Capstone 1 Predicting Molecular Properties

# Data Wrangling

Python was used as the major coding language, with libraries such as pandas, numpy, seaborn and matplotlib.pyplot.

* What kind of cleaning steps did you perform?

Two .csv files (train.csv and structures.csv) were imported as dataframes using pandas, namely *train* and *structures* respectively.

Explanatory data analysis was performed using seaborn figures, df.describe(), df.duplicated().any() and df.info() etc. These two datasets were clean as no null values or no duplicate rows were contained, the column names are consistent, and column data types are reasonable.

Dataframe train was rearranged for future analysis: the “type” column was broken down into three columns “bond”, “atom\_x” and “atom\_y”.

Two dataframes were merged and redundant columns were dropped. A ready dataframe *final* was used for further analysis.

Distance between 2 atoms were calculated using cartesian coordinates and assigned to a new column “distance”. A boxplot based on distance of different coupling types were generated.

* How did you deal with missing values, if any?

There were no null values. However, dataframe *structures* contained more rows than does dataframe *train*, because in the original Kaggle competition *structures* also included information for molecules in test.csv (used for testing machine learning model and project submission). Since in this project test.csv would not be used, *train* and *structures* were merged by left join and only molecules that appear in *train* would remain for the rest of project.

* Were there outliers, and how did you handle them?

From the boxplot, we can see lots of outliers, which is possible as scalar coupling constant varies dramatically for different molecules. Outliers in the case should not be removed. Overall data appeared in the right range, for example, all one-bond couplings had positive scalar coupling constant, and most of 3JHH couplings had positive constant, with a small portion having arbitrarily small negative values.