Comparison of CATS with Principal Component Analysis (PCA)

Principal Component Analysis (PCA) was performed on the amorphous basic calcium carbonate example datasets demonstrated in section 3.1 of the manuscript. Python codes were written to perform the presented analysis, which utilized the Scipy PCA library in sklearn.decomposition. The number of components is set to 2, and defaults employed for all other parameters.

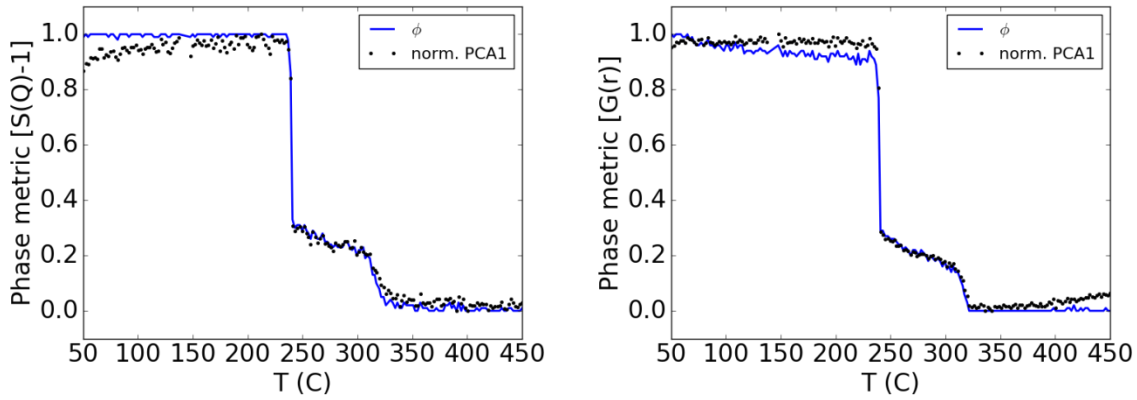

Figure 1 : The resultant phase-defining metric (ϕ for CATS and PCA1 for PCA) from analyzing the A(B)CC dataset with CATS and PCA in (left) reciprocal- and (right) real-space.

In order to directly compare the PCA to CATS, the phase metric employed in each case is the ϕ value for CATS (which varies between 0 and 1) and a normalized version of the PCA1 metric (such that it also varies between 0 and 1). When plotted in this way, as seen in Fig. 1, it is clear that both methods detect the phase transitions and relative phase fractions with comparable sensitivity. The PCA derived Cartesian coordinates (PCA1 vs. PCA2) are plotted in Fig. 2, which again demonstrate the clear transition at T = 230 C.

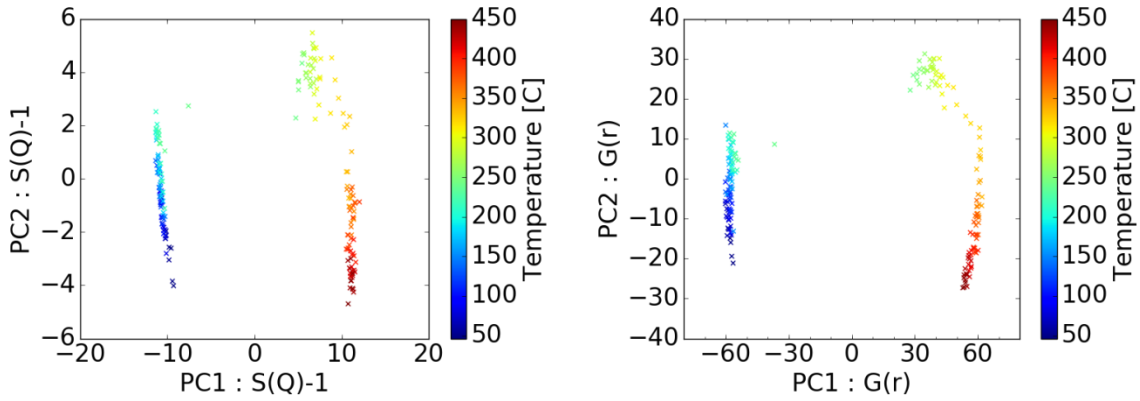

Figure 2 : The PCA derived Cartesian coordinates of the weighting between PCA1 and PCA2 for the A(B)CC dataset looking at both (left) reciprocal- and (right) real-space.

In the case of the PCA analysis, the derived components are calculated by the method itself, based on the principles of maximizing variance in weights between the components while

minimizing the residual and thus, producing the highest quality fit to the data. The CATS method, by contrast, begins by supplying the parent components (as ϕ = 0 and 1 models), and then fits the data in a best-as-it-can fashion. Weighting schemes in the CATS method are typically limited between 0 and 1, and inversely related between components. In PCA, weights can be any real value found by the algorithm (positive or negative). For these reasons, the derived individual components in PCA need not directly relate to any single phase, only that their constituent sum, under different weighting schemes, fit the data well. The resultant residual from the PCA and CATS analysis of the ACC data are shown in Fig. 3.

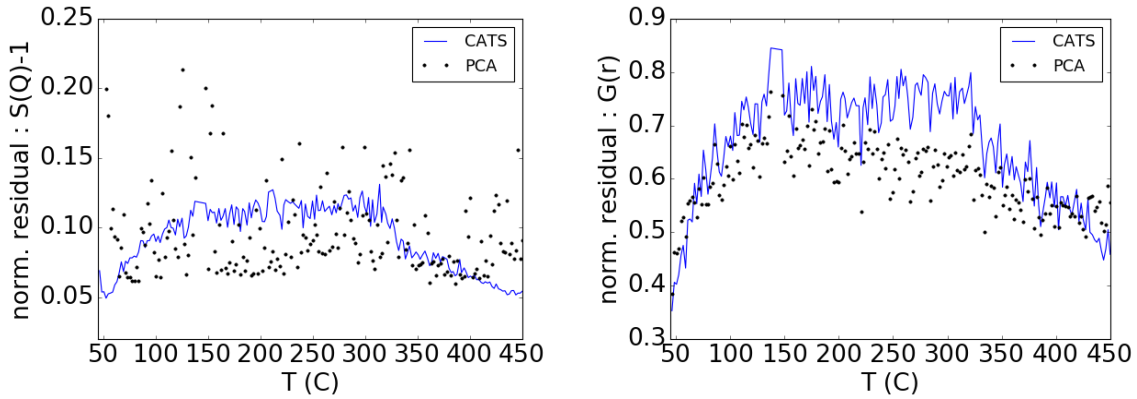

Figure 3 : The associated residuals, normalized by number of data points, from fitting the full set of A(B)CC data with the refined PCA and CATS solutions in both (left) reciprocal- and (right) real-space.

We see that both methods fit the data relatively well in both real- and reciprocal-space, with the PCA producing slightly lower residuals compared to the CATS analysis. This is not surprising, as PCA determines a best set of principal components (similar to parent datasets in CATS) to best fit the data as a whole. CATS, by contrast, is limited to physical models provided by the data. A comparison of the PCA derived components and parent datasets from CATS are shown in Fig. 4.

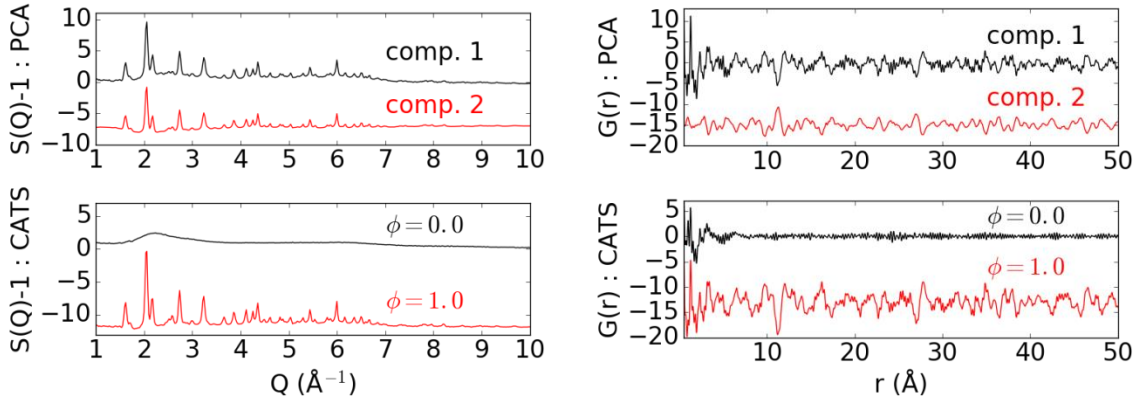

Figure 4 : The PCA derived components (comp. 1 and comp. 2) and employed CATS parents ($\phi = 0/1$) from the analysis of the A(B)CC dataset in both (left) reciprocal- and (right) real-space.

Here, we see that the components from PCA, while able to fit the data to high quality (under the calculated weighting scheme) are not directly related to real models or system under study. This demonstrates the strength of CATS analysis for its intended purpose; one is able to probe how a physically relevant selection of parents (measured amorphous and crystalline structures) contributes across a varying experimental coordinate (temperature). Those regions of the dataset which are NOT well fit by CATS can help suggest the existence of additional structures, which CATS cannot readily accommodate by deriving its own ideal basis set of eigenvectors (i.e. components or parents).

We performed the same analysis on the $BaTiO_3$ dataset of section 3.3 (PCA vs. CATS) and found similar results. Both CATS and PCA fail to detect the phase change via phase metric (see Fig. 5) however the residual in both cases suggests the underlying transitions (Fig 8). While the average residual is lower for the PCA (Fig. 8), the derived components from the PCA were not physically relevant. We note that the residual from the CATS analysis is structurally similar to the Cartesian pathway derived from PCA.

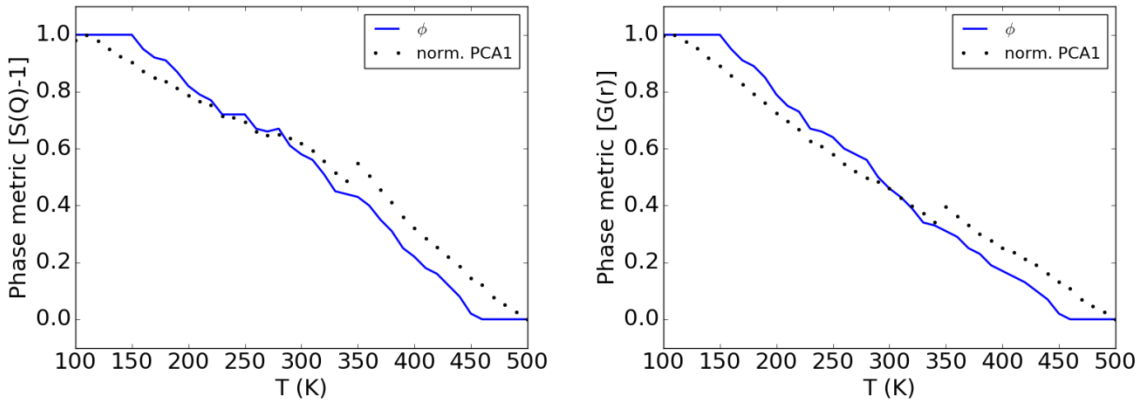

Figure 5 : The resultant phase-defining metric (ϕ for CATS and PCA1 for PCA) from analyzing the $BaTiO_3$ dataset with CATS and PCA in (left) reciprocal- and (right) real-space.

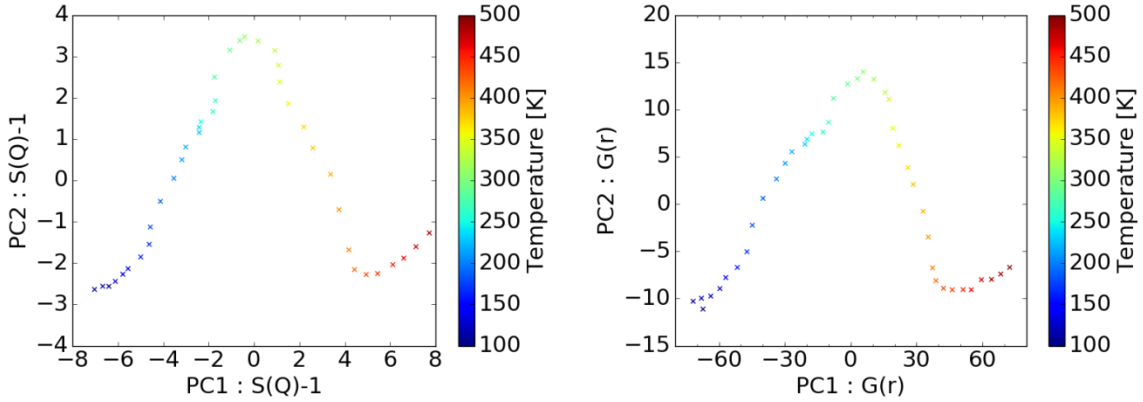

Figure 6 : The PCA derived Cartesian coordinates of the weighting between PCA1 and PCA2 for the $BaTiO_3$ dataset looking at both (left) reciprocal- and (right) real-space.

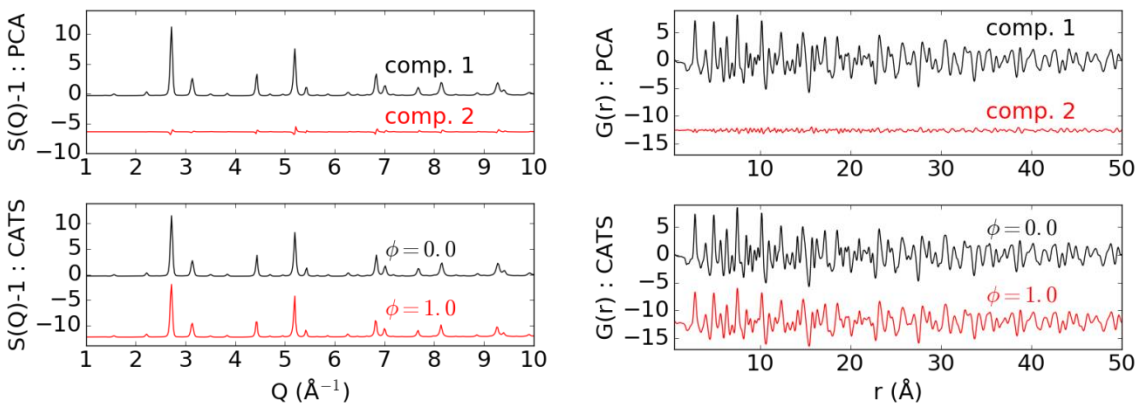

Figure 7 : The PCA derived components (comp. 1 and comp. 2) and employed CATS parents ($\phi = 0/1$) from the analysis of the $BaTiO_3$ dataset in both (left) reciprocal- and (right) real-space.

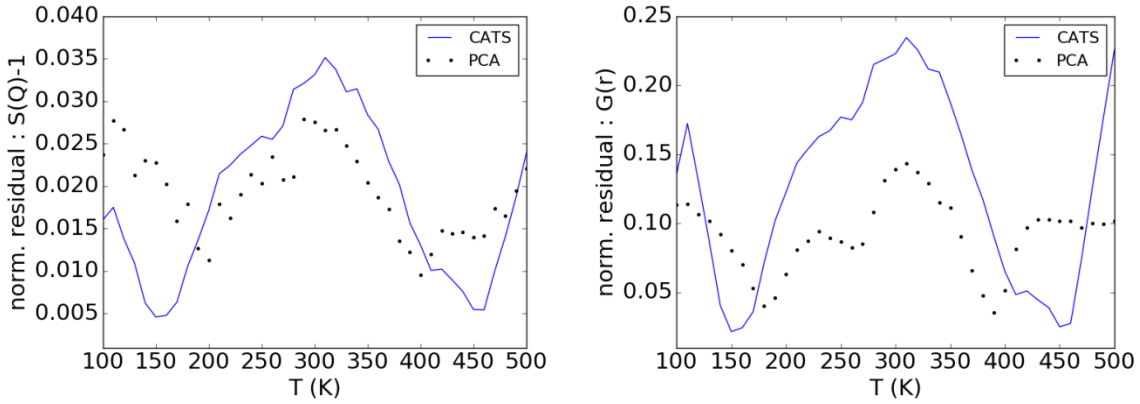

Figure 8 : The associated residuals, normalized by number of data points, from fitting the full set of $BaTiO_3$ data with the refined PCA and CATS solutions in both (left) reciprocal- and (right) real-space.

Example of CATS models compared with data

We here demonstrate the as-fit CATS models to data for the PDFs from the ACC data series. Note that these measured patterns were limited in statistics, and as can be seen from Fig 9a, this results in noisy PDFs. The CATS derived models are compared to the data here at temperatures before (200 C – 9b.), during (280 C – 9c.), and after (350 C – 9d.) the transition. Individual spectra at these temperatures are highlighted in red in Fig. 9a.

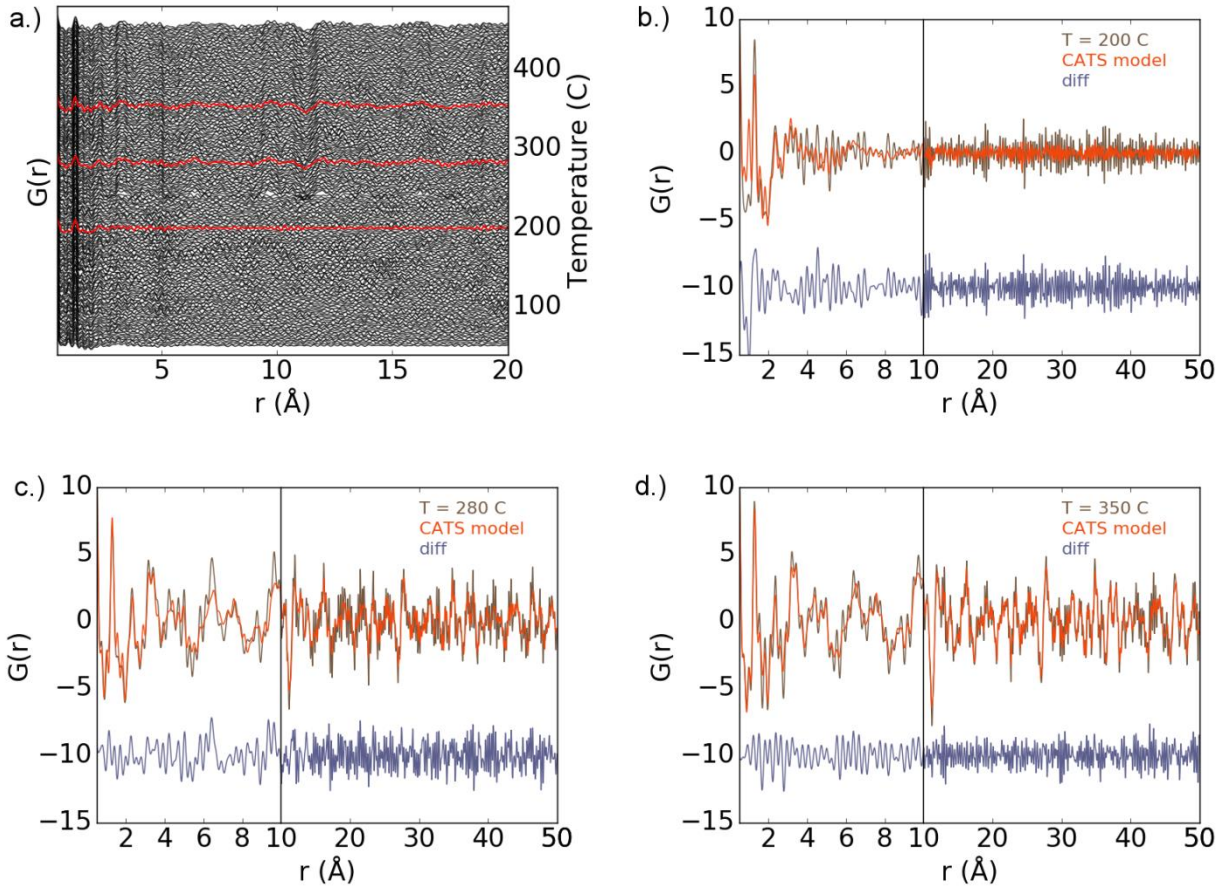

Figure 9 : (a) The full ACC dataset plotted as a waterfall plot, with the data from T = 200, 280, and 350 C highlighted in red. (b-d) The comparisons of the CATS derived models to the data at those temperatures, with the differences offset.

As can be seen, the CATS analysis does approximatively model the data, despite being a simple linear combination of the first and last phases of the material. Note that the modeling of data in this way is meant to indicate features of interest for further study, and not as final models. Here, that CATS analysis suggests that the intermediary transitionary region, from T = 240-325 C, contains coexistence of the initial amorphous phase and final crystalline phases.