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

Complete and return via email to brian_reich@ncsu.edu by 15 January 2018

Team Composition (2 people max.)

Name	Department	Institution	Email
Nicholas Rego (graduate student)	Biochemistry and Molecular Biophysics	University of Pennsylvania	nrego@mail.med.upenn.edu
Zachary Varley (undergraduate) OR Amish Patel (PI)	Chemical and Biomolecular Engineering	University of Pennsylvania	zvarley@seas.upenn.edu / amish.patel@seas.upenn.edu

Project Title

Characterizing Protein Hydrophobicity Using High-Dimensional Descriptors

Project Synopsis (approx. 100 words)

Protein hydrophobicity informs its interactions and assemblies. However, empirical approaches for describing protein hydrophobicity using low-dimensional descriptors have failed to capture it with sufficient accuracy. Recent work has explained this failure by highlighting that protein hydrophobicity depends sensitively on the nanoscopic topographical and chemical pattern displayed by protein surfaces. Capturing such high dimensional protein hydrophobicity will require a combination of molecular simulations (with enhanced sampling approaches) to generate the requisite data, and the state-of-the-art data science approaches to capture the complex functionality present in the data. If successful, this work will open up applications in the high-throughput screening of ligands for drug discovery as well as the high-throughput prediction of protein interaction interfaces.

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

Low (<10) - dimensional descriptors have failed to reliably predict protein hydrophobicity. An understanding of the molecular determinants of protein hydrophobicity and the development of molecular simulation methods for characterizing such determinants, along with advances in high-performance computing have opened the door for the development of high (>100)-dimensional descriptors for predicting protein hydrophobicity. Deep learning appears to be well suited for addressing this challenge. Although enhanced sampling molecular simulation methods can generate the large amounts of requisite data, efficiently making use of this data to construct a computationally economical model for capturing protein hydrophobicity will require expert wielding of the data science toolkit: questions pertaining to the quality of data, the smallest number of dimensions needed, as well as the amount of training data that will be necessary will have to be addressed.

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

For select proteins (number of proteins, N = O(10) to begin with), the positions of all protein atoms (number of protein atoms, S is roughly 200) as well as their atom types will serve as input (to simulations as well as for machine learning). This data will be obtained from the Protein Data Bank, and will be used to perform biased molecular dynamics (MD) simulations, using an external potential that seeks to systematically displace waters from the protein hydration shell. Simulations will be performed for M = 10 – 20 potential strengths, and for every potential strength, the number of waters that remain in the individual protein atom hydration shells will be recorded. Our training data set will thus have dimensions of N x M x S, and will be obtained from N x M biased MD simulations. The data and will be shared with data science teams in the format that is most convenient.

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

Introduction and Motivation The hydration and interactions of complex molecules, such as cavitands, dendrimers, drugs, and proteins, which display chemical and topographical heterogeneity at the nanoscale, play a central role in numerous phenomena, ranging from supramolecular host-guest chemistry to biomolecular recognition. The hydrophobic effect, which refers to the favorable interactions between non-polar moieties in water, plays an important role in the interactions and assemblies of such complex molecules. However, quantifying the hydrophobicity of such molecules in manner that informs their interactions has proven to be challenging. In particular, here we will focus on proteins as the archetypical heterogeneous molecules, and tackle the challenge of predicting their hydrophobicity.

Data Dimensionality Challenge Recent work from our group and others has shown that the challenge in accurately characterizing protein hydrophobicity, or how unfavorable protein-water interactions are, stems from the fact that proteins disrupt the inherent structure of water in countless different ways, depending not only on the chemistry of the underlying protein surface, but also on the precise topography and chemical pattern of amino acids. In other words, the chemical and topographical cues presented by the protein surface induce a collective, many-body response from water molecules, which is challenging to capture. This body of work also explains why protein hydrophobicity is not captured reliably by approaches, such as hydropathy scales, which employ low-dimensional descriptors (e.g., list of protein residues near protein patch of interest).

Data Quality Challenge All-atom molecular dynamics simulations with explicitly represented waters are capable of generating the large amounts of data that would be necessary for uncovering the high-dimensional description of protein hydrophobicity. However, the data generated by equilibrium molecular simulations is not adequate to inform protein hydrophobicity. Recent work has shown that the hydrophobicity of a surface is manifest, not in average quantities, e.g., water density near the surface, but in rare fluctuations away from the average; in particular, it is the statistics of low-density fluctuations, which result in the formation of a cavity near the surface, that captures its hydrophobicity. Because creating a cavity adjacent to a protein disrupts protein-water interactions, the corresponding free energy also serves as an estimator of the strength of those interactions; the easier it is to displace interfacial waters, the smaller the cavity formation free energy, and the more hydrophobic the protein surface.

Simulation Approach To characterize the strength of protein-water interactions or protein hydrophobicity in a way that captures the many-body water response and informs protein hydrophobicity, we perform all-atom protein simulations in explicit water, and systematically disrupt protein-water interactions by applying an unfavorable biasing potential, φN_V, which attempts to displace waters from the protein hydration shell; here, N_V, is the number of waters in the entire protein hydration shell, V. The response of the hydration waters to the strength of the applied potential, φ, contains a wealth of information, including the free energetic cost of disrupting protein-water interactions, and the order in which those interactions are disrupted. In particular, the regions of the protein that interact weakly with water (hydrophobic) dewet first (at low φ), whereas those that are highly hydrophilic, hold on to waters even at large φ. The collective water response obtained from such "φ-ensemble simulations" thus captures protein hydrophobicity, and enables the prediction of protein interactions.

Data Science Approach Deep learning provides an exciting avenue for capturing the many-body response of protein hydration waters, and predicting the high-dimensional protein hydrophobicity. However, to successfully address the protein hydrophobicity challenge using deep learning, a number of questions that bear on both the data science and the physics of the problem will first need to be addressed: How high-dimensional does the descriptor need to be? How much training data will be needed? How do we use enhanced sampling methods to efficiently obtain the requisite data?

Description of Data In addition to the molecular topology, the positions of all protein atoms (number of protein atoms, S is roughly 200) as well as their atom types (which encode non-bonded parameters, such LJ sigma, LJ epsilon, charge) serve as input to molecular simulations, and will serve as an upper bound on the dimensionality of the input node. For select proteins (number of proteins, N = O(10) to begin with), this data will be obtained from the Protein Data Bank, and will be used to perform the φ-ensemble simulations described above. As the strength of the potential is increased (number of φ-ensembles that will be simulated per protein, M = 10 - 20), protein hydration waters are systematically displaced, and for every potential strength, the number of waters that remain in the individual protein atom hydration shells will be recorded. Our training data set will thus have dimensions of N x M x S, and will be obtained from N x M biased MD simulations.