MATDAT18: Materials and Data Science Hackathon
MATERIALS SCIENCE TEAM APPLICATION FORM

Complete and return via email to brian_reich@ncsu.edu **by 19 January 2018**

Team Composition (2 people max.)

Name	**Department**	**Institution**	**Email**
Sapna Sarupria	Chemical and Biomolecular Eng.	Clemson University	ssarupr@clemson.edu
Ryan DeFever	Chemical and Biomolecular Eng.	Clemson University	rdefeve@clemson.edu

Project Title

Using machine learning to enhance the efficiency of rare event sampling methods

Project Synopsis (approx. 100 words)

Our research focuses on studying phase transitions in aqueous systems such as gas hydrates nucleation and develop methods that facilitate these studies in molecular simulations. We have used advanced sampling techniques, specifically forward flux sampling combined with molecular dynamics to generate statistically relevant number of transition paths capturing the liquid-to-hydrate transition. In this process, we have generated significant amounts of data in the form of transition paths and configurations in the phase space between the metastable liquid and hydrate structure. In our analysis process, we have discovered that traditional methods of characterizing hydrate structures limit the information we can extract from this data. Therefore, we are interested in developing machine learning techniques to characterize these configurations and transition paths. This will enable us to discover structural features hidden through traditional analysis and gain further insights into the mechanisms of the transition.

Identified Data-Science Collaborative Need (approx. 100 words)

Sarupria group has significant experience in performing rare event simulations and in the field of nucleation in aqueous systems. We however, are not experts in machine learning algorithms and therefore, would like to participate in the MATDAT workshop. We hope to establish collaborations with experts in machine learning and related methodologies to assess, apply and develop machine learning techniques for structural determination. The sampling from our large-scale FFS calculations provide tremendous datasets for this pursuit. The approaches developed in the scope of our work have the potential to significantly improve the efficiency of sampling rare events in molecular simulations.

Data Origin and Access (*data must be available and sharable with data science teams* – please **address: data source/origin, access privileges, sharing privileges)**

The data we will use for this project was generated by Sarupria group at Clemson and we have complete access and ownership of the data. The data will comprise of the configurations generated from forward flux sampling calculations. These are output from GROMACS (MD simulation software) and are available both in the binary form and in ASCII-form. The data is hosted on the Clemson supercomputer and will be accessible to us. We will make it accessible to our data science partners as needed. We have the privileges to modify the data (i.e. convert formats etc if needed) as well as to share the data.

Using machine learning to enhance the efficiency of rare event sampling methods

I. Project Description (approx. 1.5 pages, plus figures and references; please describe data size, form, dimensionality, uncertainties, number of examples, etc.)

Our research focuses on developing computationally efficient methods to sample rare events. Rare events refer to the class of events that have a low probability of occurance within accessible observation time but have tremendous impact when they occur. In molecular simulations, an example of a rare event is nucleation – birth of a new phase from a metastable phase. We use advanced sampling technique called Forward Flux Sampling (FFS)[1,2] to sample such rare events in molecular simulations. FFS breaks down the initial-to-final state transition into a series of transitions between interfaces that define sub-regions of phase space between the initial and final state. We have developed software that integrates the implementation of FFS with the capabilities of Hadoop to perform large scale FFS calculations with computational and user efficiency in high-performance computing environment.[3,4] This has enabled us to perform some of the largest scale FFS calculations on various processes namely crystallization of Lennard Jones like particles, heterogeneous nucleation of ice on different surfaces and nucleation of gas hydrate structures from THF-like solute and water solution[5]. However, in doing so we have discovered that although we have generated significant amount of data the information we are able to extract from it is limited. We illustrate this by considering our studies on gas hydrate nucleation.

Gas hydrates are crystalline structures formed by guest molecules entrapped in cages formed through hydrogen bonding between water molecules (Fig 1). The molecular process through which this occurs remains an open question. While several straightforward molecular dynamics (MD) simulations have been performed to probe the mechanisms, their insights are limited due to the few nucleation events sampled. In our study,[5] we used FFS to generate a statistically relevant number of nucleation trajectories. Further, we performed extensive committor probability analysis to characterize the transition state and hence, develop insights into the reaction coordinates associated with this phase transition and the mechanism underlying gas hydrate nucleation. Overall, we generated 1101 transition paths, 4189 configurations that are part of the transition paths and 10,099 configurations that were sampled along the transition but did not make it to the final state. Interestingly, we found that all the transition paths came from 8 initial configurations (configurations at the first interface) even though we have 778 configurations at that interface. Even more intriguing is the fact that of those 8 configurations, one spawned over 1021 transition paths.

The obvious and key question that arises from our results is – what makes the 8 initial configurations reactive relative to the other 770 configurations? What is so special about that one configuration amongst the 8 reactive configurations? (By most available methods to assess basin (i.e. initial) sampling, we find that our basin sampling is sufficient.) To probe this, we visualized and characterized the initial configurations using most available methods to identify hydrate-like structures. We evaluated the structures based on 33 order parameters and linear combinations thereof. Snapshots of our results are shown in Fig. 2. The values of five different order parameters for all the 778 configurations at the first interface are shown. The reactive configurations are shown in red and blue triangles, where the red triangle is the configuration that spawned majority of the transition paths. Interestingly, no distinct feature of the reactive versus non-reactive configurations was obtained. This highlights the need for methods that enable us to extract structural features from a given dataset without prior knowledge. Such methods will truly enable us to maximize the information we can extract from the dataset of configurations we have generated using FFS.

Motivated by this, the goal of our research is to develop approaches to be able to (i) identify key structural features, which appear to remain hidden in the current hydrate structure identification methods. (ii) identify and correlate those structural features to reactive versus non-reactive trajectories. Based on the previous

work in related fields (specifically, ice nucleation[6] and reaction coordinate determination[7]), we hypothesize that machine learning techniques can provide us such approaches.

Geiger and Dellago[6] recently demonstrated that properly trained artificial neural networks can be used for structure detection. They were able to detect amorphous and crystalline structures with high accuracy even in cases of complicated atomic arrangements, such as ice structures, for which traditional structure detection can become unreliable. The strength of this approach is that several basic units of structural fingerprints can be assessed for a given configuration enabling the search of subtle structural patterns. The weakness however, is that it requires the knowledge of reference structures to train the neural network. This is challenging in case of gas hydrates where the nucleation process is thought to occur in multiple steps such as metastable solution to amorphous solid; and amorphous solid to crystalline structure.[8–11] Recently, Ferguson and coworkers[12] have developed a methodology for structure determination without reference structures for colloidal systems. We anticipate that similar approaches could be applied to systems such as gas hydrates and ice nucleation.

Machine learning approaches have also been used to detect transition states and identify the best reaction coordinate. Reaction coordinate determination remains one of the biggest challenge in rare event simulations and most robust methods are computationally prohibitive. Ma and Dinner[7] developed neural network based method to determine the functional dependence of the committor probability (p_B) on a set of coordinates. The neural network used as input the data from transition paths sampled using transition path sampling technique and calculations of p_B performed on configurations selected from the transition paths. Using this approach, the authors were able to screen through $>$5000 candidates to four key physical variables to describe C_{7eq}-to-α_R isomerization of the alanine dipeptide. We anticipate that we could use similar approaches to assess reactive versus non-reactive configurations obtained from our FFS calculations of gas hydrates. We have performed careful calculations of the committor probability for the configurations obtained from FFS sampling. This provides us good dataset for developing machine learning methods to differentiate reactive vs non-reactive configurations.

It is worth mentioning that we also have such FFS based data along with committor probability calculations for crystal nucleation in Lennard Jones like particles. These can provide us a simpler test case system to begin developing the machine learning methods prior to testing them on more complex structures such as gas hydrate crystals. The ability to identify hidden structural features and relate them to reactive versus non-reactive trajectories through machine learning can provide a route to further enhancing the sampling techniques for rare events. This will make it possible for us to study processes and transitions that continue to remain beyond the reach of molecular simulations. This is the long-term goal of our research.

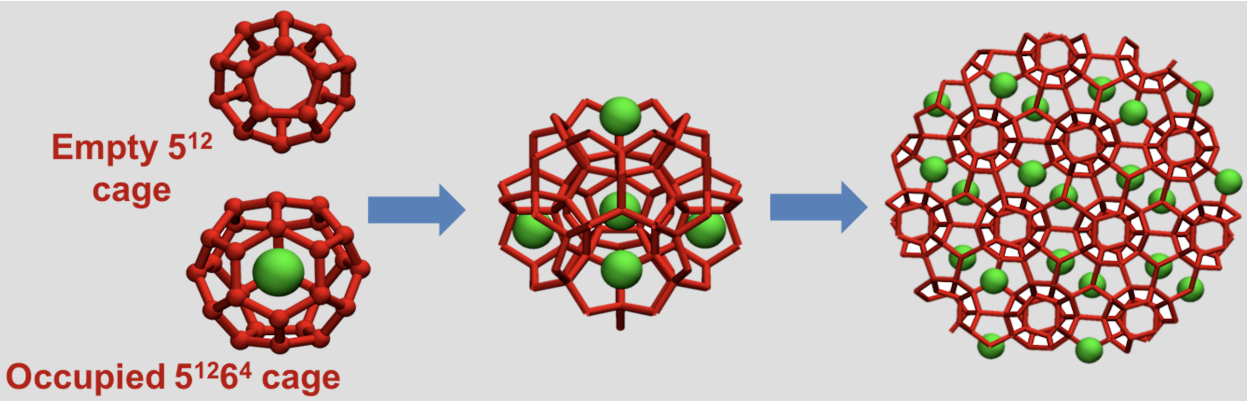

Figure 1: Schematic illustrating structure of gas hydrates. Left panel: cages that form the basic building units of hydrate crystal. Middle panel: Unit cell of a hydrate cyrstal structure. Right panel: sII hydrate structure. The red spheres indicate water molecules adn the red bonds represent hydrogen bonds between the water molecules. Green spheres indicate guest molecules.

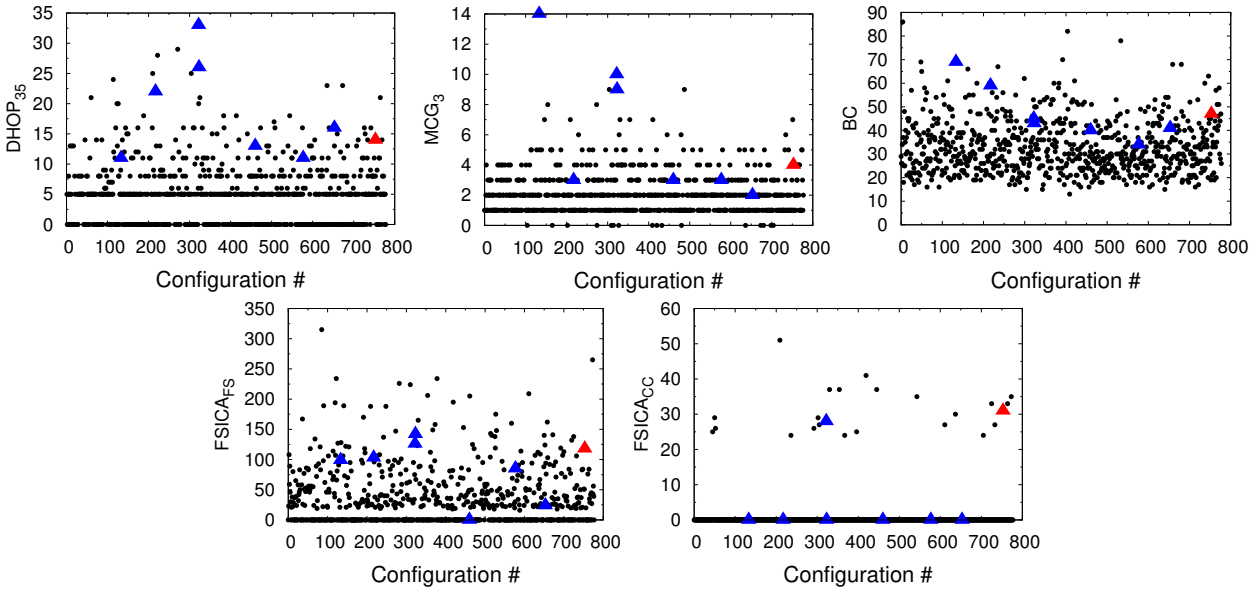

Figure 2: Value of several order parameters (OP) for all 778 configurations at initial interface. The configuration that spawned majority of the transitions paths are shown as red triangle and the seven other configurations that spawned at least one transition path are shown as blue triangles. All other configurations are shown with black points. DHOP_{35}: OP based on dihedral angle between water molecules; MCG_3: OP based on mutually-coordinated guest molecules; BC: OP based on tetrahedrality and 5-membered rings of water molecules; FSICA_{FS}: OP based on planar rings the water molecule participates in; FSICA_{CC}: OP based on number of complete cages the water molecule participates in. The OP are used to characterized the water molecule as hydrate-like or not and then the largest cluster size of hydrate-like water molecules is calculated. This is the reported y-axis value. Reproduced from Ref.[5]

References Cited

[1] C. Valeriani, R. J Allen, M. J. Morelli, D. Frenkel, and P. R. ten Wolde. Computing stationary distributions in equilibrium and nonequilibrium systems with forward flux sampling. *J. Chem. Phys.*, 326(11):114109, 2007.

[2] R. J. Allen, C. Valeriani, and P. R. ten Wolde. Forward flux sampling for rare event simulations. *J. Phys. Cond. Matt.*, 21(56):463102, 2009.

[3] P. Xuan, Z. Yueli, S. Sarupria, and A. Apon. Sciflow: A dataflow-driven model architecture for scientific computing using hadoop. In *IEEE BigData Conference: Big Data and Science*, 2013.

[4] W. Hanger, R. S. DeFever, L. Ngo, A. Apon, and S. Sarupria. Scalable Forward Flux Sampling, ScaFFS: Software platform to study rare events in molecular simulations. In *Supercomputing 2015 Conference*, 2015.

[5] DeFever, R. and Sarupria, S. Nucleation mechanism of clathrate hydrates of water-soluble guest molecules. *J. Chem. Phys.*, 147:204503, 2017.

[6] P. Geiger and C. Dellago. Neural networks for local structure detection in polymorphic systems. *J. Chem. Phys.*, 139:164105, 2013.

[7] A. Ma and A. R. Dinner. Automatic method for identifying reaction coordinates in complex systems. *J. Phys. Chem. B*, 109:6769, 2005.

[8] M. R. Walsh, C. A. Koh, E. D. Sloan, A. K. Sum, and D. T. Wu. Microsecond simulations of spontaneous methane hydrate nucleation and growth. *Science*, 326:1095–1098, 2009.

[9] L. C. Jacobson, W. Hujo, and V. Molinero. Amorphous precursors in the nucleation of clathrate hydrates. *J. Amer. Chem. Soc.*, 132:11806–11811, 2010.

[10] J. Vatamanu and P. G. Kusalik. Observation of two-step nucleation in methane hydrates. *Phys. Chem. Chem. Phys.*, 12:15065–15072, 2010.

[11] S. Sarupria and P. G. Debenedetti. Homogeneous nucleation of methane hydrate in microsecond molecular dynamics simulations . *J. Phys. Chem. Lett.*, 3:2942–2947, 2012.

[12] W. F. Reinhard, A. W. Long, M. P. Howard, A. Ferguson, and A. Z. Panagiotopoulos. Machine learning for autonomous crystal structure identification. *Soft Matter*, 13:4733–4745, 2017.