017*), or in other words, those that identify the most important residues with respect to the slowest processes. If all features are used, it results in poor clustering performance, which will adversely affect the MSM. Further, when the number of frames in simulation data is less than the number of features, it results in poor generalization of the MSM due to over-fitting (*Dy & Brodley, 2004*). This will necessitate the need for increased sampling at a higher computational cost (*Malmstrom et al., 2014*). However, the increased sampling is in direct opposition to one of the primary goals of accelerating the single long simulation through the use of short parallel simulations. Manual selection of CVs based on prior information about the system and human intuition have also been used to construct MSMs (*Lovera et al., 2012*; *Ahalawat & Mondal, 2018*). Nonetheless, this can lead to loss of information and very slow convergence in thermodynamic and kinetic property calculations. The automatic feature selection methods of highly-discriminant non-redundant features, like the one proposed in this study, provide a more generalized model with reduced computational complexity (*García & García-Pedrajas, 2018*).

In summary, and to highlight some of the principal points of the discussions above, in our approach, the initial feature selection is performed by Laplacian scoring (*He, Cai & Niyogi, 2006*; *Sgourakis et al., 2011*), and the model is optimized using a GA (*Chen et al., 2018*) on this subset. The time-structure independent components (tiCs) in TICA is used to identify the states in the system by grouping kinetically similar structures (*Schwantes & Pande, 2013*; *Pérez-Hernández et al., 2013*). During the MSM construction step, if the number of features is larger than the number of samples in the dataset, the co-variance matrix used for TICA is not guaranteed to be a positive definite. It means that some of the features can be expressed as a linear combination of the other features. This emphasizes the need for use of feature selection before dimensionality reduction to avoid over-fitting. Hence, in this study we use Laplacian scoring to pre-rank and select the features based on its importance. The ability of Laplacian score to identify only the significant features (*Sgourakis et al., 2011*), avoids the need for large amount of simulation data and over-fitting of data in the MSM constructed. Hence, these Laplacian scored feature subsets are used as a heuristic initialization of the zeroth population in the GA. The GA further improves the MSM efficiency, by adopting natural evolution strategies such as crossover and mutation (*Chen et al., 2018*; *Mittal & Shukla, 2017*). It is important to note that our method only provides an extra tool to perform effective GA-based searches for MSM based studies. We have shown that this improves the results in specific datasets used in this work. It can be seen as an adjunct tool that could be opportunistically used alongside previous techniques mentioned above.

MATERIALS AND METHODS

We have applied Laplacian scores driven winnowing, and the subsequent, GA based feature selection to improve the MSM models of the following complexes (previously studied, further detailed in subsection 'Datasets to evaluate the approach') Villin (PDB ID:2F4K), Fs Peptide (Ace-A_5(AAARA)_3A-NME), WW domain (PDB ID: 2F21), Piperidine-Thrombin with Thrombin taken from PDB ID: 3D49 and Benzamidine-Trypsin (PDB ID:3PTB). This work was carried out on the Intel vLab Knights Landing cluster that

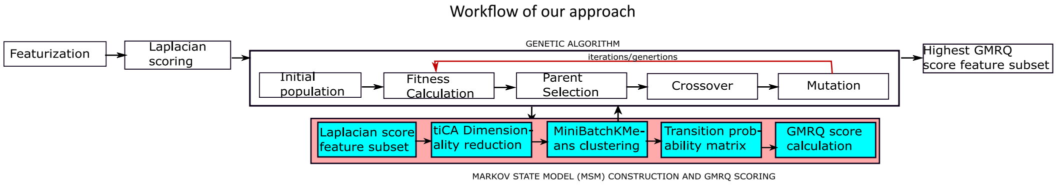
Workflow of our approach
Featurization
Laplacian scoring
GENETIC ALGORITHM
iterations/genertions
Initial population
Fitness Calculation
Parent Selection
Crossover
Mutation
Highest GMRQ score feature subset
Laplacian score feature subset
tiCA Dimensionality reduction
MiniBatchKMeans clustering
Transition probability matrix
GMRQ score calculation
MARKOV STATE MODEL (MSM) CONSTRUCTION AND GMRQ SCORING

Figure 1 **The proposed methodology for the identification of the optimal features.**

Full-size DOI: 10.7717/peerjpchem.9/fig-1

consists of 252 Intel Xeon-Phi nodes connected with Omni-Path (Intel OPA) 100 Series interconnect and academic desktop with Intel Core i7-7500U @2.70 GHz with Nvidia GPU GTX TITAN Black @889 GHz. An outline of MSMs (see *Husic & Pande, 2018*) for a detailed review of MSM), and the application of MSM in adaptive sampling MD follows. We have inherited this part of the workflow from previous studies. Finally, we also provide details about the additional components of our workflow.

Markov State Models (MSM)

MSMs are used to model randomly changing systems. They have been used to study time-series events in which the current state depends only on the previous state and not on any other states before it (*Tang, Bevan & Grover, 2017*). It has a wide variety of applications in the field of physics, chemistry, medicine, finance, management etc. It has been used for biological modelling such as the simulation of brain function (*George & Hawkins, 2009*), viral infection of single cells (*Gupta & Rawlings, 2014*), analysis of bacterial genome (*Skewes & Welch, 2013*), as well as MD simulations using adaptive sampling techniques (*Doerr et al., 2016*; *Doerr & De Fabritiis, 2014*) etc.

Figure 1 represents the workflow adopted towards the application of MSM to analyze MD trajectories of bio-molecules. The goal is to find meta-stable conformers, with a few variations as explained below. Prediction accuracy of MSM, as indicative by GMRQ score, can be improved by selecting optimal features from the simulation trajectory data.
MSM is represented by an Nx N matrix which gives information about the probability of transition between N states (*Singhal, Snow & Pande, 2004*; *Noé & Fischer, 2008*). In Fig. 2, x_{ij} is the average of transition count matrix and its transpose. The probability of each row in the average transition count matrix is calculated to obtain the MSM. From the MSM, the eigen-flux is calculated. Eigen flux (-n1 for s4, n2 for s3, n3 for s2 and n4 for s1) of each state shows the transition from source state to the sink state, and this is used to identify the slowest process (*Beauchamp et al., 2011*). The N states in the MSM represents the state decomposition into which the simulation data is clustered. The ideal MSM, or in other words, the ideal state decomposition reveals even the slowest dynamical process. In this specific use case, it is the conformational change occurring in the simulated system. The probability with which transition to a state j occurs is calculated as

$$p_j(t+\tau)=\sum_{i=1}^{N} p_i(t)T_{ij}(\tau) \qquad (1)$$

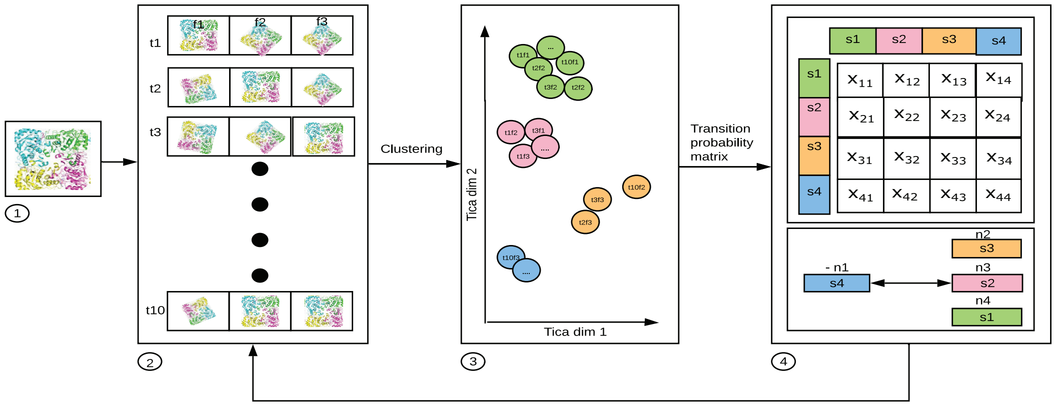

Figure 2 **1. Starting structure (frames f) of the bio-molecule used for parallel simulation. 2. 10 (arbitrary number for illustration) parallel trajectories (t1 through t10) produced by MD software, such as ACEMD, GROMACS,AMBER, etc. 3. The frames per structure, having reduced to 2 dimensions using TICA. 4. MSM constructed with each of the values in a cell of the matrix, x_{ij}, where i is the source state, and j is the destination state showing the probability of transition. from i to j.** The figure has been drawn based on Fig. 1 in *Husic & Pande (2018)*.

Full-size DOI: 10.7717/peerjpchem.9/fig-2

It can be generalized as

$$p(t+\tau)=\sum_{i=1}^{N}p(t)T(\tau) \tag{2}$$

where $p_j(t+\tau)$ is a population vector after time τ; τ is lag time; $p_i(t)$ is a population vector at time t; $T_{ij}(\tau)$ is the probability of transition from state i to state j; $T(\tau)$ is the probability transition matrix that represents the MSM.

The eigenvalues λ_i and eigenfunctions ψ_i of $T(\tau)$ can decompose the transition probability matrix (*Shukla et al., 2015*) as in Eq. (3). The eigenfunctions represents the slowest processes that occurs in the simulated system.

$$T(\tau)\circledast\psi_i=\lambda_i\psi_i \tag{3}$$

Generalized Matrix Rayleigh Quotient (GMRQ) derived from the variational principle of conformation dynamics (*Noé & Nuske, 2013*) is a metric used to evaluate the efficiency of the MSM (*McGibbon & Pande, 2015*). MSMs with higher GMRQ scores provide higher discriminatory abilities towards identifying the slowest dynamical process occurring in the system (*McGibbon & Pande, 2015*). Variational principle is used in the calculus of variations, or changes to develop a function that maximizes the variables dependent on it. Based on variational principle, the eigen values and eigen functions are estimated for the MSM constructed using the simulated data. GMRQ score of the MSM is the sum of the first p eigenvalues of $T(\tau)$. They also identify the slow processes, p. The upper boundary for the GMRQ score is set by variational principle as shown below (*Husic & Pande, 2017*; *Noé & Nuske, 2013*):

$$GMRQ\equiv\sum_{i=1}^{p}\hat{\lambda}_i\leq\sum_{i=1}^{p}\lambda_i \tag{4}$$

where $\hat{\lambda}_i$; λ_i are estimated and real i-th eigenvalues respectively. The efficiency of MSM is improved by trying to reach the upper boundary set by the sum of real eigenvalues (λ_i). MSMs with higher GMRQ scores are better able to identify the slowest conformational change in the system (*Noé & Nuske, 2013*; *McGibbon & Pande, 2015*).

Adaptive sampling MD based on MSM and Optimized feature selection for MSM

We have used simulation trajectories that are produced by adaptive sampling MD, based on the MSM. In this protocol, the first step in the construction of MSM involves the extraction of feature, namely, dihedral angles, distance between amino-acid residues, root mean square deviation (RMSD) from the raw Cartesian coordination, etc. This step is called featurization. The selection of optimal features or collective variables have a critical impact on the thermodynamic and kinetic property calculations predicted by the model, and is explored in many studies (*Shamsi, Cheng & Shukla, 2018*; *Mittal & Shukla, 2017*; *Sultan et al., 2014*). Consequently, using the optimal feature set for the construction of MSM affects its prediction accuracy. Identification of the most critical dihedral or contact features assist in identifying the optimum conformers. These conformers are then used as starting structures for the next epoch of adaptive sampling in MSM based adaptive sampling MD. Appropriate feature selection circumvents issues due to energy barriers leading to convergence of thermodynamic properties (such as free energy of binding in the case of protein-ligand binding simulation), or folding of the unfolded protein (*Doerr & De Fabritiis, 2014*).

Feature selection is an unsupervised machine learning task. One method of categorizing them identifies three sub-classes, namely, filters (*Devakumari & Thangavel, 2010*; *Mitra, Murthy & Pal, 2002*; *Tabakhi & Moradi, 2015*), wrappers (*Dy & Brodley, 2004*; *Dutta, Dutta & Sil, 2014*; *Breaban & Luchian, 2011*), and hybrids (*Solorio-Fernández, Carrasco-Ochoa & Martínez-Trinidad, 2016*; *Li, Lu & Wu, 2006*). In filter methods, the importance of a feature is studied based on the intrinsic properties of the data. They are not selected by training on the model on which it is to be used. In wrapper-based methods, the features are first used to train the model. Then their impact on the performance of the model is evaluated. Filter methods are computationally less expensive than wrapper methods. Hybrid methods combine the advantages of both filter and wrapper methods (*Li et al., 2008*). In this method, the feature subset works best when the same model used for feature selection is used subsequently (*Solorio-Fernández, Carrasco-Ochoa & Martínez-Trinidad, 2020*). Laplacian-score based feature selection is a filter-based method. This method finds the features that have the power to preserve the clusters in the data. Features that can maintain the structure of the plotted nearest neighbor graph are selected.

Laplacian scores have been used to reveal hidden structures in the complex conformational space of the intrinsically disordered peptide, Aβ(1-42) (*Sgourakis et al., 2011*). Laplacian scores identified crucial interactions in these conformers, which provided new drug design hypothesis that can, potentially, be used to discover drugs that inhibit the oligomers and fibril formation critical to the progression of Alzheimer's disease. In this study, inspired by the success of the study mentioned above, the initial population of a GA's chromosomes have genes pre-filtered through Laplacian scores-based ordering. In bio-molecules, conformational changes happen in localized regions (*Fan et al., 2015*). Laplacian score ranks the features based on its ability to preserve the local structure of the graph constructed based on the k-nearest neighbor algorithm (*He, Cai & Niyogi, 2006*). Due to this property of the Laplacian score, it is used to identify a subset of the features. The feature subset selection is further optimized through the GA.

Datasets to evaluate the approach

Molecular dynamics trajectories of five systems are used to validate our approach by analyzing the kinetic and thermodynamic properties predicted by the MSMs. The features selected were derived from the GA run, whose initial population was based on our Laplacian score ranking scheme. The simulation dataset of the two protein-ligand complexes are Piperidine-Thrombin *Doerr et al. (2016)* and Benzamidine-Trypsin *Scherer et al. (2015)*. The protein folding trajectories of Fs Peptide (*Beauchamp et al., 2011*), WW domain (*Lindorff-Larsen et al., 2011*) (generously shared by D.E Shaw Research), and Villin simulation *Doerr et al. (2016)* were also analyzed as a part of this study.

Our approach

Figure 1 provides a detailed view of the protocol we have implemented, while the principal choices for each component of the workflow are shown in Table 1. The hyperparameters for the protocol were chosen based on the published research, HTMD and Msmbuilder tutorial (*Chen et al., 2018*; *Doerr et al., 2016*; *Beauchamp et al., 2011*). Scikit-learn (*Pedregosa et al., 2011*) was used for feature preprocessing, and to calculate the Laplacian scores, IPython (*Pérez & Granger, 2007*) for the execution of the python scripts in the form of a pipeline. MDTraj (*McGibbon et al., 2015*) was used to analyze the simulation trajectories, and HTMD (*Doerr et al., 2016*) was used for the construction of free energy heatmaps, standard free energy plots and cktest graphs for the MSMs. These graphs were constructed for selected residues to implement our approach, and for all residues to re-implement previous work (for comparison). MSMBuilder (*Beauchamp et al., 2011*) was used for MSM construction and GMRQ scoring.

- **Featurization**: Featurization is performed to transform rotational and transitional motion from the simulation data to a vector. Dihedral angles and distance between CA atoms of amino-acid residues, from the Cartesian coordinates were extracted as features from the simulation data (for the 5 bio-systems). Dihedral featurization comprises backbone and side chain dihedral angles, along with sine and cosine of these angles (*Beauchamp et al., 2011*). All of the above are calculated for each frame in the simulation trajectory. In our approach, we used the sine and cosine of back bone dihedral angles for dihedral features. It has been proved to be one of the important metrics that is able to identify different conformers in an MD trajectory (*Cossio, Laio & Pietrucci, 2011*). Contact featurization calculates the distance between each pair of amino acid residue, or the distance between the specified pair of residues for each frame (*Beauchamp et al., 2011*). In this study, we implemented contact featurization as a vector of distances between CA atom of each amino acid residue and ligand for the protein-ligand systems,

Table 1 **Attributes used in each stage of the proposed approach.**

Featurization: Types

- Dihedrals- measured as phi and psi angles of respective amino acid residues
- Contacts- measured as distance between CA atoms of amino acid residues

Feature Selection Method

Genetic algorithm with initial population selected using Laplacian scored features

Dimensionality Reduction Method

TICA algorithm with 4 components and lag time of 2 ns

Clustering Method

Mini-Batch K-means with 200 clusters

MSM Model- Hyperparameters

- N_timescales = 5 Lag time (ns) = 50
- Scoring using GMRQ

Genetic Algorithm- Hyperparameters

- Crossover Probability = 0.8
- Mutation Probability = 0.2

and the distance between the CA atoms of amino acid residues for the protein-folding simulation data.

- **Laplacian scoring**: An average, across all trajectories, of the Laplacian scores obtained for each feature is calculated. The features were then sorted based on this average Laplacian score (of each feature). This ordered list is used to build the initial population for the GA. Laplacian Score (LS) is fundamentally based on Laplacian Eigenmaps and Locality Preserving Projection. The algorithm for calculating the Laplacians (*He, Cai & Niyogi, 2006*) is given below. Let L_r denote the Laplacian score of the rth feature. Let f_{ri} denote the ith sample of the rth feature; where $i = 1, \ldots, m$.
  1. Construct a nearest neighbor graph—G with m nodes—. The ith node corresponds to x_i. We connect the edge between nodes i and j, if x_i and x_j are close, i.e., x_i is among k nearest neighbors of x_j, or x_j is among k nearest neighbors of x_i.
  2. The weight matrix S of the graph, models the local structure of the data space. $\mathrm{S}_{ij} = \mathrm{e}^{\frac{(x_i x_j)^2}{t}}$, when nodes i and j are connected, and $\mathrm{S}_{ij} = 0$ otherwise.
  3. For the r^{th} feature, we define: $f_r = [f_{r1}, f_{r2}, \ldots, f_{rm}]^T, D = diag(S\mathbf{1}), \mathbf{1} = [1, \ldots, 1]^T, L = DS$ where the matrix L is called graph Laplacian (*Chung, 1997*). Let
  $$\tilde{f_r} = f_r - \frac{{f_r}^T D\mathbf{1}}{\mathbf{1}^T D\mathbf{1}}\mathbf{1} \tag{5}$$
  4. The Laplacian score of the r^{th} feature is:
  $$L_r = \frac{{f_r}^{-T} L\tilde{f_r}}{{f_r}^{-T} D\tilde{f_r}} \tag{6}$$
- **Genetic Algorithm (GA)**:
  - Initial population - Features are shortlisted based on Laplacian score. To avoid over-fitting or under-fitting, we use the thumb rule of selecting $\frac{1}{10}$th of the number

of samples (*Sánchez & García, 2018*) as the target number of features. This subset is then used as the initial population for GA. Under-fitting occurs when the number of features that represent the model are significantly lesser than the number of data points available. In MD centered protein systems, the number of features or the dimensionality is high due to the nature of the problem, and thus probability of under-fitting is negligible.

  - Fitness calculation - The fitness of the feature subset is scored based on GMRQ score of the respective MSM. Dimensionality reduction is performed using TICA (*Pérez-Hernández et al., 2013*). Dimensionality reduction produces a linear combination of the features, and the top components (called time structure-based independent components or tICs) capture the most prominent processes in the simulation. Further, clustering of frames is performed using MiniBatchKmeans clustering (*McGibbon & Pande, 2015*). The MSM is constructed using the clusters identified and the transition probability between these clusters. The number of clusters, lag time of TICA and MSM, number of components used for TICA and n_timescales which represents the timescale of the n slow processes is presented in Table 1. K-fold cross-validation ($K = 5$) is used to avoid over-fitting. The value of GMRQ score for the MSM is calculated based on the k-fold cross validation.
  - Selection - Selection of parents is performed for the production of off-springs for the next generation. This is performed based on the fitness of the feature subset. This helps in the selection of ideal traits, which in this case is the important CVs.
  - Crossover - The crossover probability decides the number of parents taking part and, in effect, the number of off-springs generated. We have used a single point crossover.
  - Mutation - It is performed to increase the gene variations in each of the generations. Single point mutation is carried out, in which a randomly selected gene is replaced with the highest ranked feature based on its Laplacian score, which is not already present in the chromosome.
  - Repeat the above four steps for the specified number of generations

- **MSM construction and GMRQ scoring**: The MSM is constructed using the most optimal CVs obtained after the GA. This is then compared with MSM constructed using all the features in terms of the GMRQ score, and also compared using the implied timescale. The methodology is summarized in the Fig. 1

Details on Bio-systems studied

Five systems were analyzed in this work (The original complexes have been uploaded in figshare).

- Benzamidine-Trypsin (*Doerr et al., 2016*) : This work studies the binding of serine protease beta-trypsin to inhibitor benzamidine. It comprises 10 microseconds of simulation data, with component trajectories of 200 nanoseconds each.

- Piperidine-Thrombin (*Scherer et al., 2015*) : Thrombin is a serine protease that acts in the coagulation pathway to convert factor XI. The simulation dataset of Piperidine-Thrombin comprises 810 trajectories of 200 nanoseconds each, resulting in a cumulative simulation length of 162 microseconds.
- Fs Peptide (*Beauchamp et al., 2011*) : Fs peptide is a system widely used to study intricacies of protein folding. The simulations were carried out using the AMBER99SB-ILDN force field with OpenMM 6.0.1. The Fs peptide dataset includes 28 trajectories of 500 nanoseconds each, totalling to 14 microseconds of data.
- Villin(*Doerr et al., 2016*) : Villin is an actin binding protein. The Villin dataset consists of 1,374 trajectories of 100 nanoseconds each and total simulation data of length 137.4 microseconds.
- WW domain (*Lindorff-Larsen et al., 2011*) : WW domain is a protein domain that plays an important role in the interaction between protein ligands. The WW-domain dataset consists of 325 trajectories with a total of 1,137 microseconds.

RESULTS

The effectiveness of MSMs that use features selected through GA with Laplacian-based initialization is compared to MSMs that used all features. The metrics used are GMRQ scores, and the implied timescales obtained for the slowest process for the two different sets of features. Recall that dihedrals and contacts are used to generate the feature set. The features selected by our approach in systems where we saw an improvement MSM metrics are shown in Fig. 3.

Among the important residues identified for Villin, are 47 and 69. These are two of the most important hydrophobic residues that play an important role in the folding of Villin (*Frank et al., 2002*). In the Benzamidine-Trypsin system, residues 79, 175, 180, 190 are observed to play important roles in ligand-binding process (*Buch, Giorgino & De Fabritiis, 2011*) and calcium binding loop (*Plattner & Noé, 2015*). Indeed, these were also identified as important residues using our workflow. This validates that the proposed approach is able to identify residues important to the biological process being studied.

Comparison of GMRQ scores

In this section, we have compared the GMRQ scores of MSM models obtained when the full set of dihedral angle-based features and the full set of features calculated using all contacts are considered, against the ones where we use only selected features using our approach. The summary of these results are shown in Table 2. For the Benzamidine-Trypsin system, the model with select features has a higher GMRQ score of 3.21 and 2.59. The GMRQ scores of MSMs constructed for Villin and Fs Peptide systems with reduced set of dihedral and contact features are both higher by 19.15%, 19.95%, 17.44%, 11.25% respectively, as compared to when constructed with all the features. This indicates that the features identified by our method is able to capture the most important features and reduce the noise by avoiding unwanted features in certain cases. However, in the case of Piperidine-Thrombin and WW-domain systems, the GMRQ score of the MSM constructed with reduced set of selected features is less than the score obtained with all the

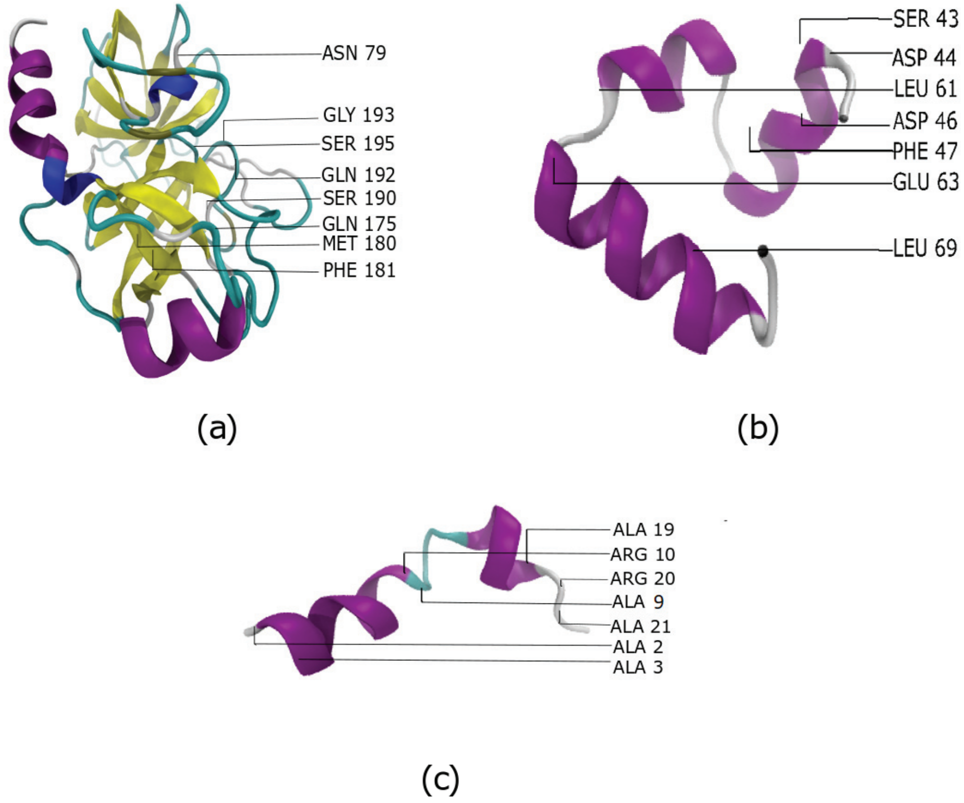

Figure 3 The most important features identified by our approach for three systems (A) Benzamidine - Trypsin (B) Villin and (C) Fs Peptide.

Full-size DOI: 10.7717/peerjpchem.9/fig-3

features selected. This may be due the fact that the interactions captured by the reduced set of features is not enough to capture the slowest process that occurs in the respective simulation data.

The Table 3 shows that our approach is able to mitigate model inaccuracies caused by the over-fitting that occurs when the length of simulation (and hence, number of observations) is relatively shorter, while the number of features are higher. The variance in the GMRQ score for the model with selected residues is 5 to 10 times lesser than when all residues are used for MSM construction. This shows that the MSM constructed using selected residues is able to generalize better, and overcome over-fitting issues. The WW-domain is not given in the Table 3 as simulation length of 1,137 microseconds is adequate, and hence, over-fitting does not occur.

Some studies use the Chapman–Kolmogorov property (*Noé et al., 2009*; *Prinz et al., 2011*; *Bowman, Pande & Noé, 2013*) as a measure of self-consistency of individual MSMs. Nonetheless, since our goal was to find the optimum MSM that identifies the slowest implied timescale, we have used GMRQ scores for the primary analysis. The analysis of Chapman–Kolmogorov test is given for the three systems that produced better results (see Table 2 for the ranking) through our approach is provided in form of figures in the (Figs. S1 through S3).

Table 2 **Comparison between GMRQ score of different features and the efficiency of MSM improved with feature selection using Laplacian score and GA.**

System	Feature	GMRQ score (all) used as Benchmark	GMRQ score (selected with Laplacian score and GA)
Benzamidine-Trypsin	Dihedral	3.54	3.21
	Contacts	1.87	2.59
Piperidine-Thrombin	Dihedral	5.68	5.65
	Contacts	5.25	5.09
Villin	Dihedral	3.76	4.48
	Contacts	3.71	4.45
WW-domain	Dihedral	4.66	4.46
	Contacts	5.29	5.23
Fs Peptide	Dihedral	4.53	5.32
	Contacts	4.8	5.34

Table 3 **Compares the variance in the GMRQ score when all features are used and features selected using Laplacian with GA is used.** This shows that using selected features helps to avoid over-fitting.

System	Variance in GMRQ score (all features)	Variance in GMRQ score (selected with Laplacian score and GA)
Benzamidine-Trypsin	0.27	0.04
Piperidine-Thrombin	0.06	0.005
Villin	0.42	0.08
Fs Peptide	0.92	0.18

Standard free energy prediction

The free energy surface is indicative of the thermodynamic and kinetic properties of the system. The free energy surface and the standard free energy plots of the meta-stable states, identified as macro-states in MSMs, of the Villin system are given in Fig. 4. One of the main advantages of an MSM is its ability to predict the thermodynamic and kinetic properties of the system. The meta-stable states identified by the MSM model map to different conformers of the bio-molecule in the simulation. These are identified as different points on the free energy surface. The stable, lowest energy state is the bonded state in protein-ligand binding simulation, and folded state in protein folding simulation. For the sake of brevity, the same for the other two high scoring MSM models, namely, Benzamidine-Trypsin system, Fs Peptide and Villin (dihedral features only) are given in Figs. S4 to S6.

The free energy surfaces of the Villin system for full-set and selected of features are given in Figs. 4A and 4C respectively. The standard free energy of each of the meta-stable states

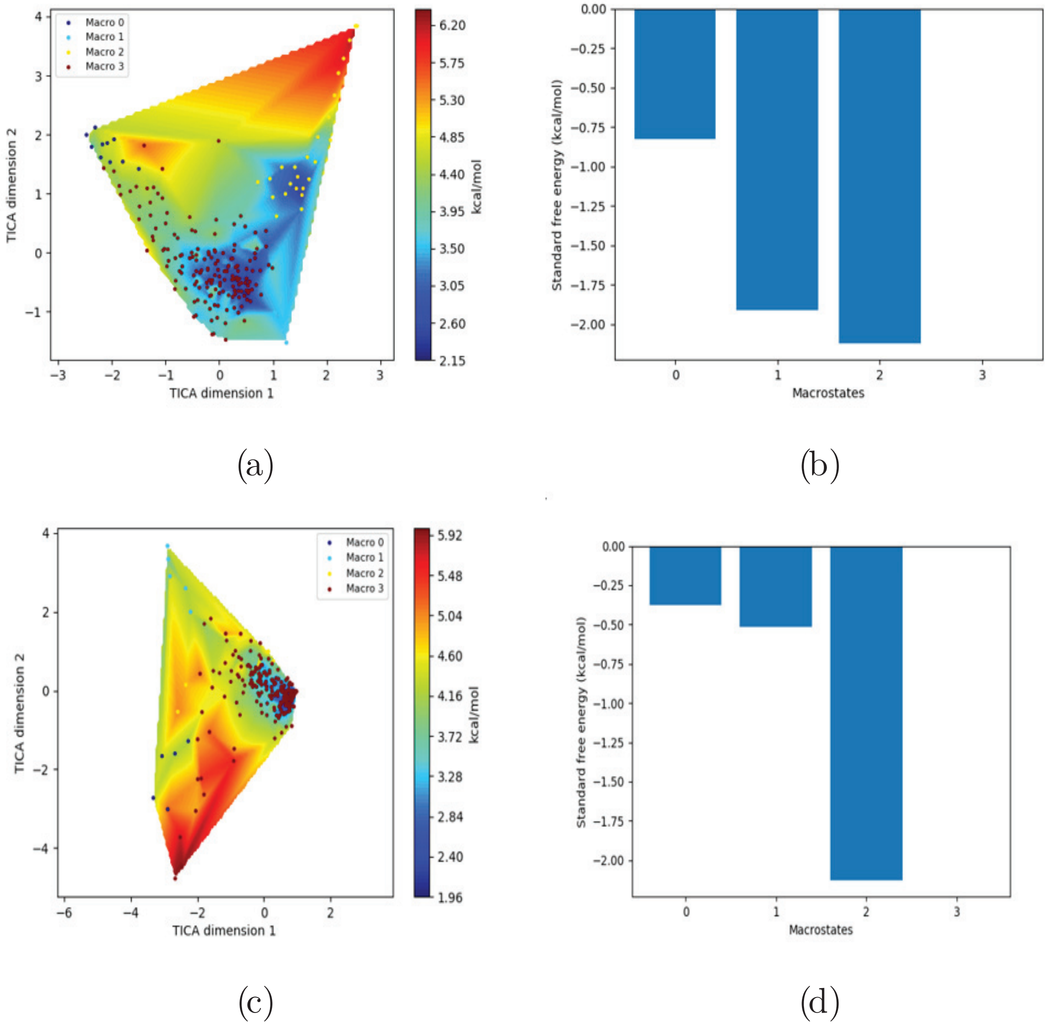

Figure 4 Free energy surface projected on the first two TICA dimensions and the corresponding standard free energy of meta-stable states of Villin for the MSM models. (A) Free energy surface of MSM constructed with all contacts, (B) Standard free energy of meta-stable states plotted in (A), (C) Free energy surface of MSM constructed with selected with selected contacts, (D) Standard free energy of meta-stable states plotted in (C).

Full-size 🖼 DOI: 10.7717/peerjpchem.9/fig-4

identified in the corresponding free energy heat map are shown for full-set and selected of features in Figs. 4A and 4D. Figure 4A shows four meta-stable states on the free energy surface that was plotted using all the contact features. The standard free energy of each of the meta-stable states identified in Fig. 4A) is shown in Fig. 4B). The lowest energy of −2.29 kcal/mol, observed for the meta-stable state 2, is identified as the sink state by the MSM model. Similarly, four meta-stable states are identified (Fig. 4C) for the MSM model that used contact features between selected residues that were identified by our approach. The standard free energy of each of the meta-stable states is plotted in Fig. 4D, and lowest energy of −2.31 kcal/mol is observed for the meta-stable state 2, identified as the sink state

Table 4 **Comparison of implied timescale between different features and the efficiency of MSM improved with feature selection using Laplacian score and GA.**

System	Feature	Implied timescale in ns (all features) Benchmark	Implied timescale in ns (selected with Laplacian score and GA)
Benzamidine-Trypsin	Dihedral	3,980.45	1,069.21
	Contacts	5,571.51	14,082.11
Piperidine-Thrombin	Dihedral	2,453.09	1,530.06
	Contacts	5,017.45	4,055.23
Villin	Dihedral	1003.65	1,725.33
	Contacts	788.59	1,169.06
WW-domain	Dihedral	7,934.28	7,101.21
	Contacts	8,679.26	8,535.81
Fs Peptide	Dihedral	1,515.48	2,034.63
	Contacts	1,943.93	2,198.82

by the MSM model. The lower standard energy stable state (−2.31 kcal/mol) is identified by the MSM model constructed using selected residues identified by our approach.

Implied timescale comparisons

Implied timescale values refer to the time taken by the slowest processes captured by MSMs. Table 4 shows that the MSMs constructed using the selected dihedral and contact features of Benzamidine-Trypsin, Villin and Fs Peptide are able to capture slowest processes. Since our goal is to capture the slowest timescale that represents the folded or the protein-ligand bonded state, the other four slower timescales is given in the Fig. S7, for the interested reader. Five timescales are compared for systems with better MSMs as identified by our approach, namely, Benzamidine-Trypsin, Fs Peptide and Villin.

The MSM constructed using the full-set of all the contacts with the ligand, in the Piperidine-Thrombin system, is able to capture the slower process. Nonetheless, it can be seen that, in this case, GMRQ score for the model from the reduced set is not significantly higher against the one with the full-set of contact-based features.

DISCUSSION

Optimal feature selection to identify an optimal set of collective variables, in other words, residues of the simulated bio-molecules, ensures that the co-variance matrix calculated during dimensionality reduction phase is a positive definite matrix. Additionally, avoiding over fitting, due to the lack of sampling data relative to a large number of features in each frame of the samples is highly desirable. In this work, we show that enhanced feature selection using Laplacian scoring addresses both these issues i.e., the requirement of increased sampling and avoiding over fitting.

In the protein-ligand system, Benzamidine-Trypsin, the contact features between selected residues (Fig. 3A) are able to identify the slowest processes correlated to the binding of Benzamidine to Trypsin, better than when all the features are selected. This may be because using all the contact features adds noise to the model when the MSM is constructed with all contacts. Results also show that contact features are relatively more discriminant towards identifying the slowest processes, as compared to dihedral features. This suggests that researchers should empirically check the set of features that best fit the system in hand, when using Laplacian scoring. In this system, the number of features (for example, 888 dihedrals) are very large as compared to the number of frames (200). Our proposed methodology, allows the use of selected features for construction of efficient MSM, and thus shows that this approach is useful when the size of the molecules are larger, and have smaller length simulations.

In the protein folding systems used, when the amount of simulation data is more than the number of dihedral and contact features extracted, the GMRQ score of MSM constructed with all the features is higher than when using a reduced set of features selected by our approach. The GMRQ score of MSM constructed for WW-domain with all features is higher. Whereas, in the case of Villin and Fs Peptide, MSM with the reduced set of features has higher GMRQ score, and is able to identify the slowest process.

CONCLUSIONS

Identification of CVs, that is feature selection, is one of the the critical steps in the construction of MSM. The features determines the ability of MSM to capture the slowest processes and thus the hidden conformers in the molecular dynamics simulation data. The GMRQ score of the MSM provides a metric to calculate the accuracy of the prediction made by the MSM. In this study, we have shown that, in some systems, more efficient MSMs are constructed using Laplacian score with GA using GMRQ score as fitness score. This method helps to identify which of the features, between dihedrals or contacts, is to be chosen for the construction of an efficient MSM. The most significant advantage of this method is that it helps to reduce the amount of sampling and overcomes the bottleneck of long simulation. It has the potential to circumvent over fitting caused by large dimensionality of the simulated data. This approach has been applied to simulation data involving folding of protein and protein-ligand binding. In this approach, MSM building had the goal of finding out the slowest process in the simulation. In protein folding, the movement from the unfolded state to the folded state, and in protein-ligand binding the binding of ligand to the protein are the slowest process in the simulation data identified by MSM.

The time required to find out the optimal subset of features is significant due to use of wrapper method, along with filter based feature selection. The GMRQ score of MSM constructed with all the features is higher for longer simulations. Nonetheless, this approach significantly reduces the amount of sampling required to construct an efficient MSM as shown in the case of Benzamdine-Trypsin, Villin and Fs Peptide systems, and hence helps to take advantage of adaptive sampling MD which involves a large number of short simulations. This method helps to construct MSM with higher GMRQ score and find out the optimum features that maximally affects the slowest process in the simulation data. In summary, our study shows that heuristic initialization of the GA population can automatically select the essential features that contribute to the construction of MSM with improved GMRQ score. To our knowledge, this is the first approach that uses Laplacian score along with GA to automatically select features to construct MSMs with a reduced set of features. The code used is available at GitHub: https://github.com/anuginu/MSM_latest.

ACKNOWLEDGEMENTS

The authors thank D.E. Shaw Research for MD simulation trajectories of WW-domain. The authors also thank Intel Vlab cluster for providing the clusters for running the experiment. We thank Priya Chekco for proof-reading the article.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the Department of Science and Technology (Government of India) IndoSwedish Collaborative Project (Grant Reference Number: INT/SWD/VINN/p-10/2015, AG/MP); and the Department of Biotechnology, Government of India (Grant Reference Number: BT/16476/7/631/2016, PA). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Department of Science and Technology (Government of India) Indo–Swedish Collaborative Project: INT/SWD/VINN/p10/2015, AG/MP.
Department of Biotechnology, Government of India: BT/16476/7/631/2016, PA.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Anu George conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Madhura Purnaprajna conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the paper, critical feedback on the validity of the models, results. And specific suggestions on what needs to improve in the validation, and approved the final draft.
- Prashanth Athri conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The code is available at GitHub: https://github.com/anuginu/MSM_latest.

Thrombib-piperidine: De Fabritiis, Gianni (2020): Protein Ligand Binding Assay. figshare. Collection. 10.6084/m9.figshare.c.4958759.v1

Villin: De Fabritiis, Gianni; Doerr, Stefan (2020): Protein folding of villin by Acellera. figshare. Collection. 10.6084/m9.figshare.c.4951794

Benzamidine-trypsin PDB file: PyEmma, PyEmma (2020): Benzamidine-Trypsin PDB file. figshare. Dataset. 10.6084/m9.figshare.12044853.v2

Benzamidine Trypsin trajectory Dataset: PyEmma, PyEmma (2020): Benzamidine Trypsin trajectory generated using ACEMD by pyemma. figshare. Dataset. 10.6084/m9.figshare.12034965.v2

Fs Peptide PDB file: MSMBuilder, MSMBuilder (2020): Fs Peptide PDB file. figshare. Dataset. 10.6084/m9.figshare.12044862.v2

Fs MD Trajectories: McGibbon, Robert T. (2014): Fs MD Trajectories. figshare. Dataset. 10.6084/m9.figshare.1030363.v1

WW-Domain Data: databank, RCSB protein (2020): WW domain PDB file. figshare. Dataset. 10.6084/m9.figshare.12044055.v1

WW- domain Trajectory dataset: Lindorff-Larsen, Kresten (2020): WW- domain Trajectory dataset. figshare. Collection. 10.6084/m9.figshare.c.4946391.v2.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-pchem.9#supplemental-information.

REFERENCES

Ahalawat N, Mondal J. 2018. Assessment and optimization of collective variables for protein conformational landscape: GB1 β-hairpin as a case study. *The Journal of chemical physics* **149**:094101-1–094101-10 DOI 10.1063/1.5041073.

Alder BJ, Wainwright TE. 1957. Phase transition for a hard sphere system. *The Journal of Chemical Physics* **27**:1208–1209 DOI 10.1063/1.1743957.

Beauchamp KA, Bowman GR, Lane TJ, Maibaum L, Haque IS, Pande VS. 2011. MSM-Builder2: modeling conformational dynamics on the picosecond to millisecond scale. *Journal of Chemical Theory and Computation* **7**:3412–3419 DOI 10.1021/ct200463m.

Bolhuis PG, Chandler D, Dellago C, Geissler PL. 2002. Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annual Review of Physical Chemistry* **53**:291–318 DOI 10.1146/annurev.physchem.53.082301.113146.

Bowman GR, Pande VS, Noé F. 2013. *An introduction to Markov state models and their application to long timescale molecular simulation*. vol. 797. Amsterdam: Springer Science & Business Media.

Breaban M, Luchian H. 2011. A unifying criterion for unsupervised clustering and feature selection. *Pattern Recognition* **44(4)**:854–865 DOI 10.1016/j.patcog.2010.10.006.

Buch I, Giorgino T, De Fabritiis G. 2011. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proceedings of the National Academy of Sciences of the United States of America* **108**:10184–10189 DOI 10.1073/pnas.1103547108.

Buch I, Harvey MJ, Giorgino T, Anderson DP, De Fabritiis G. 2010. High-throughput all-atom molecular dynamics simulations using distributed computing. *Journal of Chemical Information and Modeling* **50**:397–403 DOI 10.1021/ci900455r.

Bussi G, Laio A, Parrinello M. 2006. Equilibrium free energies from nonequilibrium metadynamics. *Physical Review Letters* **96(9)**:090601.

Chen Q, Feng J, Mittal S, Shukla D. 2018. Automatic feature selection in markov state models using genetic algorithm. ArXiv preprint. arXiv:1806.09723.

Childers MC, Daggett V. 2017. Insights from molecular dynamics simulations for computational protein design. *Molecular Systems Design & Engineering* **2(1)**:9–33.

Chung FR. 1997. Spectral graph theory. Providence: American Mathematical Society.

Cossio P, Laio A, Pietrucci F. 2011. Which similarity measure is better for analyzing protein structures in a molecular dynamics trajectory? *Physical Chemistry Chemical Physics* **13**:10421–10425 DOI 10.1039/c0cp02675a.

Da L-T, Pardo-Avila F, Xu L, Silva D-A, Zhang L, Gao X, Wang D, Huang X. 2016. Bridge helix bending promotes RNA polymerase II backtracking through a critical and conserved threonine residue. *Nature Communications* **7**:11244 DOI 10.1038/ncomms11244.

Dakka J, Farkas-Pall K, Balasubramanian V, Turilli M, Wright DW, Wan S, Zasada S, Coveney PV, Jha S. 2018. Rapid, concurrent and adaptive extreme scale binding free energy calculation. ArXiv preprint. arXiv:1801.01174.

De Vivo M, Masetti M, Bottegoni G, Cavalli A. 2016. Role of molecular dynamics and related methods in drug discovery. *Journal of Medicinal Chemistry* **59**:4035–4061 DOI 10.1021/acs.jmedchem.5b01684.

Devakumari D, Thangavel K. 2010. Unsupervised adaptive floating search feature selection based on contribution entropy. In: *2010 international conference on communication and computational intelligence (INCOCCI)*. Piscataway: IEEE, 623–627.

DiMasi JA, Grabowski HG, Hansen RW. 2016. Innovation in the pharmaceutical industry: new estimates of R&D costs. *Journal of Health Economics* **47**:20–33 DOI 10.1016/j.jhealeco.2016.01.012.

Doerr S, De Fabritiis G. 2014. On-the-fly learning and sampling of ligand binding by high-throughput molecular simulations. *Journal of Chemical Theory and Computation* **10**:2064–2069 DOI 10.1021/ct400919u.

Doerr S, Harvey M, Noe F, De Fabritiis G. 2016. HTMD: high-throughput molecular dynamics for molecular discovery. *Journal of Chemical Theory and Computation* **12**:1845–1852 DOI 10.1021/acs.jctc.6b00049.

Dutta D, Dutta P, Sil J. 2014. Simultaneous feature selection and clustering with mixed features by multi objective genetic algorithm. *International Journal of Hybrid Intelligent Systems* **11**:41–54.

Dy JG, Brodley CE. 2004. Feature selection for unsupervised learning. *Journal of Machine Learning Research* **5**:845–889.

Fabritiis GD, Geroult S, Coveney PV, Waksman G. 2008. Insights from the energetics of water binding at the domain-ligand interface of the Src SH2 domain. *Proteins: Structure, Function, and Bioinformatics* **72**:1290–1297 DOI 10.1002/prot.22027.

Fan Z, Dror RO, Mildorf TJ, Piana S, Shaw DE. 2015. Identifying localized changes in large systems: Change-point detection for biomolecular simulations. *Proceedings* *of the National Academy of Sciences of the United States of America* **112**:7454–7459 DOI 10.1073/pnas.1415846112.

Flocco MM, Mowbray SL. 1995. Cα-based torsion angles: a simple tool to analyze protein conformational changes. *Protein Science* **4**:2118–2122 DOI 10.1002/pro.5560041017.

Frank BS, Vardar D, Buckley DA, McKnight CJ. 2002. The role of aromatic residues in the hydrophobic core of the villin headpiece subdomain. *Protein Science* **11**:680–687.

García GC, García-Pedrajas N. 2018. Boosted feature selectors: a case study on prediction P-gp inhibitors and substrates. *Journal of Computer-Aided Molecular Design* **32**:1273–1294 DOI 10.1007/s10822-018-0171-5.

George D, Hawkins J. 2009. Towards a mathematical theory of cortical micro-circuits. *PLOS Computational Biology* **5**:e1000532 DOI 10.1371/journal.pcbi.1000532.

Grant BJ, Gorfe AA, McCammon JA. 2010. Large conformational changes in proteins: signaling and other functions. *Current Opinion in Structural Biology* **20**:142–147 DOI 10.1016/j.sbi.2009.12.004.

Gupta A, Rawlings JB. 2014. Comparison of parameter estimation methods in stochastic chemical kinetic models: examples in systems biology. *AIChE Journal* **60**:1253–1268 DOI 10.1002/aic.14409.

Harvey MJ, De Fabritiis G. 2012. High-throughput molecular dynamics: the powerful new tool for drug discovery. *Drug Discovery Today* **17**:1059–1062 DOI 10.1016/j.drudis.2012.03.017.

He X, Cai D, Niyogi P. 2006. Laplacian score for feature selection. In: *Advances in neural information processing systems*. Cambridge: MIT Press, 507–514.

Hollingsworth SA, Dror RO. 2018. Molecular dynamics simulation for all. *Neuron* **99**:1129–1143 DOI 10.1016/j.neuron.2018.08.011.

Husic BE, Pande VS. 2017. Note: MSM lag time cannot be used for variational model selection. *The Journal of Chemical Physics* **147**:176101–176102 DOI 10.1063/1.5002086.

Husic BE, Pande VS. 2018. Markov state models: From an art to a science. *Journal of the American Chemical Society* **140**:2386–2396 DOI 10.1021/jacs.7b12191.

Kohlhoff KJ, Shukla D, Lawrenz M, Bowman GR, Konerding DE, Belov D, Altman RB, Pande VS. 2014. Cloud-based simulations on Google Exacycle reveal ligand modulation of GPCR activation pathways. *Nature Chemistry* **6**:15–21 DOI 10.1038/nchem.1821.

Laio A, Parrinello M. 2002. Escaping free-energy minima. *Proceedings of the National Academy of Sciences of the United States of America* **99**:12562–12566 DOI 10.1073/pnas.202427399.

Lecina D, Gilabert JF, Guallar V. 2017. Adaptive simulations, towards interactive protein-ligand modeling. *Scientific Reports* **7**:8466 DOI 10.1038/s41598-017-08445-5.

Leelananda SP, Lindert S. 2016. Computational methods in drug discovery. *Beilstein Journal of Organic Chemistry* **12**:2694–2718 DOI 10.3762/bjoc.12.267.

Li G, Hu X, Shen X, Chen X, Li Z. 2008. A novel unsupervised feature selection method for bioinformatics data sets through feature clustering. In: *2008 IEEE international conference on granular computing*. IEEE, 41–47.

Li Y, Lu B-L, Wu Z-F. 2006. A hybrid method of unsupervised feature selection based on ranking. In: *18th international conference on pattern recognition (ICPR'06)*. IEEE, 687–690.

Lindorff-Larsen K, Piana S, Dror RO, Shaw DE. 2011. How fast-folding proteins fold. *Science* **334**:517–520 DOI 10.1126/science.1208351.

Lovera S, Sutto L, Boubeva R, Scapozza L, Dolker N, Gervasio FL. 2012. The different flexibility of c-Src and c-Abl kinases regulates the accessibility of a druggable inactive conformation. *Journal of the American Chemical Society* **134**:2496–2499 DOI 10.1021/ja210751t.

Malmstrom RD, Lee CT, Van Wart AT, Amaro RE. 2014. Application of molecular-dynamics based markov state models to functional proteins. *Journal of Chemical Theory and Computation* **10**:2648–2657 DOI 10.1021/ct5002363.

Martinez JD, Parker MT, Fultz KE, Ignatenko NA, Gerner EW, Donald E, Abraham J, Wiley IJ. 2003. *Burgers medicinal chemistry and drug discovery*. Hoboken: John Wiley & Sons.

McGibbon RT, Beauchamp KA, Harrigan MP, Klein C, Swails JM, Hernández CX, Schwantes CR, Wang L-P, Lane TJ, Pande VS. 2015. MDTraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophysical Journal* **109**:1528–1532 DOI 10.1016/j.bpj.2015.08.015.

McGibbon RT, Pande VS. 2015. Variational cross-validation of slow dynamical modes in molecular kinetics. *The Journal of Chemical Physics* **142**:03B621_1.

Meyer AG, Sawyer SL, Ellington AD, Wilke CO. 2014. Analyzing machupo virus-receptor binding by molecular dynamics simulations. *PeerJ* **2**:e266 DOI 10.7717/peerj.266.

Mitra P, Murthy C, Pal SK. 2002. Unsupervised feature selection using feature similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**:301–312 DOI 10.1109/34.990133.

Mittal S, Shukla D. 2017. Predicting optimal deer label positions to study protein conformational heterogeneity. *The Journal of Physical Chemistry B* **121**:9761–9770 DOI 10.1021/acs.jpcb.7b04785.

Mohan S, Sheena A, Poulose N, Anilkumar G. 2010. Molecular dynamics simulation studies of GLUT4: substrate-free and substrate-induced dynamics and ATP-mediated glucose transport inhibition. *PLOS ONE* **5**:e14217 DOI 10.1371/journal.pone.0014217.

Molgedey L, Schuster HG. 1994. Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters* **72**:3634–3638 DOI 10.1103/PhysRevLett.72.3634.

Namboori PK, Vasavi C, Gopal KV, Gopakumar D, Ramachandran K, Narayanan BS. 2010. Thermal analysis of nanofluids using modeling and molecular dynamics simulation. In: *Conference Proceedings*. AIP, 407–412.

Noé F, Clementi C. 2017. Collective variables for the study of long-time kinetics from molecular trajectories: theory and methods. *Current Opinion in Structural Biology* **43**:141–147 DOI 10.1016/j.sbi.2017.02.006.

Noé F, Fischer S. 2008. Transition networks for modeling the kinetics of conformational change in macromolecules. *Current Opinion in Structural Biology* **18**:154–162 DOI 10.1016/j.sbi.2008.01.008.

Noé F, Nuske F. 2013. A variational approach to modeling slow processes in stochastic dynamical systems. *Multiscale Modeling & Simulation* **11**:635–655 DOI 10.1137/110858616.

Noé F, Schütte C, Vanden-Eijnden E, Reich L, Weikl TR. 2009. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proceedings of the National Academy of Sciences of the United States of America* **106**:19011–19016 DOI 10.1073/pnas.0905466106.

Pande VS. 2014. Understanding protein folding using Markov state models. *Advances in Experimental Medicine and Biology* **797**:101–106 DOI 10.1007/978-94-007-7606-7_8.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay é. 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* **12**:2825–2830.

Pérez F, Granger BE. 2007. IPython: a system for interactive scientific computing. *Computing in Science & Engineering* **9**:21–29.

Pérez-Hernández G, Paul F, Giorgino T, De Fabritiis G, Noé F. 2013. Identification of slow molecular order parameters for Markov model construction. *The Journal of Chemical Physics* **139**:015102–015113.

Plattner N, Noé F. 2015. Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and Markov models. *Nature Communications* **6**:7653 DOI 10.1038/ncomms8653.

Prinz J-H, Wu H, Sarich M, Keller B, Senne M, Held M, Chodera JD, Schütte C, Noé F. 2011. Markov models of molecular kinetics: Generation and validation. *The Journal of Chemical Physics* **134**:174105–174129 DOI 10.1063/1.3565032.

Salomon-Ferrer R, Götz AW, Poole D, Le Grand S, Walker RC. 2013. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald. *Journal of Chemical Theory and Computation* **9**:3878–3888 DOI 10.1021/ct400314y.

Sánchez JS, García V. 2018. Addressing the links between dimensionality and data characteristics in gene-expression microarrays. In: *Proceedings of the international conference on learning and optimization algorithms: theory and applications*. ACM, 1.

Scherer MK, Trendelkamp-Schroer B, Paul F, Pérez-Hernández G, Hoffmann M, Plattner N, Wehmeyer C, Prinz J-H, Noé F. 2015. PyEMMA 2: a software package for estimation, validation, and analysis of markov models. *Journal of Chemical Theory and Computation* **11**:5525–5542 DOI 10.1021/acs.jctc.5b00743.

Schwantes CR, Pande VS. 2013. Improvements in Markov state model construction reveal many non-native interactions in the folding of NTL9. *Journal of Chemical Theory and Computation* **9**:2000–2009 DOI 10.1021/ct300878a.

Schwantes CR, Shukla D, Pande VS. 2016. Markov state models and tICA reveal a non-native folding nucleus in simulations of NuG2. *Biophysical Journal* **110**:1716–1719 DOI 10.1016/j.bpj.2016.03.026.

Sgourakis NG, Merced-Serrano M, Boutsidis C, Drineas P, Du Z, Wang C, Garcia AE. 2011. Atomic-level characterization of the ensemble of the aβ (1–42) monomer in water using unbiased molecular dynamics simulations and spectral algorithms. *Journal of Molecular Biology* **405**:570–583 DOI 10.1016/j.jmb.2010.10.015.

Shamsi Z, Cheng KJ, Shukla D. 2018. Reinforcement learning based adaptive sampling: REAPing rewards by exploring protein conformational landscapes. *The Journal of Physical Chemistry B* **122**:8386–8395 DOI 10.1021/acs.jpcb.8b06521.

Shaw DE, Grossman J, Bank JA, Batson B, Butts JA, Chao JC, Deneroff MM, Dror RO, Even A, Fenton CH, Forte A, Gagliardo J, Gill G, Greskamp B, Ho CR, Ierardi DJ, Iserovich L, Kuskin JS, Larson RH, Layman T, Lee L-S, Lerer AK, Li C, Kilebrew D, Mackenzie KM, Mok SY-H, Moraes MA, Mueller R, Nociolo LJ, Peticolas JL. 2014. Anton 2: raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. In: *Proceedings of the international conference for high performance computing, networking, storage and analysis*. IEEE Press, 41–53.

Shukla D, Hernández CX, Weber JK, Pande VS. 2015. Markov state models provide insights into dynamic modulation of protein function. *Accounts of Chemical Research* **48**:414–422 DOI 10.1021/ar5002999.

Singhal N, Snow CD, Pande VS. 2004. Using path sampling to build better Markovian state models: predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. *The Journal of Chemical Physics* **121**:415–425 DOI 10.1063/1.1738647.

Skewes AD, Welch RD. 2013. A Markovian analysis of bacterial genome sequence constraints. *PeerJ* **1**:e127 DOI 10.7717/peerj.127.

Śledź P, Caflisch A. 2018. Protein structure-based drug design: from docking to molecular dynamics. *Current Opinion in Structural Biology* **48**:93–102 DOI 10.1016/j.sbi.2017.10.010.

Solorio-Fernández S, Carrasco-Ochoa JA, Martínez-Trinidad JF. 2016. A new hybrid filter–wrapper feature selection method for clustering based on ranking. *Neurocomputing* **214**:866–880 DOI 10.1016/j.neucom.2016.07.026.

Solorio-Fernández S, Carrasco-Ochoa JA, Martínez-Trinidad JF. 2020. A review of unsupervised feature selection methods. *Artificial Intelligence Review* **53**:907–948 DOI 10.1007/s10462-019-09682-y.

Stone JE, Phillips JC, Freddolino PL, Hardy DJ, Trabuco LG, Schulten K. 2007. Accelerating molecular modeling applications with graphics processors. *Journal of Computational Chemistry* **28**:2618–2640 DOI 10.1002/jcc.20829.

Sugita Y, Okamoto Y. 1999. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters* **314**:141–151 DOI 10.1016/S0009-2614(99)01123-9.

Sultan MM, Kiss G, Shukla D, Pande VS. 2014. Automatic selection of order parameters in the analysis of large scale molecular dynamics simulations. *Journal of Chemical Theory and Computation* **10**:5217–5223 DOI 10.1021/ct500353m.

Tabakhi S, Moradi P. 2015. Relevance–redundancy feature selection based on ant colony optimization. *Pattern Recognition* **48**:2798–2811 DOI 10.1016/j.patcog.2015.03.020.

Tang X, Bevan MA, Grover MA. 2017. The construction and application of Markov state models for colloidal self-assembly process control. *Molecular Systems Design & Engineering* **2**:78–88 DOI 10.1039/C6ME00092D.

Tiwary P, Berne B. 2016. Spectral gap optimization of order parameters for sampling complex molecular systems. *Proceedings of the National Academy of Sciences of the United States of America* **113**:2839–2844 DOI 10.1073/pnas.1600917113.

Tiwary P, van de Walle A. 2016. A review of enhanced sampling approaches for accelerated molecular dynamics. In: Weinberger C, Tucker G, eds. *Multiscale materials modeling for nanomechanics. Springer series in materials science*, vol. 245. Cham: Springer.

Vanatta DK, Shukla D, Lawrenz M, Pande VS. 2015. A network of molecular switches controls the activation of the two-component response regulator ntrc. *Nature Communications* **6**:7283.

Vargiu AV, Ruggerone P, Magistrato A, Carloni P. 2008. Dissociation of minor groove binders from dna: insights from metadynamics simulations. *Nucleic Acids Research* **36(18)**:5910–5921.

Wang W, Cao S, Zhu L, Huang X. 2018. Constructing markov state models to elucidate the functional conformational changes of complex biomolecules. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **8(1)**:e1343.

Zhang BW, Dai W, Gallicchio E, He P, Xia J, Tan Z, Levy RM. 2016. Simulating replica exchange: Markov state models, proposal schemes, and the infinite swapping limit. *The Journal of Physical Chemistry B* **120**:8289–8301 DOI 10.1021/acs.jpcb.6b02015.

Zheng S, Pfaendtner J. 2015. Enhanced sampling of chemical and biochemical reactions with metadynamics. *Molecular Simulation* **41**:55–72 DOI 10.1080/08927022.2014.923574.