CLEAN ENERGY INSTITUTE
UNIVERSITY of WASHINGTON

DIRECT
Data Intensive Research
Enabling Clean Technologies

DIRECT: Data Intensive Research Enabling Clean Technologies

A graduate training program of the Clean Energy Institue for UW graduate students interested in data-enabled discovery and design of materials for clean energy.

Discovering new materials that will generate and store renewable energy in a low-cost, environmentally-benign, and scalable fashion is perhaps the most important technological challenge facing society today. However, all phases of this scientific process – design, synthesis, and characterization – are routinely stymied by the same challenge: researchers are not equipped to handle the deluge of data coming from their labs and high performance computers. DIRECT, a Clean Energy Institute training program, funded by the National Science Foundation, provides training that will equip a new generation of energy researchers to handle the massive data sets arising from all stages of materials discovery.

DIRECT provides UW graduate students the opportunity to receive data science training, practice applying new tools and skills in a project-based learning environment, and gain experience working with industry and academic partners on a team-based capstone project.

WHAT IS A CAPSTONE PROJECT?

The final step of the DIRECT program is completing a team capstone project, where trainees tackle real-world problems for clients from industry, Pacific Northwest National Laboratory, and UW labs. Trainees apply machine learning and data science skills developed during the DIRECT program to projects related to data management, processing and statistics, while working to advance research and discovery in materials for clean energy.

HAVE A DATA INTENSIVE PROJECT?

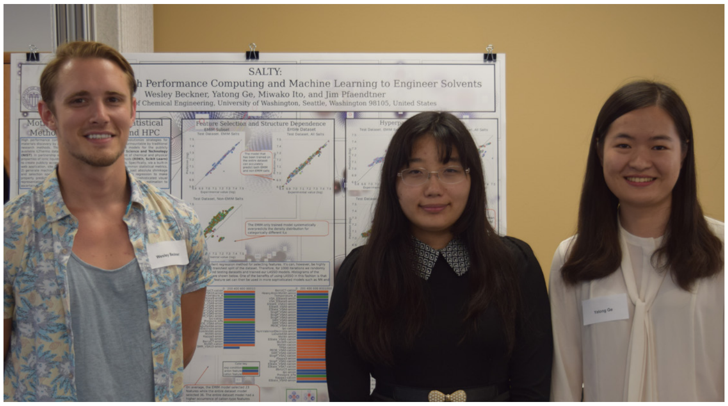
We will be looking for new projects for cohort 3 starting in Autumn 2018. If you have a data intensive project for our trainees, please contact us:

Jim Pfaendtner, PI and Director, jpfaendt@uw.edu
Kelly Thornton, Program Manager, kthorn2@uw.edu

2018 CAPSTONE PROJECTS:

Automating Molecular Bond Perception Through Natural Population Analysis
Project Sponsor: Pacific Northwest National Laboratory
Project Team: Luke Gibson (Chemical Engineering), Nathan Laurie (Materials Science and Engineering), Chenggang Xi (Chemical Engineering)

Currently, there exists tools for interpreting a file containing atomic XYZ coordinates and producing a molecular graph describing the bonding topologies. However, these tools operate on heuristic algorithms, which do not always not lead to correct bond assignments. We have created a tool (Bond Locator Utilizing Electronic Structure - BLUES) which operates without these restraints and relies on natural population analysis to determine occupancies of electrons in molecular orbitals, assign bond orders, and ultimately produce a chemical identifier string (SMILES and/or InChI). In addition to this workflow, BLUES also features 2D and 3D interactive molecular orbital visualization tools and can be used to find systematic errors in other open-source packages conversion algorithms. Our tool will aid in high-throughput analysis and design of molecules.


SALTY:
h Performance Computing and Machine Learning to Engineer Solvents

Wesley Beckner, Yatong Ge, Miwako Ito present their 2017 capstone project titled *Salty: Application of Performance Computing and Machine Learning to Engineer Solvents.*

2018 CAPSTONE PROJECTS CONTINUED:

Binary Classification of Protein Expression from DNA and Amino Acid Sequences Using a Long-Short Term Memory/Convolutional Neural Network
Project Sponsor: Jon Rue (Novo Nordisk)
Project Team: Jay Rutherford (Chemical Engineering), Josh Smith (Chemical Engineering), Christopher Nyambura (Chemical Engineering)

In medicine and biology, there are many applications requiring expression of synthetic proteins, such as biopharmaceuticals and proteomics. To find effective expression systems, high throughput experimentation is used to determine expression and solubility. Predicting properties of synthetic sequences using high performance computing can have a major impact on rapid screening of nucleotide sequences resulting in reduced usage of resources during optimization of biologic products made from transfected eukaryotic or prokaryotic expression platforms. Previous efforts to predict protein properties in silico typically rely on features derived from the protein nucleotide and amino acid sequences. In this work, we predict protein expression and solubility directly from the nucleotide or amino acid sequence using a neural network architecture that combines convolutional and long-short term memory layers. Our approach simplifies the process of protein property screening in silico and outperforms previously reported models on a classification task for an industrial-scale peptide expression data set. By predicting expression and solubility computationally, high throughput experiments can be focused on sequences with higher likelihood of successful expression, possibly resulting in faster process validation of large scale protein production.

Data Science Tool for Prediction of Chiller Plant Efficiency
Project Sponsor: Optimum Energy
Project Team: Ted Cohen (Molecular Engineering), Caitlin Parke (Chemical Engineering), Maitri Uppaluri (Chemical Engineering)

This project uses machine learning methods to predict the efficiency of chiller plants and rank operating parameters based on their contribution to plant performance. Currently, efficiency models are created in large Excel sheets that can take months to build. This python package allows users to easily clean large datasets while providing insight into the plant's operating conditions. It then uses gradient boosting machines for accurate prediction of chiller plant efficiency, which provides valuable insights for future plant optimization.

Deep Learning of Solid-State Transformations and Reaction Pathways in 2D Materials
Project Sponsor: Oak Ridge National Laboratory
Project Team: Sarthak Jariwala (Materials Science and Engineering), Jimin Qian (Materials Science and Engineering), Yiwen Wu (Materials Science and Engineering)

Recent advances in scanning transmission electron microscopy (STEM) have allowed unprecedented insight into the elementary mechanisms behind the solid-state phase transformations and reactions. However, the ability to quickly acquire large, high-resolution datasets has created a challenge for rapid physics-based analysis of STEM images and movies. In this project, we developed a convolutional-neural-network-based framework for automated localization, classification, and visualization of the defects in 2D materials from dynamic STEM data.

Detecting Sensor Drift or Failure for Chiller Plant Operations
Project Sponsor: Optimum Energy
Project Team: Victor Hu (Chemical Engineering), Zihao Tao (Materials Science and Engineering), Nicole Thompson (Chemical Engineering)

In chiller plants, there are a multitude of sensors that have to be in-calibration for safe, proper operation and optimized efficiency. Over time, these sensors can start to fail or slowly drift out of calibration, leading to inaccurate measurements and subsequent analyses for optimization. In this project sponsored by Optimum Energy LLC, we use a data-science approach to detecting sensors that are either failing or drifting out of calibration. We are using support vector regression as our generalizable machine learning model that can predict any sensor's output based off input from important surrounding sensors.

Molecules Design using Deep Generative Models
Project Sponsor: Baosen Zhang & Alex Jen (UW)
Project Team: Yize Chen (Electrical Engineering), Xiaoxiao Jia (Materials Science and Engineering), Jiaxu Qin (Molecular Engineering)

In this project, we propose to take advantage of recent advances in generative models in machine learning area to generate small molecules. An automatic machine learning pipeline for designing and generating new molecules by using generative adversarial network (GAN) model is developed, which contains both generator and discriminator. By converting discrete representations of molecules to a multidimensional continuous representation, we approximate sampling valid structures from the distribution of input molecules, and to generate new molecules for efficient exploration and optimization through open-ended spaces of chemical compounds.

Machine Learning Thermodynamics from Graphs of Condensed Molecular Systems

Project Sponsor: Xiaosing Li (UW), Aurora Clark (WSU)
Project Team: Torin Stetina (Chemistry), Andrew Wildman (Chemistry), You Chen (Materials Science and Engineering)

Understanding and predicting the relationship between structure and thermodynamic properties in condensed phase systems is key to producing next-generation materials and systems. The typical approach uses molecular dynamics simulations to achieve this goal, but the applied methods, such as thermodynamic integration, are extraordinarily computationally expensive. In this project, molecular interactions are reduced to graph objects with nodes and edges that fluctuate in time. Using that framework, we use a neural network to predict the distribution and population of energetic states from local graph-topologic parameters, giving us access to thermodynamic properties in a computationally efficient manner.

Predicting Sentiment on Climate Change Across the Globe using Deep Learning

Project Sponsor: KPMG
Project Team: Sarah Alamdari (Chemical Engineering), Neal Dawson-Elli (Chemical Engineering), Yusong Liu (Chemistry)

The advent of social media has created an influx of data to learn from. Thanks to advancements in powerful machine learning algorithms, models can be built to recognize and learn attitudes, emotions, and opinions of the human language. To this end, twitter data was processed using neural networks to understand and predict sentiment on tweets about climate change.

Solar PV Siting Survey for Anchorage, Alaska

Project Sponsor: Erin Whitney (Alaska Center for Energy and Power)
Project Team: Jingtian Zhang (Chemical Engineering), Cheng Zeng (Materials Science and Engineering), Yuening Wang (Chemical Engineering)

This project will identify potential solar sites to install solar panels on roofs and parking lots/ structures in Anchorage. Solar energy calculation model will be built based on packages in Arcgis Desktop, including slope, aspect and solar radiation. The input should be Light Detection and Ranging (LIDAR) data, and the output will be potential solar energy calculated by the model. Also, this project takes omission of HVAC system, vents and pipes on the rooftop, parking lots on the ground into consideration, making solar calculation more accurate. Eventually, this model might apply to other cities with similar situation.

ToxNet: Predicting Chemical Toxicity Using Deep Learning

Project Sponsor: Pacific Northwest National Laboratory
Project Team: Luocheng Huang (Materials Science and Engineering), Yiming Sui (Materials Science and Engineering), Sijia Xiao (Chemical Engineering)

The understanding of chemical toxicity is paramount. Chemicals are known to be carcinogenic, mutagenic, and neurotoxic, imposing considerable harm on humans and animals. Chemical toxicity testing traditionally involves animal testing, which is not only cruel but also costly. Hence, finding alternative solutions to predict chemical toxicity has become essential to the advancement of chemical engineering. By training deep learning neural networks, the preexisting toxicity data can be used to predict toxicity level on novel molecules. In our study, three types of neural networks are created to compare their strengths and weaknesses. First, a multi-layer perceptron neural network (MLP) was created using the descriptors straight from the chemical strings. Then, a recurrent neural network was created to train and predict directly on the chemical strings alone in an orderly fashion. Lastly, a convolutional neural network (CNN) was trained from 2D representations converted from chemical strings. The results from these models have ultimately shown a very promising future for predicting chemical toxicity using deep learning.

Understanding Photophysical Processes in Solar Cells via Recovering and Visualizing Spectral Signatures and Dynamics of Charge Carriers in Transient Absorption Spectroscopy

Project Sponsor: David Ginger (UW)
Project Team: Demi Liu (Chemistry), Ian Murphy (Chemistry), Jing Tu (Materials Science and Engineering)

We have designed an open-source package to identify self consistent spectrally and temporally evolving signatures of charge carriers after photoexcitation in transient absorption (TA) data. With minimal assumption, our algorithm identifies and visualizes the spectral and kinetics information of the individual population by combining methods such as multivariate adaptive regression spline fitting and data clustering.

Visualizing and analyzing 3D AFM force data to investigate solution structure at solid-liquid interfaces

Project Sponsor: Pacific Northwest National Laboratory
Project Team: Ellen Murphy (Materials Science and Engineering), Renlong Zheng (Materials Science and Engineering), Xueqiao Zhang

Recent developments in atomic force microscopy (AFM) technology allow us to directly image hydration layers at solid-liquid interfaces using 3D force mapping. The goal is to develop algorithms for data processing, visualizing, and analyzing the raw 3D force data to answer key scientific questions. Finally the project seeks to develop a GUI that users can interact with the data by viewing a full 3D rending, slices in 3D or 2D cartesian coordinate systems, and animations of these slices.