Constraining data mining with physical models: a comparative study of voltage- and oxygen pressure-dependent transport in multiferroic nanostructures

Evgheni Strelcov[1], Alexei Belianinov[1], Ying-Hui Hsieh[2], Ying-Hao Chu[2,3], and Sergei V. Kalinin[1†]

[1]Institute for Functional Imaging of Materials and Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, United States

[2]Department of Materials Science and Engineering, National Chiao Tung University, Hsinchu 30010, Taiwan

[3]Institute of Physics, Academia Sinica, Taipei 105, Taiwan

As explained in the main text, determination of the number of significant behaviors in the dataset is possible via application of statistical (PCA), clustering methods (k-means), or by under and oversampling. The later concept is shown in Figure S1, where the ambient dataset with maximal V_p = 5 V is BLU-unmixed into 2 (undersampling), 4 (optimal *k*), and 5 (oversampling) components. Undersampling leads to incomplete separation of components: on one hand, the

† Corresponding authors: strelcove@ornl.gov

whole CFO island is ascribed one type of conductivity (En 1-2), but on the other, a significant spatial variability of component intensity is evident *within* the CFO islands. The interface, showing a clearly different behavior in C-AFM maps, is not separated from the CFO island. Likewise, the BFO matrix component (En 3-4) remains merged with the interfacial behavior, and manifests conductivity much higher than that of BFO in raw data (0.4 nA vs. 80 pA @ 5 V). We omit here case with k = 3 for the sake of saving space, but note that it also represents an undersampling, which clearly shows components of the CFO core, CFO inner interface and BFO (all non-hysteretic), but misses at identifying the outer interfacial component with memristive (hysteretic) behavior, despite the fact that it is present in the original data. Unmixing the data into 4 components seems optimal from both the looks of the endmember IV curves and loading maps: the variability within CFO islands is now split into two components, highlighting the inner interface, BFO has a low-conductive component and the outer interface is ascribed its own hysteretic component. Oversampling the data at k = 5 keeps endmembers 1, 3 and 4 almost unchanged, but gemmates a new endmember (denoted 2' in Fig. S1o) from endmember 2. This process decreases the maximal intensity of endmember 2 from more than 90% to ca. 85%. The 2' endmember's maximal intensity is about the same – lower than the maximal intensities of all other components at k = 2, 3, 4 or 5. This is a clear sign of oversampling. In addition, the spatial localization of the endmember 2' behavior (Fig. S1t) is not very different from that of endmember 2 (Fig. S1s). Endmember 2' mainly represents conductive behavior at the core of one of the CFO islands, which is already described by component 2. A further increase in k leads to a continuous gemmation of new pseudo-components from the 4 optimal ones accompanied by a decrease in their intensity.

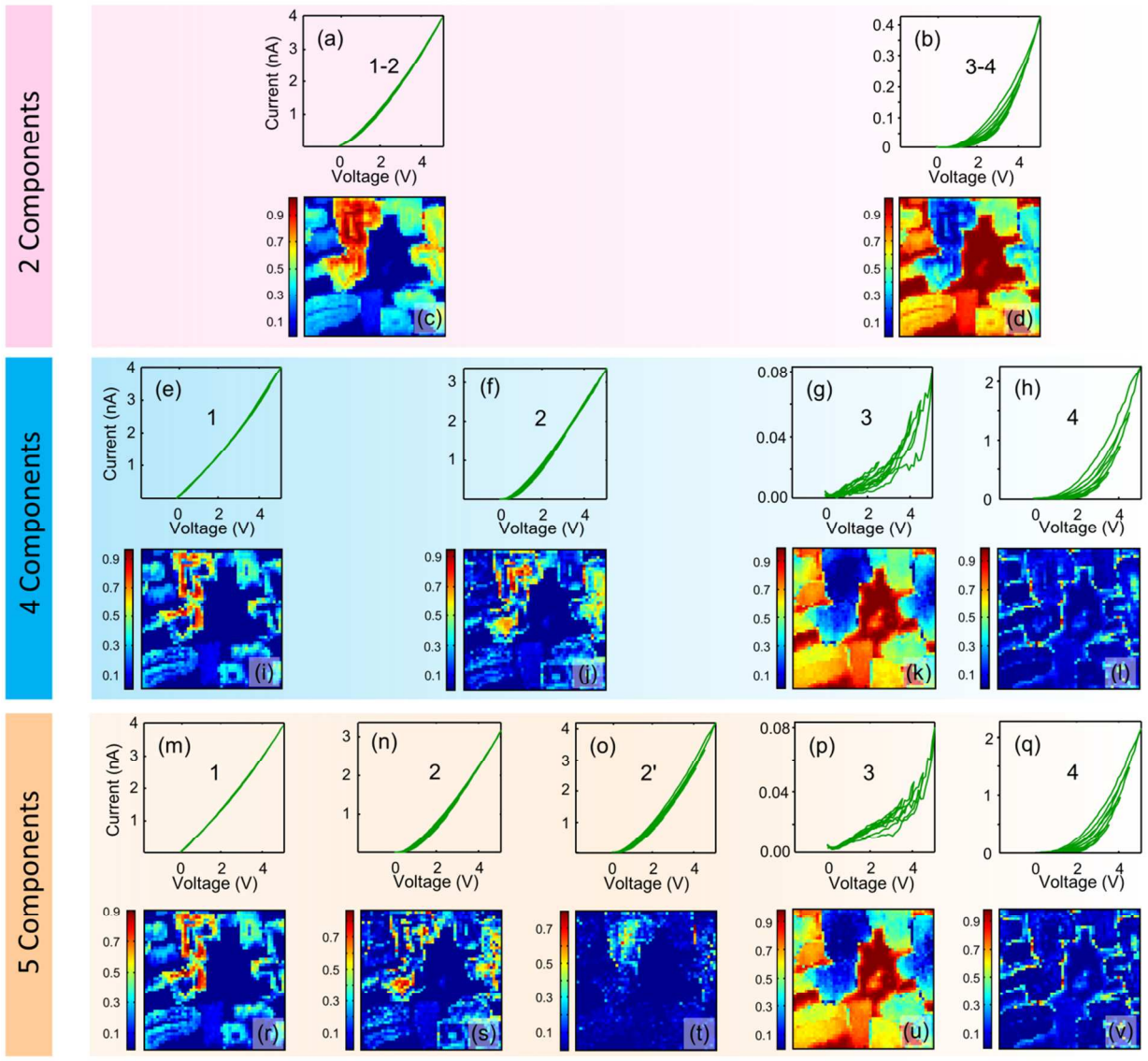

Figure S1. Unmixing the dataset into 2, 4 and 5 components with BLU algorithm illustrates the concept of under- and oversampling.