**Supporting Information**

Quantifying the Role of Qn Structural Units in C-A-S-H Gels: Insights from Molecular Dynamics Simulations and Interpretable Machine Learning

Yiting Zhanga, Shan Lia\*, Tianlei Zhanga, Huali Haoa\*

a School of Civil Engineering, Wuhan University, Wuhan, Hubei, China, 430072

Corresponding Author

Shan Li: lishan@whu.edu.cn; Huali Hao: haohuali@whu.edu.cn

**1 Models Constructed in MS**

C-A-S-H coordinate files from MS were used in this study, provided in two formats: CIF and XYZ, with 117 models in each format. The CIF files, stored in **/C-A-S-H Coordinate File/CIF files**, contain detailed crystallographic information, while the XYZ files, stored in **/C-A-S-H Coordinate File/XYZ files**, include atomic coordinates for molecular dynamics simulations and structural analysis.

**2 Nanoindentation Simulations in MD**

In the MD simulation, nanoindentation tests were carried out to investigate the mechanical properties of C-A-S-H. The input and output files of the simulation process are given below respectively:

**2.1 Input Data**

2.1.1 Data files

The initial C-A-S-H models with different Qn distributions, numbered from 1 to 117, are available in the **/MD/Input/Data files** folder. These models were constructed using Materials Studio software and subsequently converted into data file format using the msi2lmp tool.

2.1.2 Scripts for forcefield and nanoindentation

All the relevant files are located in the **/MD/Input/scripts** folder, which contains both the force field and nanoindentation simulation scripts. The files are as follows:

1-ffield.reax.choCaAlSi: Defines the essential potential functions for reactive force field simulations.

2-lmp\_control: Contains the LAMMPS simulation control settings.

3-param.qeq: A LAMMPS run script used for model equilibration, incorporating the necessary simulation settings, initial conditions, and force field parameters.

5-in.ind: Contains the settings required to perform the nanoindentation simulation.

**2.2 output files**

The output data consists of 117 sets of trajectory and output files, each corresponding to a simulation group (1-117). All files are stored in the **/MD/Output** folder, with each group of files organized in its respective subfolder. The data for each group includes the following:

1-Equilibration Trajectory Files: These are named dump.\*.relax.lammpstrj and contain the trajectory data from the equilibration process.

2-Nanoindentation Simulation Trajectory Files: These are named dump.\*.ind.lammpstrj and include the trajectory data from the nanoindentation process.

3-Energy Output Files: These are named out.dat and contain energy output during the simulation.

4-Simulation Log Files: These are named slurm-\*.out and provide the logs of the simulation run, detailing each step of the process.

**3 Training models with ML**

The dataset,four machine learning algorithms, and the script SHAP analysis in the study are included in **/ML** folder.

1-p2data250312.csv: The dataset p2data250312.csv consists of 117 data sets used for machine learning training, derived from the results of MD nanoindentation simulations.

2-MLR.py: This Python script implements the training process for the Multiple Linear Regression (MLR) model.

3-SVR.py: This Python script implements the training process for the Support Vector Regression (SVR) model.

4-RF.py: This Python script implements the training process for the Random Forest (RF) model.

5-BPNN.py: This Python script implements the training process for the Backpropagation Neural Network (BPNN) model.

6-final\_bpnn\_model.h5: This file contains the best-performing Backpropagation Neural Network (BPNN) model, saved in the HDF5 format.

7-SHAP.py: This Python script is used to perform SHAP (SHapley Additive exPlanations) analysis on the best-performing BPNN model, final\_bpnn\_model.h5.

**4 Analysis of bond lengths and bond angles**

This section provides the files related to the statistical analysis of bond lengths and bond angles in silicon-oxygen and aluminum-oxygen tetrahedra in C-A-S-H. The specific files are stored in the **/Bond length and angle analysis** folder and are as follows:

1-SiO4\_analysis.py: This script is used to identify the Q2 and Q3 sites of silicon-oxygen tetrahedra and calculate the bond lengths and bond angles.

2-AlO4\_analysis.py: This script is used to analyze the bond lengths and bond angles of aluminum-oxygen tetrahedra.

3-Q2\_bond\_kde.png: This plot shows the bond length distribution of the silicon-oxygen tetrahedra at the Q2 site, visualized using a KDE curve.

4-Q2\_angle\_kde.png: This plot shows the bond angle distribution of the silicon-oxygen tetrahedra at the Q2 site, visualized using a KDE curve.

5-Q3\_bond\_kde.png: This plot shows the bond length distribution of the silicon-oxygen tetrahedra at the Q3 site, visualized using a KDE curve.

6-Q3\_angle\_kde.png: This plot shows the bond angle distribution of the silicon-oxygen tetrahedra at the Q3 site, visualized using a KDE curve.

7-Q2\_bond\_kde\_Al.png: This plot shows the bond length distribution of the aluminum-oxygen tetrahedra at the Q2 site, visualized using a KDE curve.

8-Q2\_angle\_kde\_Al.png: This plot shows the bond angle distribution of the aluminum-oxygen tetrahedra at the Q2 site, visualized using a KDE curve.

9-Q3\_bond\_kde\_Al.png: This plot shows the bond length distribution of the aluminum-oxygen tetrahedra at the Q3 site, visualized using a KDE curve.

10-Q3\_angle\_kde\_Al.png: This plot shows the bond angle distribution of the aluminum-oxygen tetrahedra at the Q3 site, visualized using a KDE curve.

11-Results.csv: This file contains the statistical data for the average bond lengths and bond angles for Si-O and Al-O, along with the average absolute changes in bond length and bond angle at 30ps and 60ps.