# data\_science\_project

- This is the project of the thesis 'Exploring the Structure of SARS-CoV-2 Main Protease By Analyzing Molecular Dynamics Simulation Data', and the structure of the project is listed below:
- Figure\_1: The structure of all the programming files, and they can be found in the directory 'programming'. In the figure, the file with a red name is valid, the file with a blue name is used for adjustment and testing in the process, and the file with a green name is related to the display of the project.

#### **Correlation Analysis**

pearson\_correlation\_analysis.ipynb expand\_correlation\_analysis.ipynb residues\_analysis.ipynb

#### Reflection

correlation\_extraction.ipynb correlation\_threshold.ipynb Python scripts





#### Subtraction

difference\_matrix.ipynb difference\_subtract.ipynb subtract\_correlation.ipynb subtract\_example.ipynb

#### Machine Learning

logistic\_regression.ipynb S01\_scatter.ipynb subtraction\_pca.ipynb test\_ml.ipynb

#### Display

display\_generation.ipynb display\_subtraction.ipynb residues\_subtraction.ipynb

• Figure 2: The structure of results, all the files can be found in the directory 'results'.

Summary of Results



#### Reflection

- csv files containing thickness data
- session fies
- images of alpha carbons and sticks as correlations



### Correlation Analysis

- csv files of strong correlation C
  data
- images of correlation matrix heat maps

## Machine Learning

- images of points and regression lines



#### Subtraction

- csv files of differences between correlations in different systems
- images of subtraction matrix heat maps



#### Tableau

- images for visualization of correlation, reflection, and subtraction's results

