EPiC Series in Computing

Volume 60, 2019, Pages 181–189

Proceedings of 11th International Conference on Bioinformatics and Computational Biology

Connecting Molecular Energy Landscape Analysis with Markov Model-based Analysis of Equilibrium Structural Dynamics

Kazi Lutful Kabir[1], Nasrin Akhter[1], and Amarda Shehu[1,2,3,*]

[1] Department of Computer Science, George Mason University, Fairfax, VA 22030
[2] Department of Bioengineering, George Mason University, Fairfax, VA 22030
[3] School of Systems Biology, George Mason University, Fairfax, VA 22030

Abstract

Molecular dynamics simulation software now provides us with a view of the structure space accessed by a molecule. Increasingly, Markov state models are proposed to integrate various simulations of a molecule and extract its equilibrium structural dynamics. The approach relies on organizing the structures accessed in simulation into states as an attempt to identify thermodynamically-stable and semi-stable (macro)states among which transitions can then be quantified. Typically, off-the-shelf clustering algorithms are used for this purpose. In this paper, we investigate two additional complementary approaches to state identification that rely on graph embeddings of the structures. In particular, we show that doing so allows revealing basins in the energy landscape associated with the accessed structure space. Moreover, we demonstrate that basins, directly tied to stable and semi-stable states, yield to a better model of dynamics on a proof-of-concept application.

1 Introduction

Decades of experimental, theoretical, and computational research have shown that biological molecules are intrinsically dynamic. In particular, peptides and proteins harness their structural equilibrium dynamics into productive events in which they interact with molecular partners in the cell [4]. Due to its central role in governing the biological activities of a molecule, elucidating the equilibrium structural dynamics is of primary importance in molecular biology [19].

While great progress is being made, wet-laboratory techniques are able to reveal few snapshots of a molecule rearranging between different structures. Molecular Dynamics (MD) simulations are increasingly providing us with a broader view of the structure space navigated by a molecule one trajectory of structures at a time [12]. Despite algorithmic and hardware advances, typically, many MD simulations are needed to address the issue of limited sampling. Launching many MD simulations yields various trajectories that need to be integrated and analyzed in order to extract the dynamics captured in simulation.

*Correspondence: amarda@gmu.edu

O. Eulenstein, H. Al-Mubaid and Q. Ding (eds.), BiCOB 2019 (EPiC Series in Computing, vol. 60), pp. 181–189

Increasingly, Markov state models (MSM) are being proposed as being able to integrate various MD trajectories and extract the elusive equilibrium dynamics [6, 11]. MSMBuilder [7] and Emma [17, 18] are being increasingly adopted by computational biophysicists. MSMs seek to first organize structures accessed in simulation (which represent microstates) into structural (macro)states. The goal is to reveal the thermodynamically-stable and semi-stable states populated by a system of interest and then quantify transitions among identified states (as evidence in the order in which structures are visited in simulation) [6]. Current practice is to rely on off-the-shelf clustering algorithms to group structures into states.

In this paper we propose that the quality of the extracted structural dynamics (and so the accuracy of the dynamics model) can be improved by leveraging the connection between stable and semi-stable states and basins in the energy landscape associated with the structure space of a molecule. Indeed, current MSM software ignore the energetics that one obtains along with structures from MD simulations.

Inspired by success in improving decoy selection in protein structure prediction by leveraging the concept of basins [1–3], we propose to define states as basins and replace structure clustering with basin identification. We do so in a proof-of-concept setting on MD simulation data provided by collaborators. We extract models of dynamics via PyEmma [17] in three different settings, one where we employ off-the-shelf clustering algorithms in PyEmma, one where we employ clustering algorithms that leverage graph embeddings of structures, and another where we leverage the energy landscape and replace clustering with basin identification. We conduct a rigorous evaluation of the extracted models and show that the basins result in a more accurate model of structural dynamics.

2 Method

We first conceptually summarize the steps that are followed to build an MSM from various MD trajectories, devoting the rest of this section to describing the proposed basin-based state identification and the evaluation of an MSM of structural dynamics.

2.1 Summary of MSM Construction

The input to MSM building methods is a list of possibly more than one MD trajectories (we note that we focus here on simulations conducted at equilibrium, under physiological conditions). While different trajectory formats are allowed by different software, the minimum information that each trajectory needs to have is the series of structures that were accessed in order by an MD simulation and the time step expired between two consecutive structures. Structures are specified in terms of the Cartesian coordinates of the atoms constituting the molecule (or molecular system) of interest. The time step is typically a parameter of choice in an MD simulation and can vary from 1 femtosecond (fs) in detailed, all-atom simulations to several fs in coarse-grained ones. Note that the time steps can be different among the trajectories over which an MSM integrates to extract the dynamics.

Moreover, not all structures in an MD trajectory need to be used. Most MSM software allow the user to indicate a lag time that can be a multiple of the actual time scale between two consecutive structures in a trajectory; this has the effect of skipping over structures and losing some temporal and spatial resolution. The assumption is made that structures within the same state interconvert faster than the chosen lag time, and this condition can be tested by varying lag time and observing properties of the resulting MSM [6]. Once a lag time is chosen, MSM construction follows a systematic process and consists of the following four steps.

(i) A *featurizer* extracts features of interest from the given structures; many options are supported in existing software, such as Cartesian- or angle-based features computed over all or selected atoms (e.g., only heavy backbone atoms).

(ii) The dimensionality of the resulting data is reduced further. Typically, two representative linear transformations are used, Principal Component Analysis (PCA) or time-lagged Independent Component Analysis (TICA) [10]. The basic difference between them is that PCA extracts coordinates of maximal variance, whereas TICA captures coordinates of maximal auto-correlation for a given lag time. Typically, studies show TICA outperforms PCA on MD trajectory data [16].

(iii) The reduced/projected (structure) trajectories are then subjected to a clustering algorithm to group them into states. This process is also referred to as state space discretization, and alternatives are provided for the clustering algorithm. One can utilize common clustering approaches or apply their own strategy. It is here that we evaluate two different approaches and make the case that state identification needs to be tied to the energy landscape.

(iv) Once states are identified, transition probabilities are then obtained. Micro(state) trajectory(ies) can be utilized to generate a maximum likelihood estimate (MLE) of a probability distribution of transitions to (macro)states in the form of a suitable transition matrix [6]. This information can be utilized to construct the corresponding MSM, which can then be subjected to rigorous analysis, such as model selection, estimation (e.g., Bayesian) and validation. The latter tests whether the model is capable of making reliable predictions regarding the kinetics of the system under observation.

2.2 State Space Discretization: State Identification

Here we build MSMs via PyEmma due to its Python interface [17]. PyEmma is representative of other MSM building software in that it offers common clustering algorithms that ignore energies provided alongside the structures from MD simulations. Three clustering algorithms are provided: k-means, regular space clustering, and uniform time clustering. In its application of k-means, PyEmma assigns structures to the k clusters based on a Voronoi discretization of the structures space around the cluster representatives. The regular space clustering algorithm is a variant of Hartigan's leader algorithm, where each data point (note that structures are projected onto some space after the featurizer and dimensionality reduction and are now data points) (considered in some arbitrary order) is either assigned to an existing cluster (based on its proximity and a distance threshold) or opens up a new cluster as its representative. In uniform time clustering, the selection of structures/data points is conducted uniformly over time, and Voronoi discretization is used for assignment of data points.

We propose and evaluate two more state space discretization strategies in this paper. The first does not utilize energies that are available alongside structures from MD simulations but leverages concepts from community detection in social networks to organize structures into states. The second additionally leverages energies and considers the structures as points in the energy landscape (that lifts the structure space with an additional dimension, energy) where then basins can be identified and serve as states. We describe each of these next.

Network Communities as Clusters

Structures accessed in MD simulations are embedded in a nearest-neighbor graph (nngraph) that encodes the proximity among structures in the structure space. Consider an Ω set of structures to be embedded in an nngraph $G = (V, E)$ where the vertex set V is populated with structures, and the edge set E is populated by inferring a local neighborhood over each

structure/vertex. A structure is connected via edges to its nearest neighbors. The latter are identified either by setting a maximum number of such neighbors or by setting a distance threshold.

The distance between two structures is measured via root-mean-squared-deviation (*RMSD*). Using RMSD to compute the distance between two conformations, if a distance threshold ϵ is used, each vertex $u \in V$ is connected to vertices $v \in V$ if $d(u,v) \leq \epsilon$. Note that the distance here is not optimally aligning two structures under consideration to remove rigid-body motions. This process would be very time-consuming were it to be performed on every pair of structures under comparison. Instead, as current practice in other settings, the structures are first all aligned to a reference structure (arbitrarily selected to be the first in some MD trajectory) and then only RMSD is computed among every pair of structures under comparison. We note that the identification of nearest neighbors of a vertex can be done efficiently via proximity query data structures.

The resulting graph can then be subjected to community detection algorithms that are typically proposed for applications on graphs that encode relations among entities or individuals (e.g., social networks). Such algorithms effective cluster structures together based on their organization in the graph, identifying cohesive groupings among structures/vertices. There are now many such algorithms. According to a previous study that evaluates community detection algorithms on protein structures obtained in a protein structure prediction setting [8], the top two algorithms are Louvain and Greedy Modularity Maximization. In this paper, we focus our evaluation on only these two algorithms, utilizing either one to group structures into states for the purpose of building (and then evaluating) different MSM models.

Energy Landscape Basins as Macrostates

An alternative approach proposed here considers the energetics of structures obtained in simulation. The approach envisions an underlying energy landscape sampled (one structure-energy pair at a time) in simulation. The organization of the landscape offers a quantitative understanding of dynamics; afterall, the landscape contains information on how structures with similar energies interconvert into one another [14]. Indeed, the definition of a thermodynamically-stable (or semi-stable) state does not rely on structural similarity alone. Stable and (semi-stable) states correspond to basins/wells in the landscape [15].

Therefore, statistical spatial analytics of the MD-sampled energy landscape can be employed to extract basins and relate them to the states over which an MSM infers the state-to-state equilibrium dynamics. One can do so by first (again) embedding the structure ensemble Ω in an nngraph, as described above in the context of community detection. A basin is tied to a unique local minimum (its focal minimum, which is the deepest point in the basin). Let $u \in V$, and let $v \in N(u)$, where $N(u)$ denotes the 1-neighborhood of u in the graph. u is a local minimum if $\forall v \in V\ f(u) \leq f(v)$. The remaining vertices in the graph are assigned to basins as follows. Each vertex u is associated a negative gradient estimated by selecting the edge (u, v) that maximizes the ratio $[f(u) - f(v)]/d(u, v)$. From each vertex u that is not a local minimum, the negative gradient is iteratively followed (i.e., the edge that maximizes the above ratio is selected and followed) until a local minimum is reached. Vertices that reach via this process the same local minimum are assigned to the (same) basin associated with that minimum. The resulting basins are now treated as states that can be employed to build an MSM.

2.3 Analysis/Evaluation of MSM

The above approaches can result in different MSMs, since they provide different states extracted from analysis of the structure space or energy landscape probed by various MD simulations. These MSMs can be evaluated to determine the impact and effectiveness of the above state identification approaches. One way to do so is to check whether the duration of the lag time is adequate to ensure that the (micro)state discretization maintains the Markov property [11]. This property states that, if the state decomposition is accurate, structures within a state interconvert on timescales faster than the lag time and transition to other states on slower timescales.

The standard practice to test whether an obtained model satisfies the Markovian property is to visually interpret the generated implied timescale plot of the model relaxation timescale vs model lag time. The desired property is to have an exponential decay in the plot to system equilibrium. With relaxation timescales being physical properties of the system, the ideal case is for the implied timescales to be independent of the lag time. According to the variational principle of structural dynamics [13], it desired that the model have longer timescale. For an ideal model with good discretization, implied timescales plots would exhibit convergence within fewer steps. Another way to test for the Markovian property is via the Chapman-Kolmogorov test, which compares the transition probability of different (macro)states for increasing lag time steps of the MD trajectories. Again, the goal is to establish whether the lag time is sufficiently long to make the chosen state decomposition Markovian.

3 Results

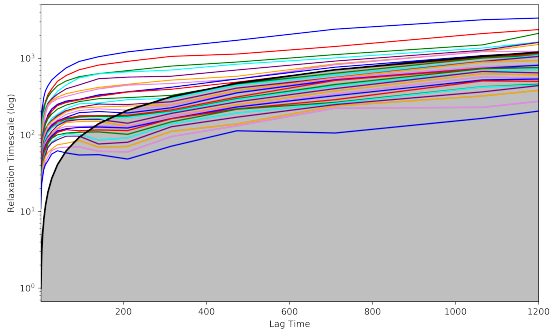
Dataset We carry out evaluation on a dataset of $30,000$ structures shared with us by our collaborators. The dataset consists of three different MD trajectories obtained at equilibrium for the Met-enkephaline (Met-Enk) peptide. The peptide is a naturally-occurring opioid that mediates pain and opiate dependency by interacting with opioid receptors [9]. Interest on Met-Enk dynamics is due to a hypothesis that the peptide is highly flexible and possibly involved in many more interactions than currently known. The MD trajectories are obtained via MD simulations in AMBER [5], using all-atom detail and explicit solvent, with a time step of 1fs.

Experimental Setup We carry out a convergence analysis to compare the quality of the models obtained with the different state space discretization approaches described in Section 2. We then visualize the best model of dynamics and analyze it for what it reveals about the equilibrium structural dynamics of the Met-Enk peptide.

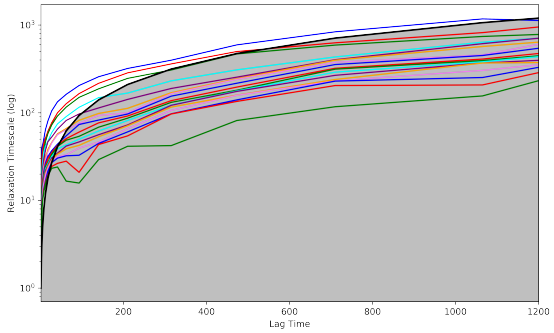
3.1 Convergence Analysis

In the spirit of standard best practices, we visualize the implied timescales plot (model relaxation timescale versus model lag time) to compare the quality of the MSMs obtained for different state identification approaches. Under each approach, we have considered two options, one where the analysis is carried out over all atoms, and another where it is carried out over only backbone atoms. In addition, for each of the three clustering algorithms in PyEmma we have considered both PCA and TICA as dimensionality reduction. A comparison of the plots shows superior performance when employing only backbone atoms over all atoms and when employing TICA over PCA. In Fig. 1 we show the implied timescales plots when state identification is carried out with k-means, with the Louvain community detection algorithm, or

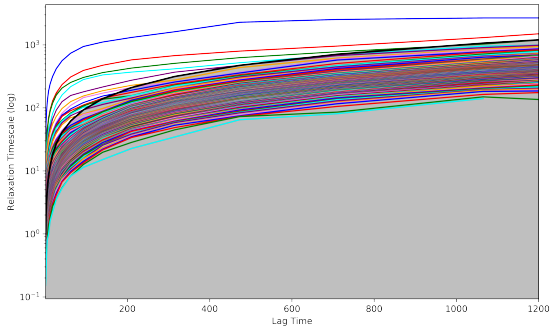
with the basin identification algorithm described in Section 2. We do not show results obtained when using the other two clustering algorithms available in PyEmma, as they are outperformed by k-means, and we only show results obtained with the Louvain community detection algorithm, as it outperforms the Greedy Modularity Maximization algorithm.



(a) k-means-based



(b) Louvain-based



(c) Basin-based

Figure 1: Implied Timescale Plot obtained when carrying out state identification with (a) Pyemma's k-means, (b) Louvain community detection algorithm, or (c) basin identification algorithm. Cutoff region (above which any curve should be) is shown in gray.

Fig. 1(a) shows that the dynamics model obtained when states are identified with k-means clustering fails to reach convergence within 1000 time steps (models obtained with states identified via regular space clustering or uniform time clustering far worse). The model obtained when states are identified with Louvain's community detection algorithm performs better than when states are identified via k-means, but Fig. 1(b) shows that even this model is unable to

exhibit convergence even after 1200 steps. In contrast, Fig. 1(c) shows that the model obtained when states are identified via the basin-based strategy reaches convergence quite early (after about 700 steps).

3.2 Visualization of Best Model

The above results suggest that the best MSM of the dynamics is obtained when states are identified as basins in the energy landscape probed in MD simulation. We visualize that model in Fig. 2. Rather than show all the 173 states (173 basins are identified), in the interest of clarity, we limit the visualization to the five states with the highest self-transition probabilities; altogether these states contain 16% of the structures. These states are shown as disks in Fig. 2 of radii proportional to the number of structures in them. Transitions among them and other not shown states are drawn as arcs, with the transition probabilities annotated, as well. The visualization of the model makes it clear that the peptide visits several long-lived states; each of the shown states roughly have self-transition probabilities ≥ 0.7. State 1 has a very high self-transition probability of over 0.9. The other states have slightly lower self-transition probabilities and comparatively-low probabilities of transition to other states.

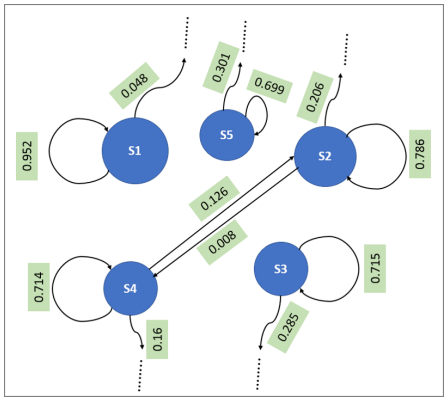


Figure 2: Best model of dynamics is visualized here, limiting the visualization to the top five states.

Each of the states is shown in Fig. 3. All the structures in a state are superimposed over one another. Each amino acid is color-coded according to its position (red denotes the N terminus and silver the C terminus). Fig. 3 shows that the states are structurally distinct and so capture different regions/basins of the energy landscape probed for this peptide in simulation. Finally, the states are compared with known wet-laboratory structures of the peptide in order

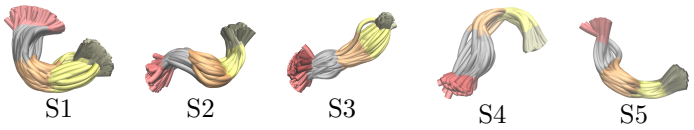


Figure 3: Structures in each of the top five states of the best model are shown superimposed over one another. Red denotes the N terminus, and silver the C terminus.

to establish which states capture already known ones and which may constitute new, unknown states of this peptide. Known structures are under Protein Databank entries 1PLW, 1PLX, and

2LWC. Table 1 compares each identified state to each known structure. We should note that each of the known structures is indeed a small NMR ensemble of 20-80 models. The first model is selected for comparison. All structures in a state are compared via RMSD (after optimal superposition removes rigid-body differences) to the first structure in each NMR ensemble. Table 1 reports the minimum and average RMSD and highlights in bold RMSDs under 1Å. Such RMSDs can be used to establish a correspondence between computationally-identified states and experimentally-resolved structures. For instance, S2 can be considered to cover 1PLW, S1 covers both 1PLX and 2LWC, S3 covers 1PLX, and S5 covers 1PLW. With a more stringent cutoff of under 0.7Å, only S1 and S2 cover the three experimentally-known structures, whereas S3-S5 constitute new states. This type of analysis suggests that the MD simulation has probed novel states of Met-Enk; at least one of them, S4, as shown in Fig. 2 transitions to a state (S2) captured in the wet laboratory.

Table 1: Structures in each state are compared to the first structure in each of the three NMR ensembles (PDB identifiers shown) deposited in the PDB. Average and minimum RMSD (Å) are reported. Entries in bold highlight the lowest RMSD establishing a correspondence between a state and a known structure.

	1PLW		**1PLX**		**2LWC**	
	avg	min	avg	min	avg	min
S1	2.813	1.181	1.920	**0.712**	2.023	**0.543**
S2	1.018	**0.516**	1.905	1.411	1.895	1.539
S3	1.540	1.167	1.904	**0.845**	1.376	**0.364**
S4	2.229	1.244	1.748	1.214	1.817	1.199
S5	1.511	**0.741**	1.575	1.321	1.480	1.280

4 Conclusion

This paper makes a contribution towards the study of functional plasticity of biological molecules in simulation. The paper does so in the context of extracting the equilibrium structural dynamics of a molecule of interest via MSM-based integration of MD trajectories. In particular, identification of basins in the energy landscape probed by various MD trajectories is proposed to replace the current clustering-based organization of structures into states. A proof-of-concept evaluation on several MD trajectories of a peptide of interest to human health reveals that basins capture structural states better and yield more accurate models of structural dynamics. While our immediate interest is not reading in great detail the elucidated dynamics, Section 3 discusses valuable insights that can be obtained for this particular peptide. The presented study represents a promising first step, and further work will investigate other molecular systems, as well as alternative approaches at extracting organizations of energy landscapes that can be directly operationalized into MSM-based models of dynamics.

Acknowledgements

This work is supported in part by NSF Grant No. 1440581 and a Jeffress Memorial Trust Award. Computations were run on ARGO, a research computing cluster provided by the Office of Research Computing at George Mason University. We are thankful to Mahmoud Namazi for the MD simulation data.

References

[1] N. Akhter, L. Hassan, Z. Rajabi, D. Barbara, and A. Shehu. Learning organizations of protein energy landscapes: An application on decoy selection in template-free protein structure prediction. In A. Kister, editor, *Protein Supersecondary Structure*, Methods in Molecular Biology. Springer Verlag, 2018.

[2] N. Akhter, W. Qiao, and A. Shehu. An energy landscape treatment of decoy selection in template-free protein structure prediction. *Computation*, 6(2):39, 2018.

[3] N. Akhter and A. Shehu. From extraction of local structures of protein energy landscapes to improved decoy selection in template-free protein structure prediction. *Molecules*, 23(1):216, 2018.

[4] D. D. Boehr, R. Nussinov, and P. E. Wright. The role of dynamic conformational ensembles in biomolecular recognition. *Nature Chem Biol*, 5(11):789–96, 2009.

[5] D. A. Case, T. A. Darden, T. E. III Cheatham, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, K. M. Merz, D. A. Pearlman, M. Crowley, R. C. Walker, W. Zhang, B. Wang, S. Hayik, A. Roitberg, G. Seabra, K. F. Wong, F. Paesani, X. Wu, S. Brozell, V. Tsui, H. Gohlke, L. Yang, C. Tan, J. Mongan, et al. Amber 14. http://ambermd.org/, 2014.

[6] J. D. Chodera and F. Noé. Markov state models of biomolecular conformational dynamics. *Curr Opin Struct Biol*, 25:135–144, 2014.

[7] M. P. Harrigan, M. M. Sultan, C. X. Hernánndez, et al. MSMBuilder: Statistical models for biomolecular dynamics. *Biophys J*, 112(1):10–15, 2017.

[8] Liban Hassan, Zahra Rajabi, Nasrin Akhter, and Amarda Shehu. Community detection for decoy selection in template-free protein structure prediction. In *ACM Conf on Bioinf and Comp Biol Workshops (BCBW)*, pages 621–627, 2018.

[9] A. Koneru, S. Satyanarayana, and S. Rizwan. Endogeneous opioids: Their physiological role and receptors. *Global J Pharmacol*, 3(3):149–153, 2009.

[10] D. G. Luenberger. *Linear and Nonlinear Programming*. Addison-Wesley, 2nd edition, 1984.

[11] Robert D Malmstrom, Christopher T Lee, Adam T Van Wart, and Rommie E Amaro. Application of molecular-dynamics based markov state models to functional proteins. *J Chem Theory Comput*, 10(7):2648–2657, 2014.

[12] T. Maximova, R. Moffatt, B. Ma, R. Nussinov, and A. Shehu. Principles and overview of sampling methods for modeling macromolecular structure and dynamics. *PLoS Comp. Biol.*, 12(4):e1004619, 2016.

[13] Frank Noé and Feliks Nuske. A variational approach to modeling slow processes in stochastic dynamical systems. *Multiscale Modeling & Simulation*, 11(2):635–655, 2013.

[14] R. Nussinov and P. G. Wolynes. A second molecular biology revolution? the energy landscapes of biomolecular function. *Phys Chem Chem Phys*, 16(14):6321–6322, 2014.

[15] K. Okazaki, N. Koga, S. Takada, J. N. Onuchic, and P. G. Wolynes. Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: Structure-based molecular dynamics simulations. *Proc Natl Acad Sci USA*, 103(32):11844–11849, 2006.

[16] G. Perez-Hernandez, F. Paul, T. Giorgino, G. de Fabritiis, and F. Noé. Identification of slow molecular order parameters for markov model construction. *J Chem Phys*, 139(1):015102, 2013.

[17] M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, Christoph. Wehmeyer, J.-H. Prinz, and F. Noé. PyEMMA 2: A software package for estimation, validation, and analysis of markov models. *J Chem Theory Comput*, 11:5525–5542, 2015.

[18] M. Senne, B. Trendelkamp-Schroer, A. S. Mey, C. Schutte, and F Noé. EMMA: A software package for markov model building and analysis. *J Chem Theory Comput*, 8(7):2223–2238, 2012.

[19] A. Shehu and R. Nussinov. Computational methods for exploration and analysis of macromolecular structure and dynamics. *PLoS Comput Biol*, 11(10):e1004585, 2015. editorial.