Project proposal

1) Team:

1. Sami Emre Erdogan, student ID 2019280513
2. Fedor Ivachev, student ID 2019280373
3. Andrei Glinskii, student ID 2019280807

2) Project: Predicting Molecular Properties. Can you measure the magnetic interactions between a pair of atoms?

Link: https://www.kaggle.com/c/champs-scalar-coupling

Problems:

In this competition the task is to develop an algorithm that can predict the magnetic interaction between two atoms in a molecule (i.e., the scalar coupling constant).

About Scalar Coupling

Using NMR to gain insight into a molecule's structure and dynamics depends on the ability to accurately predict so-called "scalar couplings". These are effectively 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.

Using state-of-the-art methods from quantum mechanics, it is possible to accurately calculate scalar 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.

Motivations:

A fast and reliable method to predict these interactions will allow medicinal 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.

Techniques:

We will use tree-based methods, neural networks and some kinds of ensemble models.

Data:

Tabular data, (312 MB)

Data description from Kaggle:

" Data Description

In this competition, you will be predicting the `scalar_coupling_constant` between atom pairs in molecules, given the two atom types (e.g., C and H), the coupling type (e.g., `2JHC`), and any features you are able to create from the molecule structure (`xyz`) files.

For this competition, you will not be predicting *all* the atom pairs in each molecule rather, you will only need to predict the pairs that are explicitly listed in the train and test files. For example, some molecules contain Fluorine (F), but you will not be predicting the scalar coupling constant for any pair that includes F.

The training and test splits are by *molecule*, so that no molecule in the training data is found in the test data.

Files

- train.csv - the training set, where the first column (`molecule_name`) is the name of the molecule where the coupling constant originates (the corresponding XYZ file is located at ./structures/.xyz), 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
- test.csv - the test set; same info as train, without the target variable
- sample_submission.csv - a sample submission file in the correct format
- structures.zip - folder containing molecular structure (xyz) files, where the first line is the number of atoms in the molecule, followed by a blank line, and then a line for every atom, where the first column contains the atomic element (H for hydrogen, C for carbon etc.) and the remaining columns contain the X, Y and Z cartesian coordinates (a standard format for chemists and molecular visualization programs)
- structures.csv - this file contains the same information as the individual xyz structure files, but in a single file”