**Practical AI/ML for Computational Biology and Chemistry Workshop**

June 13 - 17, 2022, 1:00pm – 5:00pm

Ammon Pinizzotto Biopharmaceutical Innovation Center conference room 140

Research is being revolutionized by methods from the field of Artificial Intelligence and Machine Learning (AI/ML). AI is the simulation of human intelligence processes (such as problem-solving, learning, and planning) by machines, while ML, a type of artificial intelligence, gives computers the ability to learn without explicitly being programmed. They have been applied in many areas including drug discovery, protein folding, and identification of variants from genetic data.

This workshop will provide you with

1. a conceptual understanding of various AI/ML approaches,
2. examples of practical applications of AI/ML in computational biology and chemistry,
3. hands-on exercises with emphasis on the importance of data preparation and readiness for AI/ML.

Through introductory lectures, you will be exposed to the basic concepts behind AI/ML approaches to better understand the practical applications of these data science tools, including what types of data can be used for a specific approach, and what types of outcomes can be expected. The illustrative practical applications will help you understand how these tools can be used in your research project. You will also learn about the importance of preparing the data according to the FAIR principles, namely, [Findability, Accessibility, Interoperability, and Reusability](https://pubmed.ncbi.nlm.nih.gov/26978244/). This will be addressed by providing you with first-hand experience on the issues arising when data that is not well-prepared, and covering various data formats, processing and wrangling techniques to get the data into a form where it can be utilized by Machine Learning algorithms. You will learn different visualization techniques to better understand the data at hand.

**Requirements:**

Bring your computer. Live demos and hands-on exercises will involve coding using [Google Colaboratory notebooks](https://colab.research.google.com/?utm_source=scs-index). Basic Python programming is recommended but not required.

The skills covered in the Data Carpentry Bootcamp that will be held in June 1-3 2022 meet this requirement.

**Schedule**:

Day 1 (June 13, 1:00pm – 5:00pm):

* Introduction to Workshop
  + - Welcome and roundtable introduction
    - Learning outcomes
    - Workshop overview
    - Workshop Logistics
* Lecture 1: it will present the main concepts of AI/ML and applications in Biology and Chemistry
  + Introduction to AI/ML
    - Overview and key concepts in AI/ML
    - ML workflow
    - Learning about the Python ML ecosystem (libraries and packages for ML)
  + AI/ML Applications in Biology and Chemistry
    - AI/ML methods
    - Data types
    - Multi-omics data analysis
    - Protein structure prediction with AlphaFold
* Live Demo
  + Introduction to popular Python libraries/packages for ML in Biology and Chemistry:
    - NumPy: library to support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays
    - Pandas: library for data analysis
* Hands-on Exercise
  + Basics of NumPy, Pandas

Day 2 (June 14, 1:00pm – 5:00pm):

* Lecture 2: it will cover data collection and data preparation, data readiness
  + Data Collection
    - Data formats (e.g., CSV/TSV, XML, JSON, HTML/Web Scraping, SQL)
  + Data Preparation for ML
    - Data cleaning (identifying and correcting errors in the dataset)
    - Feature selection (finding the best set of features that enables to build useful models)
    - Data transformation (converting data from one format into an ML-digestible format)
    - Feature engineering (selecting the required independent features)
    - Dimensionality reduction (reducing the number of input variables in the training data)
  + Data Readiness for AI/ML Checklist
    - Essential checks, Additional checks, Data preparation
* Live Demo
  + Demonstration of data preparation aspects, such as
    - Basic data cleaning
    - Marking and removal of missing data
    - Outlier identification and removal
    - Missing data imputation
* Hands-on Exercise
  + Data preparation related activities

Day 3 (June 15, 1:00pm – 5:00pm):

* Lecture 3: it will focus on the ML feature selection, engineering and scaling
* Live Demo
  + Scaling, Engineering on numeric/categorical data
  + Change data distribution
  + Derive new input variables
  + Select numerical, categorical input features
  + Recursive feature elimination
* Hands-on Exercise
  + Feature selection and engineering with Diabetes and Breast Cancer data

Day 4 (June 16, 1:00pm – 5:00pm):

* Lecture 4: it will focus on the ML model, from its selection to its training, evaluation and tuning
  + Model Selection, Training and Evaluation
    - Models: Linear regression, Polynomial regression, Logistic regression, Support vector machine, Decision trees, Random forest, Hierarchical clustering, Density based clustering
    - Metrics: Confusion matrix, Accuracy, Precision, Recall, F1-score, ROC, AUC, R2, MSE
  + Model Tuning, Interpretation and Deployment
* Live Demo
  + Model building and evaluation (Wisconsin Breast Cancer Dataset)
  + Model tuning, interpretation and deployment (Wisconsin Breast Cancer Dataset)
* Hands-on Exercise
  + Wine types prediction using physicochemical features
  + Wine quality prediction using physicochemical features

Day 5 (June 17, 1:00pm – 5:00pm):

* Lecture 5: it will provide an introduction to deep learning models
  + Introduction to Deep Learning
* Live Demo
  + Pfam protein sequence classification using Tensorflow and Keras
  + Predicting molecule solubility using DeepChem
* Hands-on Exercise
  + Wine types prediction using Deep Learning
  + Protein 3D structure prediction using AlphaFold2
* Course Summary