

# BIEN3300/ CENG3300 Fall 2023-24

## Syllabus - *Data Science for Molecular Engineering*

### Overview

With the explosion of information and fast development of AI technologies, data science has become an essential skill for scientists and engineers in various fields. While molecular engineering is a subject that heavily engages data collection, processing and analytics, data science is not systematically included in the curriculum. This course introduces the basics of data science in the context of molecular engineering, which is in particular relevant to BIEN, CENG, CEEV students. It aims to provide students with the necessary knowledge and skill sets to dig into masses amount of science and engineering data, with applications in rational experimental design, molecular discovery and optimization, and predictive modeling, etc. These are important preparations for the students to take more advanced data science courses, as well as take on challenges in their future careers.

### Course Intended Learning Outcomes

1. Identify problems that can be formulated as a supervised learning task
2. Process different types of data to be ready for model training
3. Understand the principles of supervised learning methods
4. Perform model training, validation and testing
5. Clearly interpret model predictions and present model results
6. Know the application of data science methods in molecular science related problems

### Tentative Weekly Schedule

#### Part 0. Introduction

Week 1 Real-world applications of data science in physical, chemical and life sciences

#### Part 1. Math Methods for Data Science

Week 2 Basic concepts in linear algebra, calculus and numerical optimization

Week 3 Review of statistics and probability

#### Part 2. Data Processing

Week 4 Introduction to Python and Pandas for Processing Tabulated Data

Week 5 Cheminformatics toolboxes – the example of RDKit

Week 6 Results visualization – molecules, chemical reactions, trees, graphs

#### Part 3. Supervised Learning

Week 7 Introduction to common supervised learning problems in molecular engineering

Week 8 Linear regression - molecular property predictions and feature importance

Week 9 Molecular feature engineering and nonlinear regression

Week 10 Introduction to deep learning and neural networks

#### **Part 4. Miscellaneous Topics**

Week 11 Introduction to unsupervised learning and reinforcement learning

Week 12 Guest lecture on real world applications of data science

Week 13 Course project presentation, Literature critique presentation

#### **Evaluation**

Homework 20%

Final exam 40%

Course project 25%

Literature critique 15%