# Multivariate Regression Analysis of Factors Regulating the Formation of Synthetic Aluminosilicate Nanoparticles

Faisal Adams[1], McNeill Bauer[1], Clement Levard[2], F Marc Michel[1]

[1]Department of Geosciences, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA
[2]Aix Marseille Univ, CNRS, IRD, INRA, Coll France, CEREGE, Aix-en-Provence, France
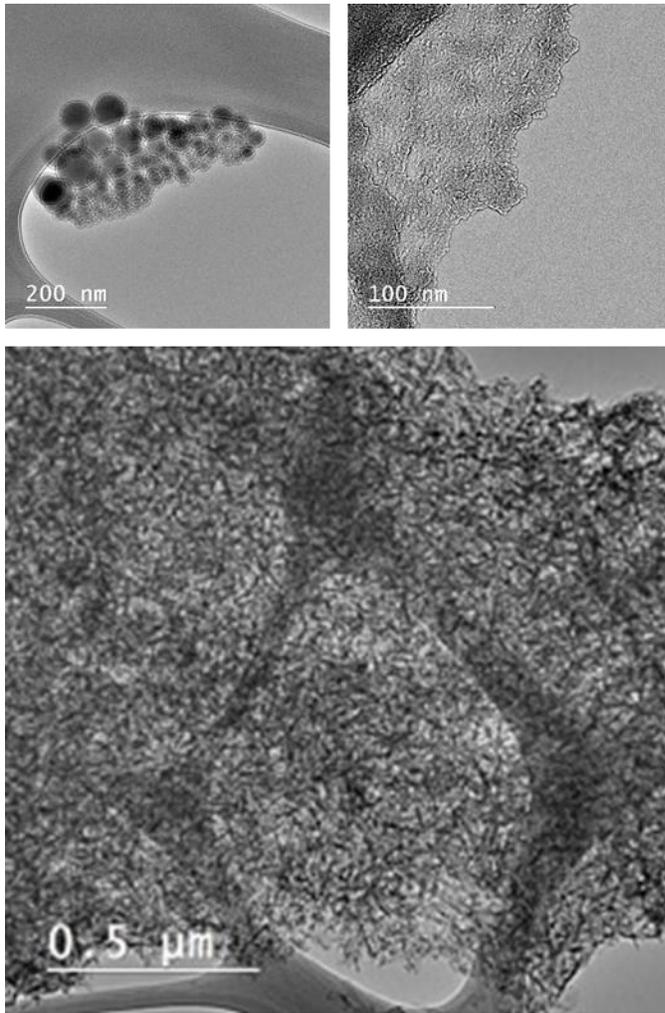
**Section I**



**Figure S1.** HR-TEM images of amorphous silica (top left), proto-imogolite (top right) and pseudo-boehmite (bottom). Imogolite is not shown here, since it is highly susceptible to the intense beams of a TEM and is commonly imaged after being embedded in epoxy or via cryo-TEM.
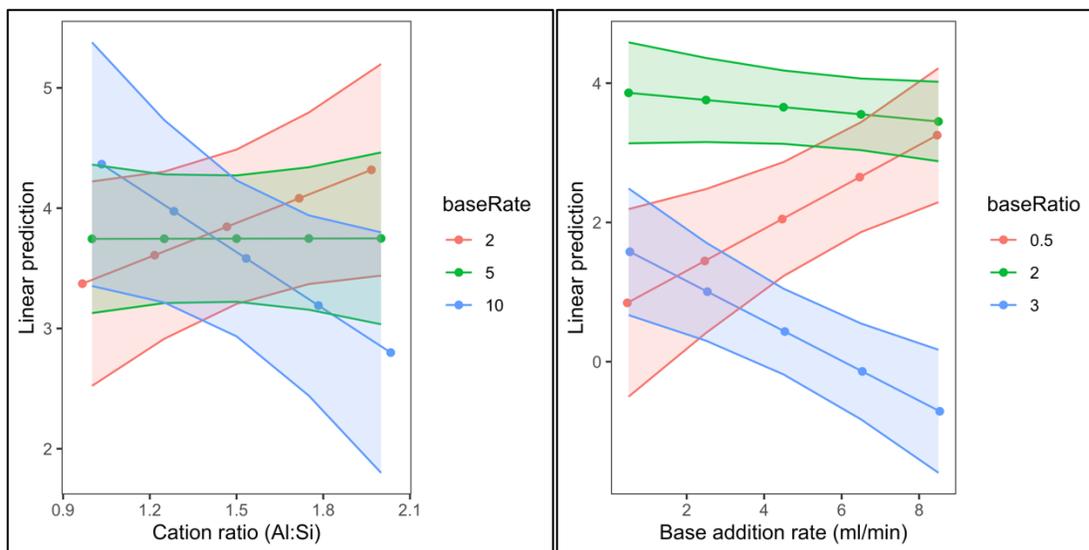
**Section II**



**Figure S2.** Plots of interaction between cation ratio and rate of base addition (left), and hydrolysis ratio and rate of base addition (right) for the imogolite model.
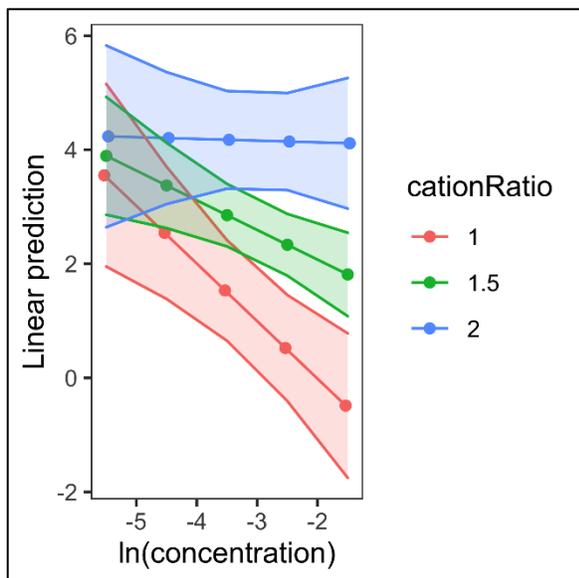


**Figure S3.** Plot of interaction between cation ratio and concentration for the pseudo-boehmite model.
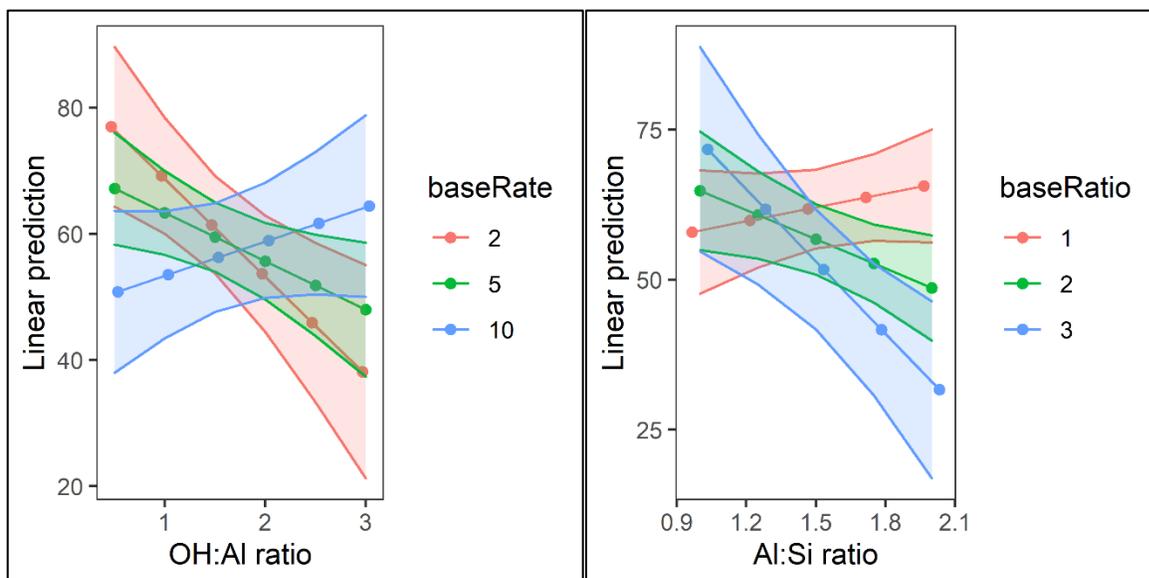
**Figure S4.** Plots of interaction between hydrolysis ratio and rate of base addition (left), and hydrolysis ratio and cation ratio (right) for the proto-imogolite model.
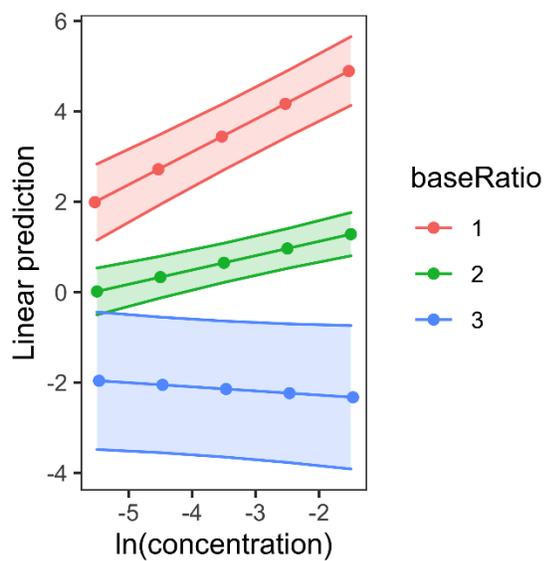


**Figure S5.** Plot of interaction between hydrolysis ratio and concentration for the amorphous silica model.

## Section III

**Table S1.** Linear regression model checks for each nanoparticle.

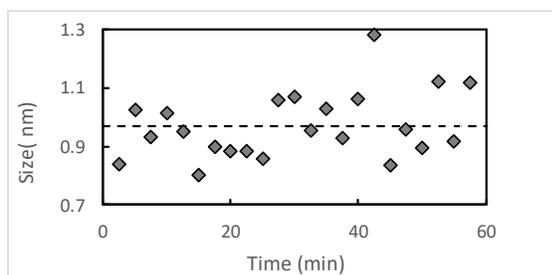| Model | adjusted $R^2$ | RMSE | corrected RMSE | bootstrapped DW statistic | DW p-value | Shapiro-Wilks test p-value | Q-Q plot |
|---|---|---|---|---|---|---|---|
| **Imogolite** | 0.85 | 0.65 | 0.27 | 2.95 | 0.13 | 0.88 | Passed |
| **Proto-imogolite** | 0.42 | 23.69 | - | 1.66 | 0.08 | 0.77 | Passed |
| **Pseudo-Boehmite** | 0.63 | 1.11 | 1.37 | 2.46 | 0.33 | 0.71 | Passed |
| **Amorphous Silica** | 0.67 | 0.73 | 0.39 | 1.89 | 0.48 | 0.96 | Passed |

## Section IV



**Figure S6.** The average particle size during synthesis at an Al:Si ratio of 2.0 and Al concentration of 0.1M. The dashed line represents the overall mean particle size of 0.97 nm.
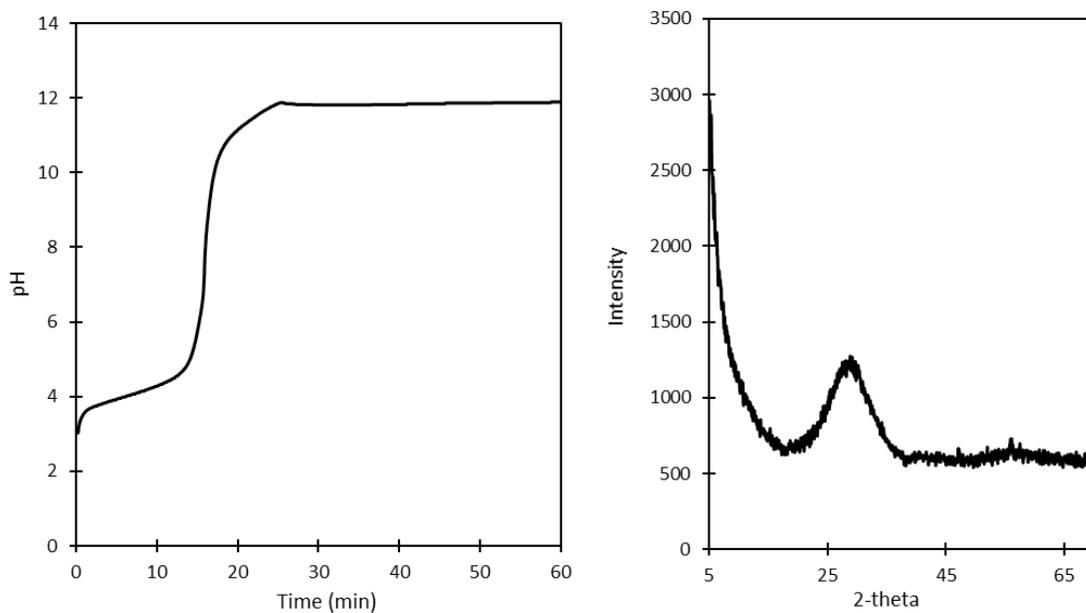
## Section V



**Figure S7.** The pH curve during synthesis (left) and pXRD pattern (right) for a sample prepared at a hydrolysis ratio of 5. The other synthesis conditions include an initial Al concentration of 0.1M, base addition rate of 10mL/min and Al:Si ratio of 1.
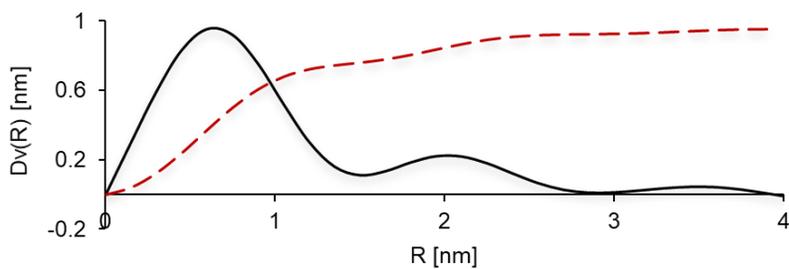
## Section VI



**Figure S8.** DvR results showing the cumulative distribution of particle sizes (dashed line) and particle size distribution (solid line).

**Section VII**



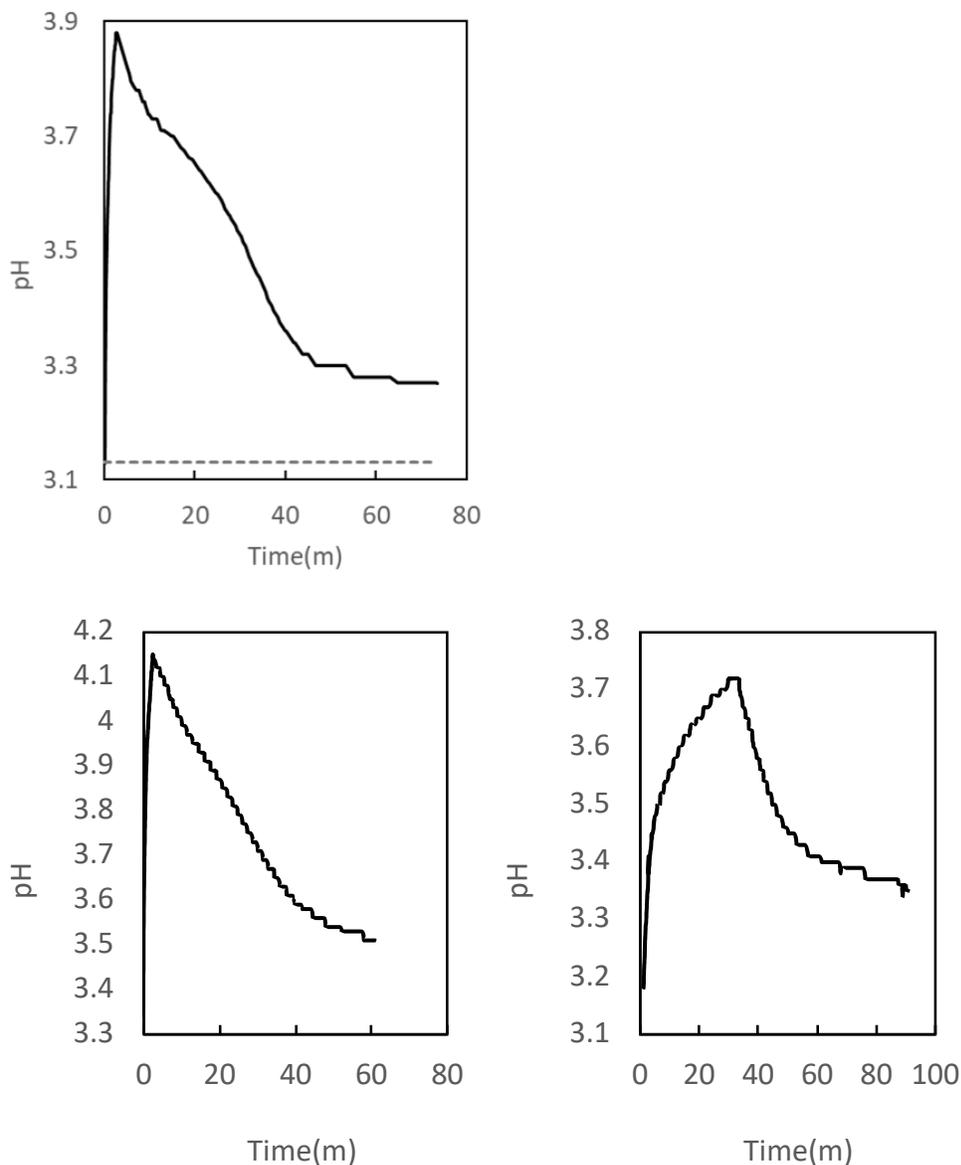**Figure S9.** Example plots of pH change observed during the first hour of synthesis. The top left plot had synthesis conditions of 0.1M AlCl3 with a 1:1 Al:Si ratio, and a 1:2 OH:Al ratio. Time zero corresponds with the starting pH and the beginning of NaOH addition. The bottom left plot has the identifier C0.05 F10 S1 H0.5, while the bottom left plot belongs to C0.125 F2 S1 H0.5. The key to these identifiers is detailed below.

**Table S2.** pH values for a set of conditions before and after oven treatment. The pre-oven pH is the value recorded at the end of the 1-hour period during synthesis, and the post-oven pH was recorded after the suspensions were removed and cooled to room temperature. A key for the label is provided beneath the table.

| Synthesis conditions | Pre-oven pH | Post-oven pH |
|---|---|---|
| C0.05 F10 S1 H0.5 | 3.51 | 2.56 |
| C0.1 F10 S2 H0.5 | 3.46 | 2.84 |
| C0.15 F10 S1.5 H3 | 6.20 | 4.21 |
| C0.1 F10 S1 H0.5 | 3.26 | 2.25 |
| C0.125 F2 S1 H1 | 3.35 | 2.65 |

C = Al concentration (M), F = flowrate (ml/min) S = Al:Si ratio (mol/mol) and H = OH:Al ratio (mol/mol). For example, C0.2_F2_S2_H3 corresponds to synthesis conditions of 0.2M Al concentration, 2 ml/min flowrate, Al:Si ratio of 2 mol/mol and OH:Al ratio of 3 mol/mol.

**Section VIII**

Statistical interpretation

The $R^2$ value quantifies the variance in the data that is explained by a regression model, and it was vital in analyzing the effectiveness of our multivariate methods. In a multivariate model, the presence of multiple independent variables in the regression increases the overall $R^2$ value, and so an adjusted $R^2$ calculation is required. The adjusted $R^2$ accounts for that increase and instead gives a true representation of how much of the data is explained by the model - it is reported in this paper for all our linear regression models.

The root mean square error (RMSE) is the standard deviation of the residual of the data and is used to provide an estimate of our models' predictive power.

Durbin and Watson (DW) statistics were used to formally test for autocorrelation in the residuals from the regression analysis [1, 2]. Random residuals are a crucial component of a predictive model, and correlation in the residuals indicates some predictive information is not described by the model. The DW statistic (d) always lies between 0 and 4. The alternative hypothesis that the residuals are autocorrelated is rejected when the p-value for our bootstrapped DW test (10,000 reps) is $> 0.05$

The independence of all main explanatory variables were tested using, a correlation matrix.

Normality of the residuals was formally tested using the Shapiro-Wilk test [3] due to its robustness compared to other tests of normality [4], and graphically confirmed with a Q-Q (quantile-quantile) plot.

Selection of input conditions

The minimum number of samples needed was determined by performing a power analysis in R, and subsequently using references from the literature to determine the range of conditions commonly covered and investigated. Using 4 independent variables (equivalent to 3 numerator degrees of freedom), 66 data points (equivalent to the denominator degrees of freedom), $f^2$ of 0.35 and a significance level of 0.05 as parameters resulted in a power of 0.989. These values mean that with a minimum of 66 data points, significant effects (f2) of the independent variables can be detected with 98.9% chance of detecting an effect (power) and at least 95% confidence in the models' ability to capture the variation in the data. The range of conditions that would maximize the formation of several phases were also inferred, with the aim of modeling a continuum of their formation.
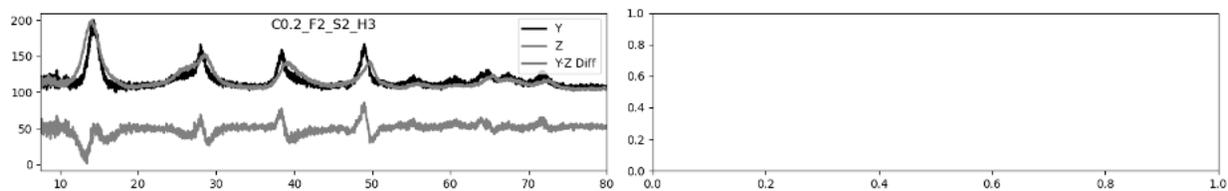
**Section IX**
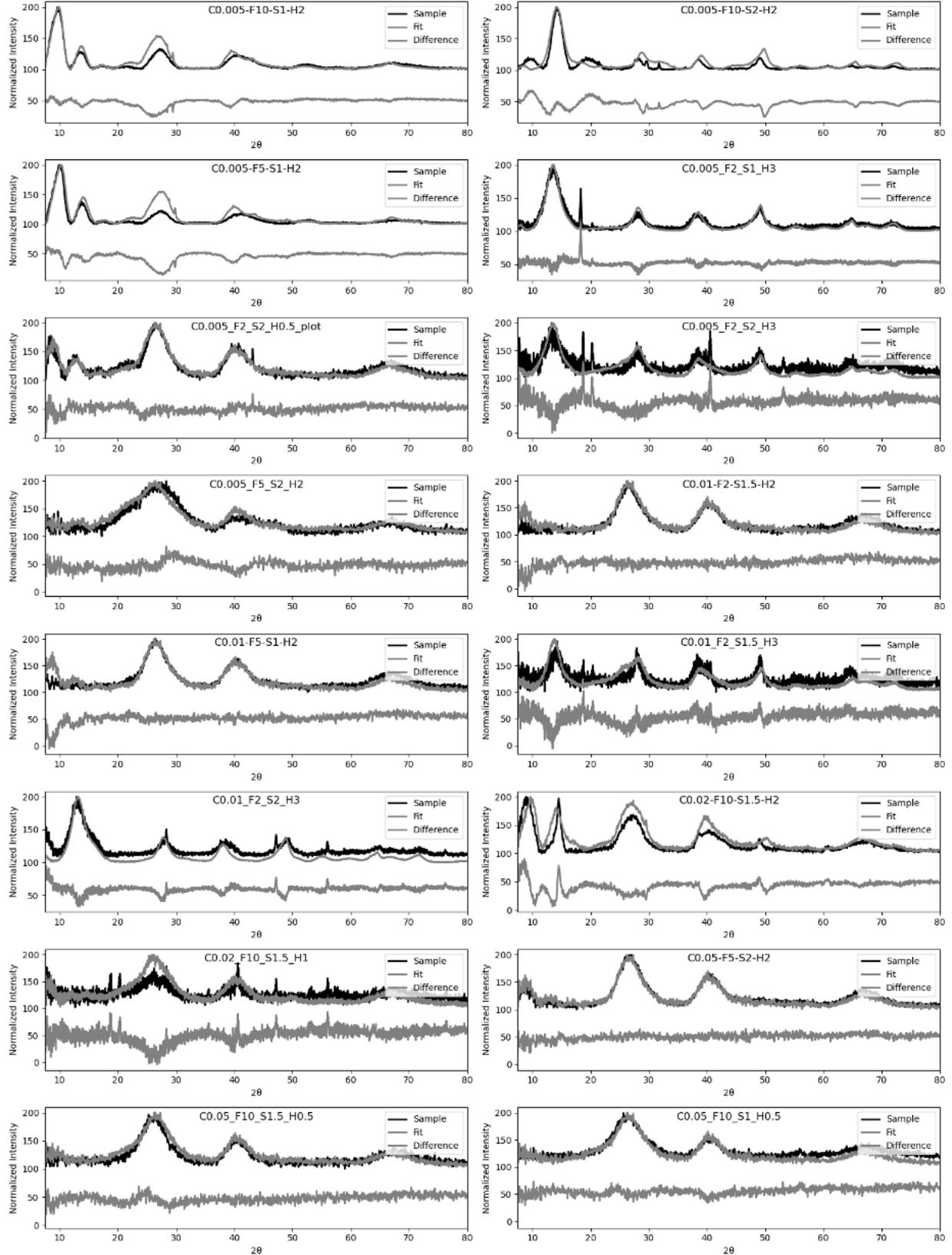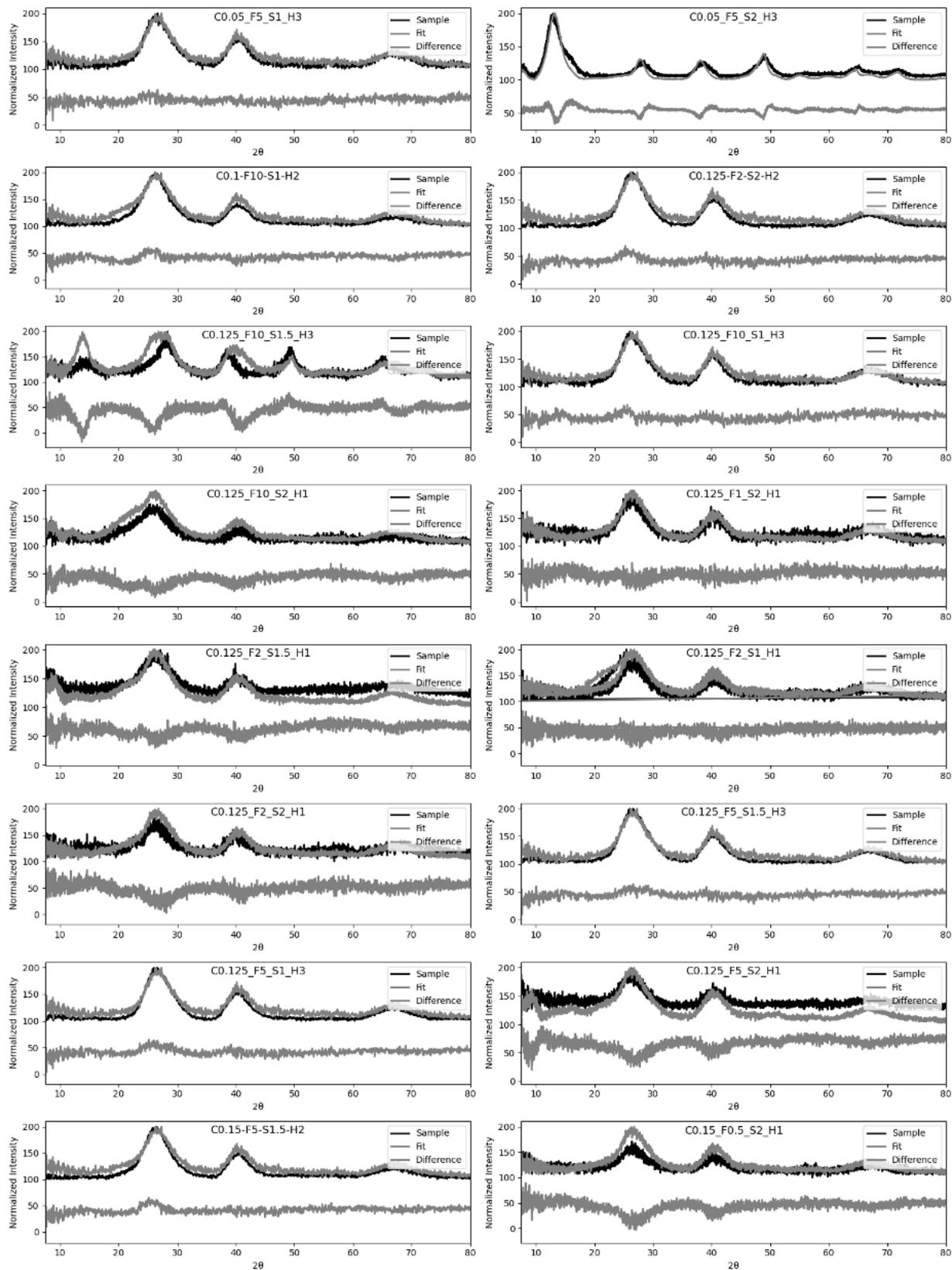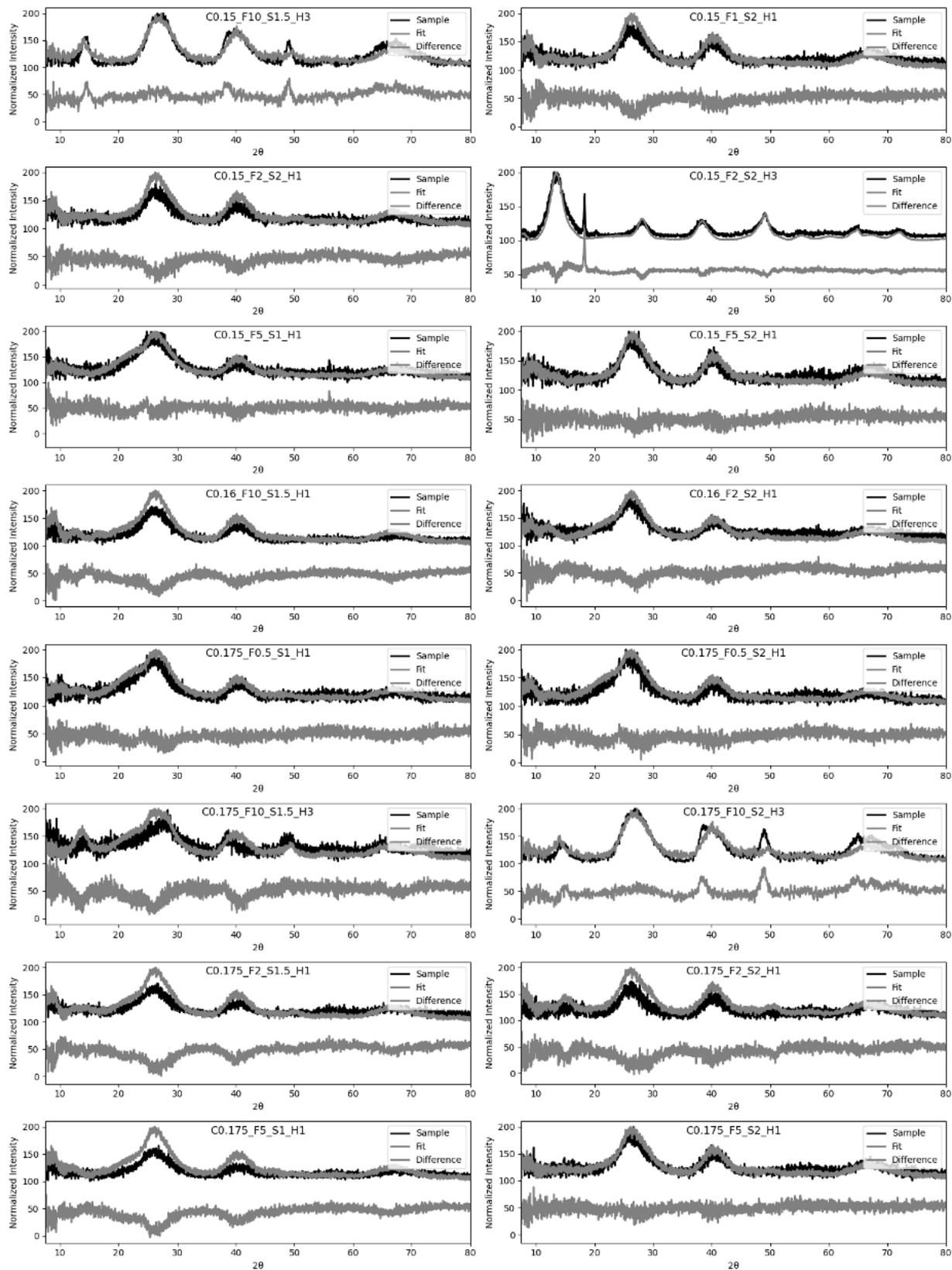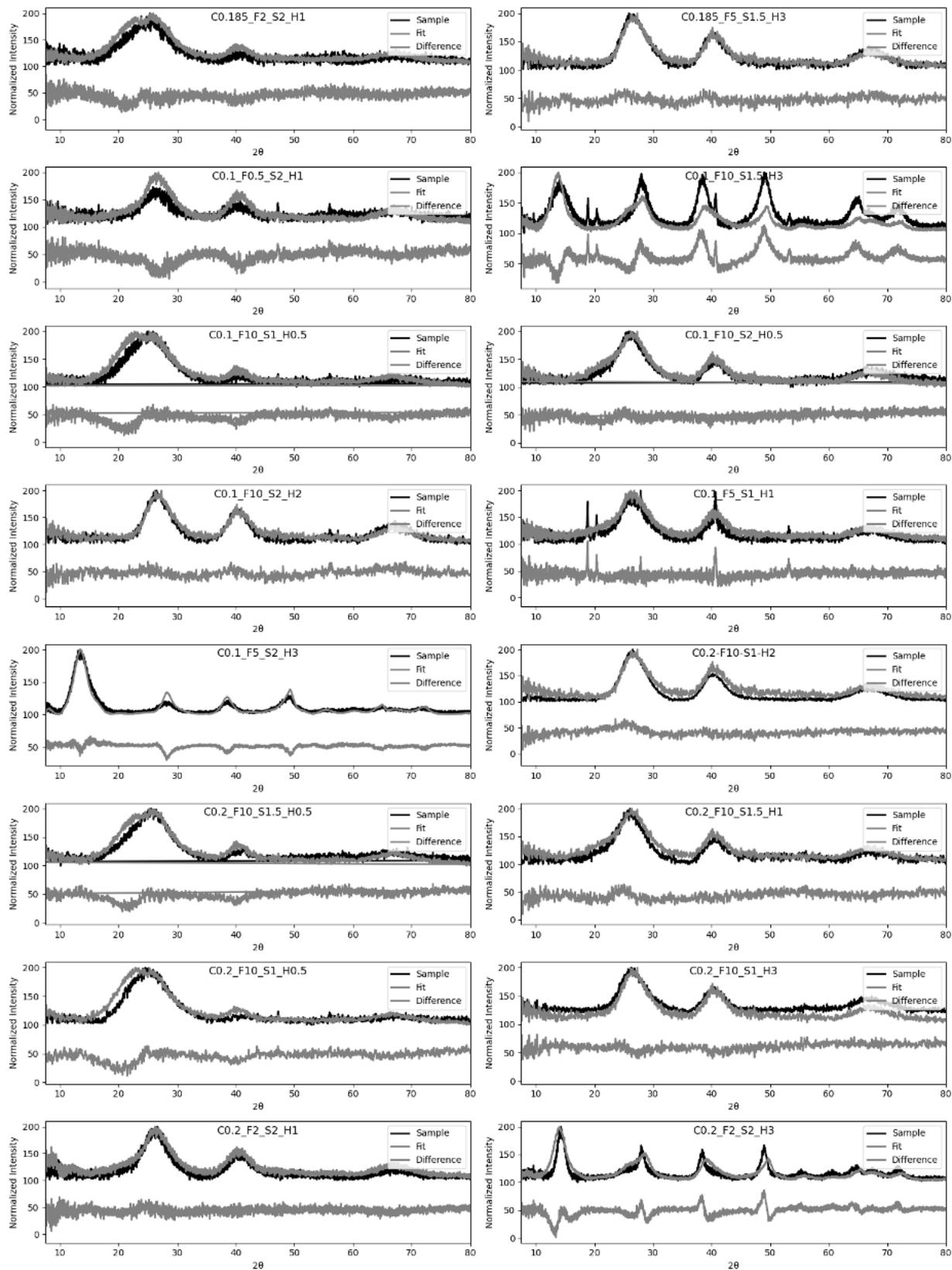
**Figure S10.** The plots above show the powder diffraction data for the samples used in this study. The solid black line corresponds to the sample pattern, the overlaid grey line shows the fit while the grey line beneath both of these shows the difference curve. The vertical axes represent normalized intensities, while the horizontal axes are recorded in $2\theta$. Each pattern is labeled with a unique identifier that corresponds to the synthesis conditions used to produce it - C = Al concentration (M), F = flowrate (ml/min) S = Al:Si ratio (mol/mol) and H = OH:Al ratio (mol/mol).

**Table S3.** Synthesis conditions reported in published studies and the corresponding phases derived. Since single phases were reported, all particles are assumed to be 100% in relative abundances for each entry.

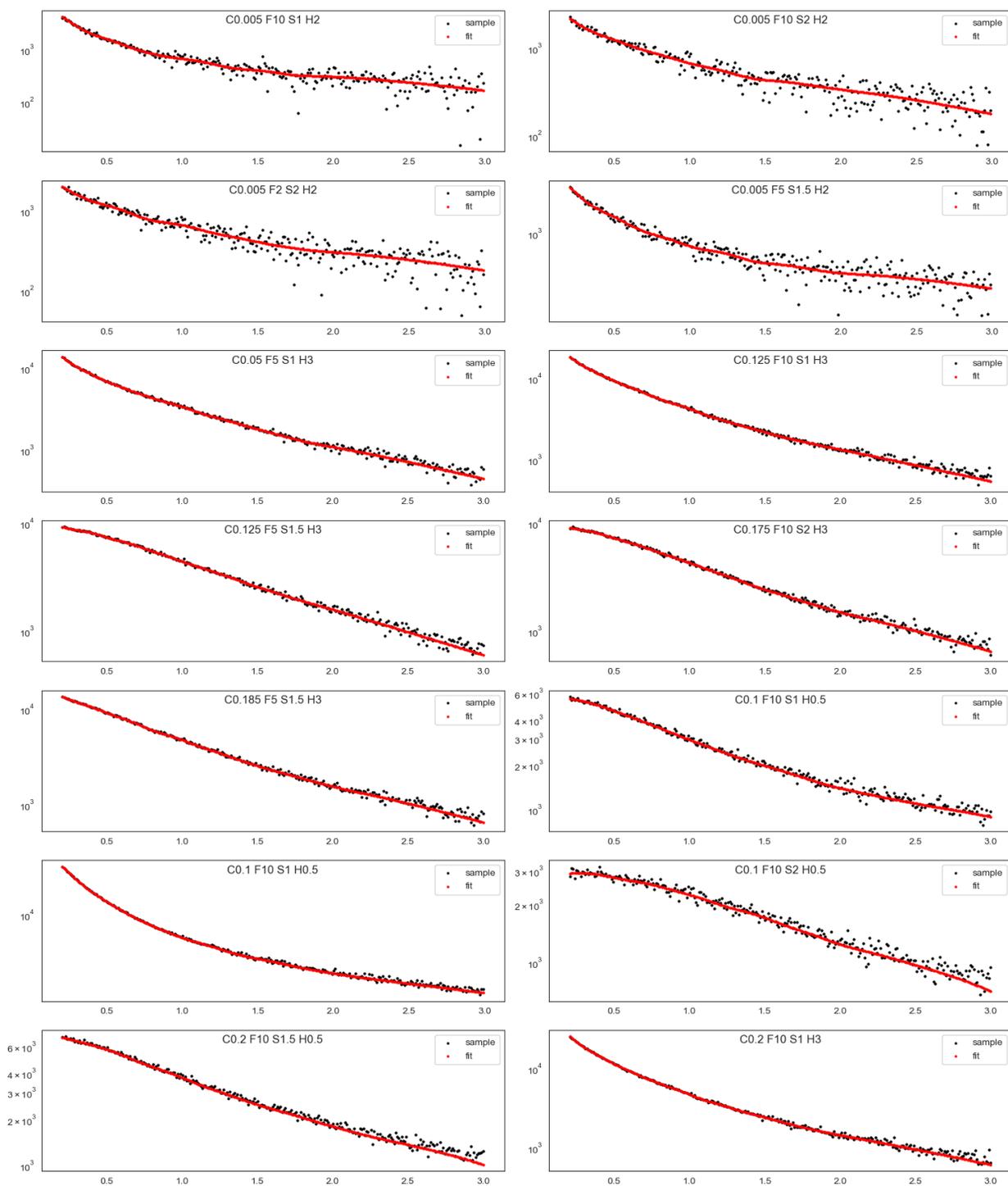| Phase | Paper | Al_conc | Al_Si | hydrolysis |
|-------|-------|---------|-------|------------|
| Proto-imogolite | Levard et al 2012 | 0.1 | 2 | 2 |
| Imogolite | Pan et al 2017 | 0.005 | 2 | 2 |
| Imogolite | Denaix et al 1999 | 0.002 | 2 | 1 |
| Proto-imogolite | Denaix et al 1999 | 0.1 | 2 | 1 |
| Proto-imogolite | Ohashi et al 2002 | 0.1 | 1.33 | 3 |
| Proto-imogolite | Ohashi et al 2002 | 0.05 | 1.33 | 3 |
| Proto-imogolite | Ohashi et al 2002 | 0.03 | 1.33 | 3 |
| Proto-imogolite | Ohashi et al 2002 | 0.01 | 1.33 | 3 |

# Section X



**Figure S11.** *In situ* SAXS data shown in black markers along with their corresponding Dv(R) fits overlaid in red. The horizontal axis is q in $nm^{-1}$, and the vertical axis is intensity in arbitrary units. The plot titles correspond to the input conditions for each pattern, and follow the naming convention detailed in previous sections.

## References

1.	Durbin, J.; Watson, G. S., Testing for serial correlation in least squares regression: I. *Biometrika* **1950,** *37* (3-4), 409-428.
2.	Durbin, J.; Watson, G. S., Testing for serial correlation in least squares regression: II. *Biometrika* **1951,** *38* (1-2), 159-178.
3.	Shapiro, S. S.; Wilk, M. B., An analysis of variance test for normality (complete samples)†. *Biometrika* **1965,** *52* (3-4), 591-611.
4.	Razali, N. M.; Wah, Y. B., Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics* **2011,** *2* (1), 21-33.