Foundational & Applied Data Science for Molecular and Materials Science & Engineering
Conference: May 22-24, 2019 | Iacocca Hall, Lehigh University

Poster Session

(presenters in bold)

Thursday, May 23, 2019 from 5:30pm

1. **COMPRESSING PHYSICAL PROPERTIES OF ATOMIC SPECIES FOR IMPROVING PREDICTIVE CHEMISTRY.**

   Authors: **John E. Herr**, Kevin Koh, Kun Yao, and John Parkhill, **Department of Chemistry and Biochemistry, The University of Notre Dame**.

   Abstract: Machine learning techniques are rapidly growing in popularity as a way to compress and explore chemical space efficiently. One of the most important aspects of machine learning techniques is representation through the feature vector, which should contain the most important information necessary to make accurate predictions. We introduce a vectorial representation of atomic species called the elemental modes for machine learning models. The elemental modes are derived by compressing physical properties unique to each element with an auto-encoder, thereby capturing underlying physical law in their features. We show that the elemental modes allow us to improve upon previous works in multiple ways including reducing data needs, scaling of costs, and opening up new potential uses of neural network potentials.

2. **STATISTICAL-LEARNING-ASSISTED FIRST-PRINCIPLES MODELING OF SINGLE ATOM CATALYSTS.**

   Authors: **Yifan Wang**, **Department of Chemical and Biomolecular Engineering, University of Delaware, Newark**, DE, Ya-qiong Su, Department of Chemical Engineering and Chemistry, Eindhoven University of Technology, Eindhoven, Netherlands Emiel Hensen, Schuit Institute of Catalysis, Department of Chemical Engineering and Chemistry,, Eindhoven University of Technology, Eindhoven, Netherlands Dionisios G. Vlachos, Chemical and Biomolecular Engineering, University of Delaware, Newark, DE.

   Abstract: Supported metal nanoparticles on oxide supports are widely applied as environmental catalysts. The catalyst activity is strongly affected by the nanoparticle size. Recently, single atom catalysts are being explored as effective and selective catalysts for several chemistries. Yet, there is often a strong debate whether single atoms, metal clusters or nanoparticles carry out the chemistry. The debate arises in part due to the inability to observe the dynamics of the catalysts operando and is further complicated by adsorbates, such as CO, which affect catalyst dynamics and sintering [1]. Density functional theory (DFT) is a powerful tool to predict the energetics of metal-support and adsorbate-metal interactions. However, the computational time needed to describe numerous clusters and sites and the long time scales for sintering make direct first-principles calculations impractical. To address this multiscale problem, we couple a three-dimensional cluster expansion (CE) with statistical learning tools such as LASSO and PCA to rapidly predict the energetic parameters using input from DFT calculations. Kinetic Monte Carlo (kMC) simulations are applied to model the dynamics of single atoms toward sintering. The methodology is applicable to any metal/support system. [1] Wang, Y., Mei, D., Glezakou, V., Li, J. & Rousseau, R. Nat. Commun. 6, 1–8 (2015).

3. **BAYESIAN EXPERIMENTAL DESIGN FOR MICROKINETIC MODELING OF HETEROGENEOUS CATALYTIC SYSTEM.**

   Authors: **Huijie Tian** and Srinivas Rangarajan, **Department of Chemical & Biomolecular Engineering, Lehigh University.**

   Abstract: Mean field microkinetic models, derived using density functional theory (DFT) calculations, are a powerful means to elucidate reaction mechanisms in heterogeneous catalysis. However, the intrinsic errors in DFT limit the predictive ability of microkinetic models. For a quantitative understanding of the reaction mechanism in

terms of reaction fluxes, most abundant surface intermediates, rate determining steps (RDS), etc., it is, therefore, necessary to correct for these errors. We introduce a machine learning model to address this problem. First, we build a Gaussian Process (GP) model to quantify the errors of the chosen functional (e.g. PW91). Then we use GP model to correct the DFT energies input to microkinetic models, and quantify the uncertainty in the predictions. An example of the water gas shift (WGS) reaction on a copper catalyst is used to demonstrate the framework. The result shows that the Gaussian Process model can correct the errors of the PW91 functional, and using the corrected energetics in mean-field microkinetic modeling is consistent with the experimental data, in terms of turnover frequency, apparent activation energy, and reaction order.

4. EFFICIENT QUANTUM MECHANICAL SCREENING OF PHOTOCHEMICAL DYES: A COMBINATORIAL APPROACH.

Authors: **Lisa A. Fredin**, **Department of Chemistry**, **Lehigh University** and Thomas C. Allison, National Institute of Standards and Technology.

Abstract: Though many transition metal light harvesting dyes are in use, the search continues for more efficient and effective compounds to reduce make commercially viable light conversion and storage devices. Computational methods have been increasingly applied to understand the dyes currently in use and to aid in the search for improved light harvesting compounds. Semi-empirical quantum chemistry methods have a well-deserved reputation for giving good quality results in a very short amount of computer time. The performance of newly optimized semi-emirical basis sets are tested against a set of molecules whose geometries were optimized using a density functional theory (DFT) method. Beyond geometry, the main metric for light harvesting dyes is their UV/Visible spectra. Here we have developed a method to calculate approximate UV/Vis spectra using a modified PM7 semi-empirical method. Using this re-parameterized method, a large set of transition metal centered light harvesters is screened, including expanded cage structures, electron donating and withdrawing groups, and various anchoring groups. These results clearly demonstrate the value of using semi-empirical methods to screen chromophore geometries and excitations.

5. A GENERALIZED DEEP LEARNING APPROACH FOR LOCAL STRUCTURE IDENTIFICATION IN MOLECULAR SIMULATIONS.

Authors: Ryan S. DeFever (a), Colin Targonski (b), **Steven W. Hall** (a), Melissa C. Smith (b), Sapna Sarupria (a) (a): Department of Chemical & Biomolecular Engineering, Clemson University, Clemson, SC 29634 (b): **Department of Electrical & Computer Engineering, Clemson University**, Clemson, SC 29634

Abstract: The interpretation of results from molecular simulations commonly relies upon the ability to reduce high-dimensional configurational data to relevant low-dimensional data from which inferences can be drawn about the nature of the process under study. An important component of this is the recognition of local structures, which can include crystal polymorphs and biomolecular conformations, among others. Typical measures of local structure rely on hand-picked, system-specific features that may only be able to distinguish a fraction of the relevant structures. We present a structural analysis method that utilizes the raw output of simulations—the atomic coordinates—to classify the local structure around each atom. In doing so, the process of developing and testing feature representations for each type of structure to be classified is eliminated. This approach is enabled by a type of neural network called a PointNet that takes as input point clouds, which are simply collections of coordinates. We show that the method can accurately and simultaneously distinguish between many crystal polymorphs and the liquid using three molecular systems: Lennard-Jones particles (four phases), water (eight phases), and a mesophase-forming binary mixture (six phases). To further demonstrate the broad applicability of the method, we characterize the hydrophobicity of protein surfaces based on the surrounding water structure and orientation.

6. IDENTIFYING THE DESCRIPTORS FOR ASSESSING THE NON-IDEALIY IN IONIC LIQUID-IONIC LIQUID MIXTURES.

Authors: Utkarsh Kapoor and **Jindal K. Shah, School of Chemical Engineering, Oklahoma State University.**

Abstract: Changing the identify of the cation, anion or the pendant group on the ions has been a conventional approach in exploiting the potential designer capability of ionic liquids. Such modifications are likely to yield ionic liquids with only discreet changes in the desired ionic liquid properties. Blending ionic liquid to form ionic liquid-ionic liquid mixtures has the potential to overcome such limitations and provides a simple yet an effective way of generating new ionic liquids. An open question in this domain is whether these ionic liquid mixtures form ideal solutions. In this contribution, we will demonstrate that many ionic liquid mixtures defy the conventional thinking

of ideality in terms of excess molar volumes and excess enthalpies. We will further show that molecular-level heterogeneities identified from molecular simulations can aid in classifying ionic liquid mixtures into ideal vs. non-ideal. Based on a large number of ionic liquid-ionic liquid mixture simulations, we have identified that the non-ideal behavior of ionic liquid mixtures depends on two descriptors: the difference in the hydrogen bonding ability of the anions and the ionic liquid molar volumes.

7. MOLECULAR SIMULATIONS OF LIQUID-LIQUID PHASE SEPARATION OF DISORDERED PROTEINS.

Authors: **Gregory L. Dignon**, Wenwei Zheng, Young C. Kim, Robert B. Best, Jeetain Mittal, **Department of Chemical Engineering, Lehigh University.**

Abstract: Cellular compartmentalization may occur through phase separation of proteins and RNA, commonly resulting in liquid-like proteinaceous assemblies within the cell, termed membraneless organelles. Many such membraneless organelles contain intrinsically disordered proteins, which lack a regular folded structure. To understand the relationship between amino acid sequence and phase separation, we develop and apply knowledge-based coarse-grained protein models to make predictions about the phase behavior by explicitly modeling amino acid sequences.

8. COMPUTING FLUID-PARTICLE INTERACTION FORCES FOR NANO-SUSPENSION DROPLET SPREADING: MOLECULAR DYNAMICS SIMULATIONS.

Authors: **Weizhou Zhou**, Baiou Shi, Edmund Webb III, **Department of Mechanical Engineering, Lehigh University.**

Abstract: Recently, there are many experimental and theoretical studies to understand and control the dynamic spreading of nano-suspension droplets on solid surfaces. However, fundamental understanding of driving forces dictating the kinetics of nano-suspension wetting and spreading, especially capillary forces that manifest during the process, is lacking. Here, we present results from atomic scale simulations that were used to compute forces between suspended particles and advancing liquid fronts.~ The role of nano-particle size, particle loading, and interaction strength on forces computed from simulations will be discussed. Results demonstrate that increasing the particle size dramatically changes observed wetting behavior from depinning to pinning.~ From simulations on varying particle size, a relationship between computed forces and particle size is advanced and compared to existing expressions in the literature. High particle loading significantly slowed spreading kinetics, by introducing tortuous transport paths for liquid delivery to the advancing contact line. Lastly, we show how weakening the interaction between the particle and the underlying substrate can change a system from exhibiting pinning behavior to de-pinning.

9. DATA-DRIVEN INVESTIGATIONS OF HIGH-ENTROPY ALLOYS.

Authors: **Ankit Roy** and Ganesh Balasubramanian, **Department of Mechanical Engineering and Mechanics, Lehigh University.**

Abstract: Edisonian approach can take decades to explore the design space sufficiently. For accelerated exploration of HEAs (high entropy alloys), less time consuming and data intensive techniques are called for which can make predictions of properties for any given composition of HEAs based on the already established data available in literature and other materials data base. Techniques have been divided into thermodynamic, machine learning (ML), combinatorial and calculation of phase diagram approaches followed by a comparison between the aforementioned.

10. LABEL FREE DETECTION OF RARE CIRCULATING TUMOR CELL FROM IMAGE MACHINE LEARNING.

Authors: **Shen Wang, Department of Mechanical Engineering and Mechanics, Lehigh University**
Dr. Yaling Liu, MEM BioE, Lehigh University.

Abstract: The characteristics of rare circulating tumor cell (CTC) in patient's blood is important information regarding the diagnosis of cancer and further treatment. A traditional way of counting CTC via the comparison between bright-field and fluorescent images is usually under a series of experimental procedures. Here we present a label-free detection of CTC from patient blood sample using convolutional neural network based on microscopy

images. The model yields 90% overall accuracy over limited amount of raw data. This technique enables a quicker and simpler process for counting and locating CTC. With more data being input, the learning model will become a more convincing for CTC analysis.

11. DATA-DRIVEN CHARACTERIZATION AND PREDICTION OF THE MOLECULAR DETERMINANTS OF HYDROPHOBICITY AT THE NANOSCALE.

Authors: **Nicholas Rego**, Amish Patel, Andrew L. Ferguson, **University of Pennsylvania.**

Abstract: Intermolecular interactions in water underlie many key processes in biology, material science, and beyond. A key factor determining the strength of such intermolecular associations is the hydrophobic effect - the thermodynamic driving force of association of hydrophobic moieties in water. However, hydrophobicity at the nanoscale is determined by the specific chemical and topographical patterns presented by the surface. The relationship between a particular chemical pattern of a surface and its underlying hydrophobicity is complex and non-trivial and predicting the relative hydrophobicity of surfaces based on chemical and topographical details alone has remained elusive; a comprehensive prescription for hydrophobicity must account for the many-bodied, collective response of water. As such, additivity and implicit-solvent models are ill-equipped to capture the nuances of water's interactions with chemically diverse surfaces. While recent advances in all-atom, explicit solvent molecular dynamics simulations have demonstrated a means to quantify the hydrophobicity of complex surfaces, predicting hydrophobicity based on surface chemistry alone has remained challenging. Here, we combine explicit-solvent simulation results quantifying the hydrophobicity of a number of chemically diverse patterned surfaces with machine learning to construct intuitive models to predict hydrophobicity based on surface patterns alone. We show that while the simplest additive models are insufficient to accurately distinguish hydrophobicity of different surfaces, with relatively simple adjustments, motivated by physical intuition, we are able to construct simple models that are capable of determining hydrophobicity of simple surfaces with striking accuracy. These results could provide valuable insight into the underlying nano-scale molecular determinants of hydrophobicity.

12. DESIGNING OPTIMAL TRAINING SETS FOR DATA-DRIVEN MOLECULAR PROPERTY PREDICTION.

Authors: **Bowen Li** & Srinivas Rangarajan, **Department of Chemical and Biomolecular Engineering, Lehigh University.**

Abstract: We consider the problem of designing the training set with the most informative molecules in a specified library to build data-driven molecular property models with least computational effort. Using a generalized sparse group additivity and kernel ridge regression as two representative classes of models, we propose a method combining rigorous model-based design of experiments and cheminformatics-based diversity-maximizing subset selection within the epsilon greedy framework to systematically minimize the amount of data needed to train models. We demonstrate the effectiveness of the algorithm on subsets of various databases, including QM7, QM9, NIST, and a catalysis dataset. For group contribution models, a balance between exploration (diversity-maximizing selection) and exploitation (D-optimality selection) leads to learning with a fraction (sometimes as little as 10\%) of the data to achieve similar accuracy as five-fold cross validation on the entire set. On the other hand, kernel ridge regression prefers diversity-maximizing selections.