Supporting information

X-ray Nanospectroscopy Reveals Binary Defect Populations in Submicrometric ZnO Crystallites

Selwin Hageraats* [a, b, c], Katrien Keune[a, c], Slavica Stankic[d], Stefan Stanescu[e], Moniek Tromp[f], and Mathieu Thoury[b]

[a]Rijksmuseum, Conservation and Science, PO box 74888, 1070DN Amsterdam, The Netherlands

[b]IPANEMA, CNRS, Ministère de la Culture et de la Communication, Université de Versailles Saint-Quentin-en-Yvelines, USR 3461, Université Paris-Saclay, 91128 Gif-sur-Yvette, France

[c]Van 't Hoff Institute for Molecular Science, University of Amsterdam, PO box 94157, 1090 GD Amsterdam, The Netherlands

[d]Sorbonne Université, CNRS, Institut des NanoSciences de Paris, INSP, F-75005 Paris, France

[e]Synchrotron SOLEIL, L'ormes des Merisiers, Saint Aubin BP-48, 91192 Gif-Sur-Yvette Cedex, France

[f]Materials Chemistry, Zernike Institute for Advanced Materials, Nijenborgh 4, 9747AG Groningen, The Netherlands

*E-mail: s.hageraats@rijksmuseum.nl

2 pages; text only

Detailed information about processing and visualization of STXM energy stacks

The total dataset obtained from the STXM analysis of zinc whites comprises six energy stacks with between 153 and 205 Zn L-edge energies each, and six two-energy stacks recorded at the highest possible resolution (20-30 nm step size) with energies chosen before and after the Zn L-edge. In addition, the ultrapure lab-synthesized ZnO sample was analyzed in the same manner to act as a reference.

All STXM data was preprocessed using the aXis200 software in two steps. First, the images inside each energy stack were aligned using a Fourier cross-correlation algorithm to account for the inherent mechanical and thermal drifts during the extensive measurement time. Second, the transmission images were transformed into optical OD images by choosing an appropriate I_0 signal in the stack (i.e. the average of a region with no ZnO present).

The aligned OD high-spectral resolution energy stacks of the six zinc whites were then loaded into Python to undergo a series of five additional preprocessing steps:

1) Resampling to a common energy axis corresponding to the dataset with lowest number of spectral points (i.e. 153)
2) Selection of spectra based on a signal-to-noise threshold of 10
3) Gaussian filtering with a filter of width σ = 2.2 eV
4) Subtraction of the average absorption between 1015 and 1021 eV
5) Normalization by each spectrum's integral

The resulting set of resampled, selected, smoothed, and normalized spectra was factorized using the simplex volume maximization (SiVM) algorithm, for which the mathematical basis was set forth by Thurau *et al*.[1] The SiVM algorithm was set to calculate a simplex with only four vertices (endmembers) with the maximum possible volume. This maximization of the simplex volume can be shown to be equivalent to the minimization of the sum of squared errors between the actual data and the data modelled linearly using the simplex vertices (endmembers) as basis vectors.

To visualize what this linear model of the data looks like, the four endmembers and the smoothed data of the zinc whites *and* the lab-synthesized ultrapure ZnO were passed to an NNLS fitting algorithm (*scipy.optimize.nnls*). The algorithm produces a data cube of reduced dimensionality that contains the non-negative coefficients describing each data point in terms of a linear combination of the four spectral endmembers. Each slice of this cube is a coefficient matrix that can be interpreted as the distribution of the compound represented by that particular endmember throughout the sample. False-color images of all seven samples could then be produced by assigning three endmember coefficient matrices of choice to the red, green, and blue channel.

The average spectra shown in figure 4 of the main text were calculated by subtracting from the coefficient matrix of one endmember the coefficient matrices of both other endmembers—producing a difference matrix—and averaging the spectra corresponding to all difference matrix elements that exceed a threshold value. From the values in the difference matrix, the threshold value was calculated by to be the percentile for which the local percentile gradient exceeds the minimum percentile gradient by a factor of 50.

1) Thurau, C.; Kersting, C. K.; Wahabzada, M.; Bauckhage, C. Descriptive Matrix Factorization for Sustainability: Adopting the Principle of Opposites. *Data Min. Knowl. Discov.* **2011**, *24*, 325–354