Introduction to a Data-driven Analysis Tool of Molecular Dynamics Self-Assembled Lipid Bilayer Trajectories

Stelios Karozis
Institute of Nuclear & Radiological Sciences and Technology, Energy & Safety,
NCSR "Demokritos"
Greece

Michael Kainourgiakis
Institute of Nuclear & Radiological Sciences and Technology, Energy & Safety,
NCSR "Demokritos"
Greece

ABSTRACT

The in-silico studies reported so far for the representation of the structure and the evaluation of the transport properties of lipid bilayers are in general based on assumptions and approaches that simplify the real system and problem. Nevertheless, the structure and organization of the lipid bilayers strongly affect transport coefficients. This is a quite important observation, showing that simulations can be meaningful only when addressing realistic structures, mimicking the actual lipid phase system as elaborately as possible.

In the current study, a computational tool is presented that uses Molecular Dynamics simulations (MD) results of spontaneous self-assembly lipid bilayer structures with different oriented and shaped lipid bilayer, in order to analyze the resulted trajectories, creating a Machine Learning (ML) ready dataset that can be used in a series of ML algorithms, depending the case. The development of the tool is in the alpha stage, where tests are performed, with a planned public release in free and open source license.

KEYWORDS

lipid bilayer, Molecular Dynamics, Machine learning

1 INTRODUCTION

As molecular simulations (MS) continue to evolve into powerful computational tool for studying complex biomolecular systems and the exponential growth of computational power, the systems under study are becoming more complex. As such, a large amount of configurations are produced with more ease that permit to diminish the uncertainty of the calculated thermodynamics properties. The main tools derive from statistical mechanics, hence the larger the sample becomes, the more accurate the calculation.

On the other hand, the large amount of MS results creates a data processing problem in terms of software and hardware capabilities. The hardware problem can be surpassed with modern solutions, such as distributed data processing systems, or by new software implementations that are more efficient in limited hardware infrastructures. Most MS simulation packages incorporate their own post processing tools or suggest the use of open source compatible softwares that are sufficient enough for most cases. MDTraj [1] is used in a range of cases or as basis for other processing software like TTClust [2], that partition thousands of frames into a limited number of most dissimilar conformations. Other tools, like TRAVIS ("Trajectory Analyzer and Visualizer") [3] and pyPcazip [4] are autonomous and were developed for a specific case, thus lacking generic applicability.

The aforementioned packages, alongside the incorporated tools of MD and MC simulation softwares, are well established and tested but they don't solve the problem of processing the big data production of MS simulations. Machine Learning (ML) algorithms are data analytics tools where no equation or pre defined model exists. The goal is to deduce ("learn") the model from the data. ML may be useful not only for managing and analyzing the big data of MS simulations but also as a new way to study systems and discover patterns that may lead to insights about the case under investigation. ML has been already used in MS in many different ways from post processing, to preparation of input parameters and the error reduction of simulation itself [5–9].

In the current paper, we introduce a computational tool of analyzing random oriented lipid bilayers derived from MD trajectories and creating a dataset ready to be used in ML algorithms. The initial data consist of spontaneous self-assembly structures of the lipid bilayer using MD simulations.

2 CAPABILITIES AND IMPLEMENTATION

The workflow under discussion consists of three distinct steps; (1) the analyzing of the MD trajectories, (2) the creation of the ML ready dataset and (3) the use of the dataset to ML algorithms (see Figure 1).

The tool is written in Python3 programming language and provides a dynamic input interface, that is capable of filling the requirements of each user case. The user have to describe the atom groups and the primary analysis for each group. Moreover, the input interface enables the combination of the results of primary analysis in order to calculate secondary properties for the system. The aforementioned inputs need to be written in python dictionary format.

In order to address the problem of different oriented and shaped lipid bilayer, which is the result of self assemblage (see Section 3), the tool performs a domain decomposition of the final configuration and identifies the atoms that belong to the user defined groups. Each group and domain becomes a sub-system that will be analyzed as a unique MD system. As such, each MD simulation may create more than one sub-systems, hence, instances in the final dataset. By breaking the system to small domains, where the assumption of no curvature, no intersection point etc can be applied, the conformation is treated as an ideal bilayer structure, and a series of MD analysis tools can be used. The resulted dataset can be used

AINST2020, September 02–04, 2020, Athens, Greece

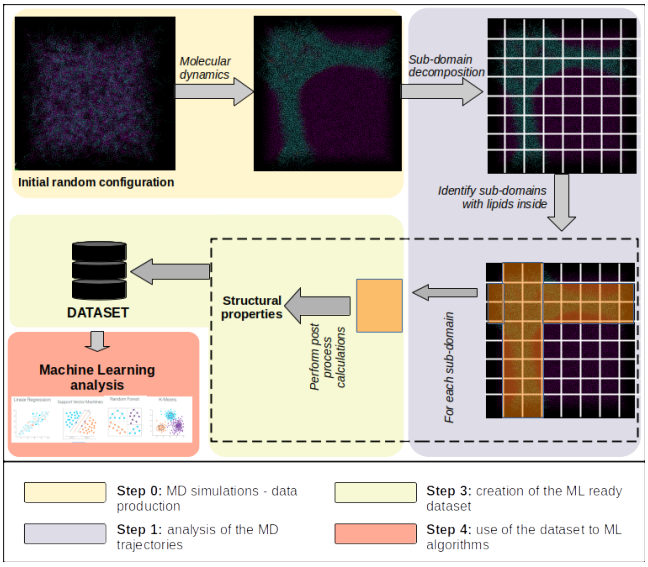

Figure 1: Illustration of the workflow process of the presented tool.

as input to ML algorithms, which enable to patterns' identification and gain insights for large and complex bilayer structures.

The tool can load efficiently trajectory and/or topology data from the format used in GROMACS [10] MD simulation tool and use many post-process tools that GROMACS provides, alongside customized calculation (primary or secondary) in order to calculate a series of properties for each sub-system. The structural characteristics that are calculated for each sub-system, are the peaks of density profile, the tilt of the order part of the order part of lipid chain, the peaks of radial distribution function of pairs of lipid groups and the order parameter of lipid chains.

3 CASE STUDY

The orientation of the lipid bilayer can be studied through MD calculations. However, such treatments are based on the a-priori placement of the lipid molecules in appropriate positions, in order to form a periodical system with appropriately oriented hydrophilic chains and hydrophobic groups. Despite the fact that the aforementioned formation saves a substantial amount of simulation time, it only represents a simplified and ideal approximation of the formation in equilibrium and does not ensure that its structural and dynamical properties are simulating accordingly the real/natural lipid phase of the system under study.

Other approaches [11] recreate the structure of the lipid bilayer using MD with random initial configurations of the molecules. This treatment aims to study the dynamics of the system while it moves towards equilibrium and to the spontaneous self-assembly of the single lipids into a bilayer, as well as simulate more realistic conformations of minimum energy. Thus, any approximation based on the a-priori placement of the lipids will be eliminated. Due to the randomness of the initial configuration which affect the resulted structure, a sufficient sampling of self-assembled systems need to be produced (10^2 order of magnitude). All of the resulted systems are far from ideal, in terms of shape and orientation, and the properties are correlated by the local composition, shape, orientation, bilayer thickness etc. The provided tools of analyzing MD trajectories lack the functionality to processing random oriented and shaped bilayer structures. The tool presented in the current paper attempts to address that problem by decompose the each simulation resulted conformation in small sub-domains, calculating structural properties of each sub-domain, such as density profile, order parameter, radial distribution function, tilt of lipid chains, and producing a ML ready dataset in order to apply data driven techniques, such as classification or clustering. The ML techniques will provide a fast, efficient and unbias way to group the different sub-domains and it will try to identify and extract the physical meaning of each resulted group via their properties. That information will lead to a recommendation of a series of distinct and well defined bilayer structure that exist simultaneously in the macroscopic the system. The recommended conformation can be reconstructed and can be taken into account in future studies of the system.

4 DISCUSSION

The capabilities of the tool serve as a bridge, connecting MD data with structural properties and ML algorithms for general data science audiences. The derived measurements constitute a domain dataset, aiming to feed ML algorithms and (i) explore patterns that may emerge by applying unsupervised learning algorithms or (ii) build a model that predicts a property of interest. Moreover, the calculated properties can be used as supplement data to a larger dataset. The tool stands out for the novel approach of examining the system as a series of sub-system, thus surpassing the problems and limitations of analyzing complex lipid bilayer structures.

The tool's development state is an alpha version, where tests and debugging are performed. As future work, the outcome and results of the a case study is planned to be used, alongside the first public release of the code under free and open source license. The tool is hosted at: https://mssg.ipta.demokritos.gr/gitlab/skarozis/toobba

ACKNOWLEDGMENTS

This research is co-financed by Greece and the European Union (European Social Fund - ESF) through the Operational Programme «Human Resources Development, Education and Lifelong Learning» in the context of the project "Reinforcement of Postdoctoral Researchers - 2nd Cycle" (MIS-5033021), implemented by the State Scholarships Foundation (IKY).

REFERENCES

[1] R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane, and V. S. Pande, "MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories," *Biophysical Journal*, vol. 109, pp. 1528–1532, oct 2015.

[2] T. Tubiana, J.-C. Carvaillo, Y. Boulard, and S. Bressanelli, "TTClust: A Versatile Molecular Simulation Trajectory Clustering Program with Graphical Summaries," *Journal of Chemical Information and Modeling*, vol. 58, pp. 2178–2182, nov 2018.

[3] M. Brehm, M. Thomas, S. Gehrke, and B. Kirchner, "TRAVIS—A free analyzer for trajectories from molecular simulation," *The Journal of Chemical Physics*, vol. 152, p. 164105, apr 2020.

[4] A. Shkurti, R. Goni, P. Andrio, E. Breitmoser, I. Bethune, M. Orozco, and C. A. Laughton, "pyPcazip: A PCA-based toolkit for compression and analysis of molecular simulation data," *SoftwareX*, vol. 5, pp. 44–50, 2016.

[5] E. Swann, B. Sun, D. M. Cleland, and A. S. Barnard, "Representing molecular and materials data for unsupervised machine learning," *Molecular Simulation*, vol. 44,

pp. 905–920, jul 2018.

[6] B. K. Carpenter, G. S. Ezra, S. C. Farantos, Z. C. Kramer, and S. Wiggins, "Empirical Classification of Trajectory Data: An Opportunity for the Use of Machine Learning in Molecular Dynamics," *The Journal of Physical Chemistry B*, vol. 122, pp. 3230–3241, apr 2018.

[7] M. Haghighatlari and J. Hachmann, "Advances of machine learning in molecular modeling and simulation," *Current Opinion in Chemical Engineering*, vol. 23, pp. 51–57, 2019.

[8] H. Sidky, W. Chen, and A. L. Ferguson, "Machine learning for collective variable discovery and enhanced sampling in biomolecular simulation," *Molecular Physics*, vol. 118, p. e1737742, mar 2020.

[9] F. Noé, A. Tkatchenko, K.-R. Müller, and C. Clementi, "Machine Learning for Molecular Simulation," *Annual Review of Physical Chemistry*, vol. 71, pp. 361–390, apr 2020.

[10] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, "GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers," *SoftwareX*, vol. 1-2, pp. 19–25, sep 2015.

[11] S. N. Karozis, E. I. Mavroudakis, G. C. Charalambopoulou, and M. E. Kainourgiakis, "Molecular simulations of self-assembled ceramide bilayers: comparison of structural and barrier properties," *Molecular Simulation*, vol. 46, pp. 323–331, mar 2020.