**Capstone 1**

**Predicting Molecular Properties**

Jingling Zhang

# Introduction

Nuclear Magnetic Resonance (NMR) spectroscopy is an analytical chemistry technique used in quality control and research for determining the content and [purity](http://chem.ch.huji.ac.il/nmr/qc.htm#impurities) of a sample as well as its [molecular structure](http://chem.ch.huji.ac.il/nmr/identify.htm). It shed light upon determination of molecular conformation in solution as well as study of physical properties at the molecular level such as [conformational exchange](http://chem.ch.huji.ac.il/nmr/techniques/other/dynamic/dynamic.html), phase changes, solubility, and [diffusion](http://chem.ch.huji.ac.il/nmr/techniques/other/diff/diff.html).

Using NMR to gain insight into a molecule’s structure and dynamics depends on the ability to accurately predict ***J*-couplings** or **Scalar couplings**, which contains information about relative bond distances, angles and connectivity of chemical bonds. *J*-couplings reflect the magnetic interactions between a pair of atoms. The strength of this magnetic interaction depends on intervening electrons and chemical bonds that make up a molecule’s three-dimensional structure.

It is possible to accurately calculate *J*-coupling constants given only a 3D molecular structure as input. However, these quantum mechanics calculations are extremely expensive (days or weeks per molecule), and therefore have limited applicability in day-to-day workflows.

A fast and reliable method to predict these interactions will allow medicinal and analytical chemists to gain structural insights faster and cheaper, enabling scientists to understand how the 3D chemical structure of a molecule affects its properties and behavior.

Ultimately, such tools will enable researchers to make progress in a range of important problems, like designing molecules to carry out specific cellular tasks, or designing better drug molecules to fight disease.

The focus of this Capstone Project would be to develop an algorithm that predict the scalar coupling constant.

# Client

This project would help analytical chemists, pharmacologists, and physicists to better understand the fundamentals of chemical structures and relevant molecular properties. It could also give insights into streamlining chemical synthesis and characterization in industry.

# Data

[https://www.kaggle.com/c/champs-scalar-coupling](https://www.kaggle.com/c/champs-scalar-coupling/overview) provided by Chemistry and Mathematics in Phase Space (CHAMPS) at the University of Bristol, Cardiff University, Imperial College and the University of Leeds.

train.csv: for training and testing models; the first column (molecule\_name) is the name of the molecule where the coupling constant originates, the second (atom\_index\_0) and third column (atom\_index\_1) is the atom indices of the atom-pair creating the coupling and the fourth column (scalar\_coupling\_constant) is the scalar coupling constant that we want to be able to predict.

structures.csv: used to extract information for training model; consist of molecule\_names, atom indexes, atoms, X, Y and Z cartesian coordinates.

# Approach

Python would be the main language and libraries such as pandas, numpy, seaborn, matplotlib etc. would be used.

Data sets do not have null values. However, some atom pairs contain Fluorine (F), which was not required for predicting the scalar coupling constant. Therefore, rows with information of Fluorine- contained atom pairs could be dropped. Structures.csv contains rows for molecules that are not present in train.csv, so inner join using pd.merge() would be necessary. The distance between atoms in a pair needs to be calculated using cartesian coordinates; the column of type of coupling (e.g. 1JHC) should be broken down into different columns containing one single piece of attribute (e.g. 1 for bond\_length, H for atom\_x, C for atom\_y). train.csv should be split into 2 parts, one for training models and another for testing out accuracy models.

Machine learning would be performed on dataframes.

# Deliverables

Major deliverables would include a prediction model, python code and PowerPoint slides.