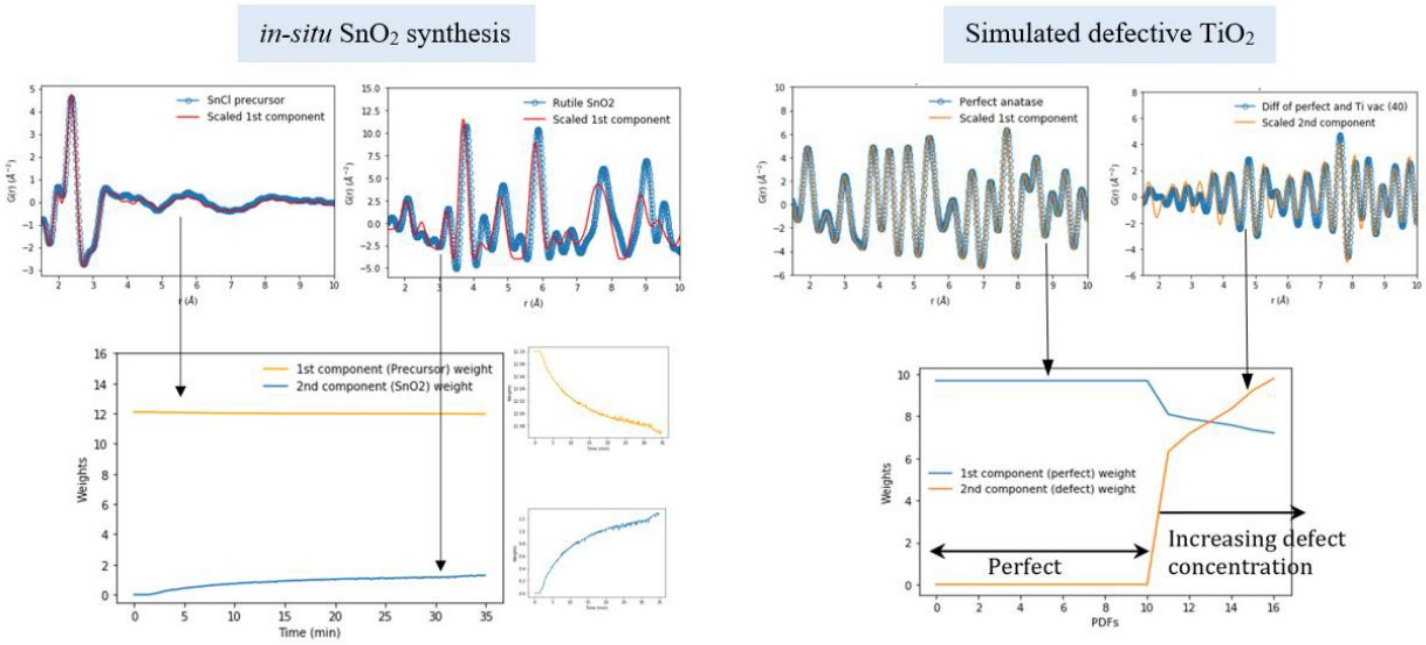
Data-driven approaches on pair distribution function data: matrix factorization and clustering

Shuyan Zhang, Jie Gong, B. Reeja Jayan, Alan J. H. McGaughey

Carnegie Mellon University, Pittsburgh, United States of America;

szhang2@andrew.cmu.edu

Advances in synchrotron X-ray scattering experiments have greatly increased the acquisition rates of pair distribution function (PDF) data. The analysis and interpretation of the data, however, are lagging behind the experimental advances because PDF analysis is met by the challenge of finding the correct structure model to fit against the data, which is a time-consuming process. We aim to apply data-driven methods to accelerate the analysis process of PDF data and the characterization of local material structures. Principal component analysis (PCA) and non-negative matrix factorization (NMF) are used to separate different features and/or constituents from the sample PDF data. We first applied these two methods on in-situ PDF measurement during tin oxide synthesis and then on the simulated PDFs of defected anatase titanium dioxide (TiO_2). It is found that for the in-situ PDF of tin oxide synthesis, NMF is able to separate constituents during different stages of the synthesis process and their relative concentrations are consistent with the experiments. For the PDF dataset of defected anatase (TiO_2), we found that NMF can separate the PDF signal of the defects from that of the perfect phase. This technique provides a tool to identify and quantify the defects from PDF data of materials.



Keywords: Defect, Pair distribution function, Data-driven, Matrix factorization