Deep data mining in a real space: Separation of intertwined electronic responses in a lightly-doped $BaFe_2As_2$

Maxim Ziatdinov[1,2,*], Artem Maksov[1,3], Li Li[4], Athena S. Sefat[4],

Petro Maksymovych[1,2], and Sergei V. Kalinin[1,2,3,#]

[1]*Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, TN, 37831*

[2] *ORNL Institute for Functional Imaging of Materials, Oak Ridge National Laboratory, Oak Ridge, TN, 37831*

[3]*Bredesen Center for Interdisciplinary Research, University of Tennessee, Knoxville, TN 37996*

[4]*Material Science & Technology Division, Oak Ridge National Laboratory, Oak Ridge, TN, 37831*

[*]E-mail: ziatdinovma@ornl.gov

[#]E-mail: sergei2@ornl.gov

Abstract

Electronic interactions present in material compositions close to the superconducting dome play a key role in the manifestation of high-T_c superconductivity. In many correlated electron systems, however, the parent or underdoped states exhibit strongly inhomogeneous electronic landscape at the nanoscale that may be associated with competing, coexisting, or intertwined chemical disorder, strain, magnetic, and structural order parameters. Here we demonstrate an approach based on a combination of scanning tunneling microscopy/spectroscopy (STM/S) and machine learning tools for an automatic separation and extraction of statistically significant electronic behaviors in the spin density wave (SDW) regime of a lightly (~1%) gold-doped $BaFe_2As_2$. We show that the decomposed STS spectral features have a direct relevance to fundamental physical properties of the system, such as SDW-induced gap, pseudogap-like state, and impurity resonance states.

Introduction. Nanoscale inhomogeneity of chemical, structural and electronic orders in a crystalline matter is expected to have a profound and non-random effect on the macroscopic properties of technologically relevant materials. Notorious examples include reduced mobility of Dirac electrons in graphene transistor devices due to formation of charge nanopuddles [1, 2], ultra-high piezoelectric response of relaxor ferroelectrics due to interaction between nanopolar domains and acoustic phonon mode [3], filamentary superconductivity [4], and fluctuating superconducting (SC) state above a transition temperature (T_c) in high-T_c cuprates associated with emergence of nanometre-sized electron pairing regions [5].

Scanning tunneling microscopy and spectroscopy (STM/S), which probes topographic and electronic properties of the surfaces with a nanometer-scale resolution, constitutes an ideal experimental tool for exploring local inhomogeneity in materials. The STM topographic images are typically recorded in a constant current regime [6], resulting in a 2-dimensional (2D) $Z(X,Y)$ dataset, where Z represents a convolution of actual height variation and electronic local density of states (LDOS) at each point (X,Y) on the surface. Meanwhile, the STS mode allows to acquire

3D $G(X,Y,V)$ datasets, where $G=dI/dV$ corresponds to a value of differential conductance proportional to LDOS at specific energy $E=eV$ at each (X,Y) point. In the simplest realizations of the bi-phase or multi-phase nanoscale systems, a separation between two or more phases is clearly visible in the STM topography, and comparison of STS spectra associated with different topographic features allows a straightforward analysis of electronic properties in these phases. Examples include STM/S measured on 2D lateral heterostructures sufficiently far from the boundary [7] or STM/S experiments on an isolated impurity embedded in otherwise ideal lattice [8]. For many strongly correlated materials, however, a complex local inhomogeneity patterns in conductance maps do not have a direct and simple connection to topographic features [see, for example, Ref. 9-11]. To complicate things even further, the morphology and chemical composition of the top-most layer of a cleaved surface in many complex compounds is usually itself a subject of controversy [12] which makes it nearly impossible to predict electronic properties in a characteristic field of view (FOV) of STM/S experiment from the first principles.

Given an ever-growing amount of multidimensional STM/S data on strongly correlated materials [12-14], there is an urgent need for developing a deep data based analysis that would allow reliable and un-biased identification and spatial mapping of statistically significant different electronic behaviors without *a priori* knowledge about the details of surface structure. Here we present a physics-robust machine learning style approach based on *k*-means clustering, principle component analysis (PCA) and Bayesian linear unmixing to uncover a wealth of "hidden" information from the STS datasets in a lightly-doped, "precursor", magnetic regime of iron-based superconductor [15]. We show how the features extracted from multivariate statistics-based decomposition of STS signal have a direct relevance to fundamental physical properties of the system, which we illustrate by uncovering a "buried" pseudogap-like phase and impurity induced double resonance states.

As a model system, we have chosen a lightly Au-doped $BaFe_2As_2$ single crystal, $Ba(Fe_{1-x}Au_x)_2As_2$ with x=0.009 (~1%). This compound shows a coupled structural and antiferromagnetic (AF) transition, from the tetragonal non-magnetic state into the orthorhombic striped SDW phase at $T_N \approx 110$ K [16]. Upon increased Au-dopants, the AF interactions becomes suppressed and the system develops into a superconductor ($T_c \approx 4$ K) at ~3%. It has been recognized that interactions present in such SDW states of the FeAs-based compounds play a

crucial role in understanding unconventional superconductivity [17, 18]. However, the details of local electronic structure at low temperatures in the non-SC phase of FeAs compounds, including the role of lattice strain, presence and origin of a pseudogap-like state, and character of impurity-induced quasiparticle states, remain a subject of a debate.

Results and discussion. We first present the STM topographic image over a relatively large FOV on a cleaved surface of 1% Au-doped $BaFe_2As_2$ in the SDW phase recorded at T=77 K [Fig. 1(a)]. The typical surface area at 77 K appears to be peppered with dark nanoscale regions. Upon cooling down to T=4 K, we found a dramatic increase in the density of the dark nanoregions as can clearly be seen from the representative STM topographic image in Fig. 1(b). In general, the variations in apparent topographic height associated with dark and bright regions can be of both topographic and electronic origin. However, we do not expect any extensive surface damage or profound changes in nanoscale chemical composition as we cool down the sample from 77 K to 4 K. Instead, the observed change in STM topographic patterns in Fig. 1(a) and 1(b) is likely related to the enhanced nanoscale electronic inhomogeneity as we approach towards a phase region with competing normal and SC orders [19] or with admixture of another form of magnetic order within the SDW phase (See Supplemental Material [20]). Such inhomogeneity shows the necessity for applying data mining tools based on multivariate statistical analysis for extracting relevant electronic behaviors in this system [21].

Zooming into a smaller FOV reveals a stripe-like reconstruction at the surface with a periodicity across the stripes of ≈0.7 nm [inset in Fig. 1(b)]. Similar unidirectional modulation of charge density has been also reported for SDW phase of $SrFe_2As_2$ [22]. While the exact origin of these charge stripes and their relation to SDW, if any, is not clear at present moment, it is worth to note that we were not able to observe similar 1D modulations at 77 K on the same cleaved surface. This suggests that the reconstruction is not cleavage-induced.

In Fig. 1(c) we show the STM topography at T=4 K measured in a region with extended quasi-1D defect which appears as a bright "diagonal" feature in the topography. The spatial extension of this defect typically exceeds ≈1 μm and we were able to reproducibly observe it in several areas of the sample. Furthermore, a similar structure was reported in another iron-based superconductor compound [23] suggesting that this defect can be a common feature of iron pnictides. High-resolution spatial maps of differential conductance G [Fig. 1(d-g)] recorded at

the area shown in Fig. 1(c) at several selective energies confirm a highly inhomogeneous electronic structure of the surface, with no one-to-one correspondence to underlying topographic data. Accordingly, averaging over even relatively small surface area can lead to a loss of significant physical information contained in individual *G* curves. Such a lossy compression of the original data is illustrated in Fig. 2 (b), where the STS curve averaged from 182 individual line spectra inside the box in Fig. 2 (a) fails to reproduce physically important features at the Fermi level seen in the 4 selected point spectra (gray, red, blue, and black spectral curves) recorded within the box area. It therefore becomes clear that the surface electronic behavior cannot be characterized reliably by a simple visual assessment of the topographic image and individual inspection of STS curves from $G(X,Y,V)$ dataset.

We now proceed to the accurate extraction of statistically significant information associated with surface electronic structure from a deep data style analysis [21]. We use the STS dataset recorded on the topographic area shown in Fig. 2(a). The dataset has dimensions of $X \times Y \times V = 50 \times 50 \times 768$, that is, it contains a stack of 768 conductance maps with a spatial resolution 50px×50px. To decorrelate the STS data in a statistically meaningful way we start with imposing a lower bound limit on the number of relevant electronic behaviors within the dataset. The smallest reasonable number of statistically significant behaviors can be estimated using *k*-means algorithm [25]. The *k*-means algorithm divides the dataset in a specified number of optimally selected clusters of curves that have similar behavior so that the within-cluster sum of squares is minimized [24, 25]:

$$\arg min \sum_{i=1}^{k} \sum_{x_j \epsilon S_i} \left\| x_j - \mu_i \right\|,$$

where μ_i is the mean of points in S_i. The selection of the number of clusters is based on the analysis of dendogram in Fig. 3 (b), in which larger vertical drops in the binary branches indicate a better cluster organization scheme in the data [25]. Based on the results shown in Fig. 3 (b), we used 3 clusters as an input in our *k*-means analysis. The resultant spatial distribution of the 3 clusters is shown in Fig. 3 (a), and the mean curves associated with each of 3 clusters are displayed by thick solid black line in Fig. 3(c-e).

We further analyzed a variance in the STS curves distributed over each nanoregion (cluster) by means of PCA [26-28]. The deviation from the mean curve within each cluster

associated with first 5 eigenvectors in PCA is shown by dashed lines in Fig. 3(c-e), and the corresponding scree plots are depicted in Fig. 3(f-h).The PCA analysis indicates that the cluster 2 and cluster 3 show a relatively moderate variance in the shape of the mean STS curve, allowing us to extract physical information from the curves. The STS curve from cluster 2 displays a metallic behavior and a well-defined dip at about 15 meV below the Fermi level. This is in a good agreement with an observation of the SDW gap centered at around -15meV in the ARPES measurements of $BaFe_2As_2$ [29]. We note that the theoretical model in [29] also showed that the SDW phase features a finite density of states at the Fermi level in the absence of the (coexisting) superconducting state, which is supported by our results. The mean STS curve from cluster 3 shows a metallic behavior and is somewhat similar to the curve from cluster 2, but with the center of the dip shifted to about -25 meV. We tentatively assign this behavior to the SDW phase whose characteristics were altered locally due to the strain induced by the quasi-1D defect [30]. Noteworthy, we did not observe similar lineshape in the regions far (>100 nm) from the defect in our experiments. The PCA-derived variance within cluster 2 and cluster 3 can be understood as relatively minor fluctuations of electronic response within a defined phase. The situation, however, is quite different for cluster 1. Here, a stronger variance in the shape of STS curves, especially in the regions close to the Fermi level, does not allow assigning any physically-defined phase. This suggests that the total number of relevant electronic behaviors is larger than estimated by the *k*-means method. However, additional, "hidden", electronic responses cannot be accurately revealed from the PCA eigenvectors, as they are constructed to be orthogonal and hence do not have a well-defined physical meaning.

To perform a more thorough and detailed separation of electronic behaviors in the STS dataset we adopt Bayesian linear unmixing (BLU) technique. This algorithm, developed by Dobigeon and co-workers, is used for separating linear mixtures of spectral sources under non-negativity and full additivity constraints [31] that allows assignment of physical meaning to the shape of the end-member curves [32]. The BLU approach assumes that a complete observation X is represented as a linear combination of independent positive endmembers, M,

$$X = MA + N,$$

where A are the relative abundances associated with each endmember, and N is an additive Gaussian noise. The detailed description of BLU method can be found in the Supplemental

Material [20]. We assume that the total STM current at each pixel in the dataset can be represented as a linear combination of the currents flowing through each of the available "channels", so that the latter can be represented by the endmembers.

The number of endmembers R in the BLU algorithm must be postulated by a researcher. The lower bound limit for a total number of endmembers has already been set by the results of k-means method. To set up the upper bound limit for possible number of relevant electronic behaviors, we refer to a general underlying physics of the problem. Here, in addition to the states associated with the SDW phase discussed in the k-means calculations, we must add states associated with (i) unidirectional modulation of surface charge density; (ii) possible presence of a different magnetic order "admixed" into the SDW phase; (iii) 2 common types of point defects on the cleaved surface; (iv) diluted concentration of Au dopants; (v) randomly scattered atoms on the surface [20, 33]. Using these constraints, coupled with the results of PCA analysis, and by performing over- and under-sampling of BLU R-components, we found that the most relevant description of electronic behavior is achieved for R=6 endmembers [20]. The BLU results with R=6, for both endmembers and abundance maps, are shown in Fig. 4. One can immediately see that endmember 4 and endmember 5 [Fig. 4(d) and 4(e)] corroborate the results on SDW-associated phase found earlier from k-means algorithm. In addition to phases already seen in the k-means, the BLU analysis revealed new features in electronic behavior that can be linked to the fundamental physical properties of the material, as described below.

The endmember 2 shows a well-defined signature of a spectral gap of $2\Delta \approx 40$ meV centered near the Fermi level [Fig. 4(b)]. We note that the gap of a similar behavior and magnitude (~30-40 meV) was observed by Madhavan and co-workers in their STM experiment on a closely related compound from AFe$_2$As$_2$ family, SrFe$_2$As$_2$, in which it was explained as the SDW-originated gap [22]. However, recent photoemission spectroscopy measurements and theoretical modelling [29] on the AFe$_2$As$_2$ type compound revealed that the SDW and SC orders must coexist spatially in order to produce a gap at the Fermi level. Otherwise, the SDW opens a gap below the Fermi level, in agreement with behavior observed in the endmembers 4 and 5. If the SDW and SC orders indeed coexist on a local scale in our sample, the formation of the electron-pairing "islands" associated with the SC order is expected to produce a continuous drop in bulk resistivity measurements [4, 34]. However, this is clearly not the case for our compound,

in which the resistance showed a small *upward* trend in the temperature range of interest [20]. This rules out a scenario in which the "admixture" of SC order leads to the gap feature seen in the endmember 2. We therefore describe the spectral weight loss at the Fermi level observed in the endmember 2 in terms of a pseudogap-like state, which is defined here as a state outside the "superconducting dome" and not directly associated with either SDW-induced gap or 'SDW+SC'-induced gap.

We next discuss a possible physical origin of the pseudogap-like state associated with endmember 2. At first glance, it is tempting to link the pseudogap-like state to the 1D striped charge order seen in the STM topographic image. However, a possibility of such direct correlation quickly falls apart as we were able to find the 1D stripes even in the areas without spectral features of a pseudogap [35]. Another explanation of a pseudogap-like state is based on the possible formation of a different, short-range, magnetic order admixed into the SDW phase, which is consistent with the presence of upturn in a magnetic susceptibility data below T_N. The formation of a pseudogap-like state may also explain a peculiar upswing in the resistivity in the SDW phase below ≈20 K [20]. Noteworthy, our finding of a $2\Delta \approx 40$ meV pseudogap-like feature, which is not directly related to SDW, correlates with photoemission spectroscopy results of Xu *et al*. [36] that showed emergence of $2\Delta \approx 36$ meV pseudogap state at the Fermi level of the underdoped $Ba_{1-x}K_xFe_2As_2$ in both SC phase and non-SC phase without long-range SDW order.

The endmembers 1 and 6 in Fig. 4 (a) and 4 (f) show a clear resonance peak features associated with impurity induced bound states in the SDW phase of this compound. The impurity-induced nature of these peaks is further confirmed by the inspection of the corresponding abundance maps that show the peak features are generally constrained to a point-like areas on the surface [Fig. 4 (g, j)]. Of particular interest is the endmember 1 which can be described by a non-magnetic impurity-induced double resonance peak model studied in [37]. We tentatively ascribe the origin of this spatially diluted double peak state to the Au dopants. It is worth noting that a spatial dependence of the energy position and intensity of the two impurity-induced peaks [37] allows in principle a further separation within the BLU scheme. Finally, the origin of endmember 3 is likely related to the minor instabilities ("noise") of the tunneling junction during the grid acquisition.

Conclusions and outlook. Our results on the identification of a surface nanoscale electronic structure in the underdoped state of FeAs-based superconductor, by means of a deep data style analysis are important for providing clues to understand how the high-temperature superconductivity may emerge in these systems. First, while there is a growing evidence of the pseudogap state formation in FeAs compounds [36, 38-40], there is still an open debate on the relation of the pseudogap to the superconducting state and on the role of magnetic correlations in the formation of the pseudogap. Our revelation of "buried" pseudogap-like spectral features in the SDW phase, combined with results of magnetic susceptibility and resistivity measurements, suggests a potential link between a pseudogap state and weak or short-range magnetism within SDW phase. Second, the real-space analysis of the electronic character of impurity-induced quasiparticle states found in the spectral unmixing of our data can be further used as a probe into the details of the strong correlations in the system. In this sense, it is natural to extend the deep data approach to the reciprocal space, which is commonly used to study quasiparticle interference pattern [41]. We expect that a nanoscale inhomogeneity in the electronic structure of the surface would produce spatially different scattering patterns at the same value of energy. The application of techniques such as sliding FFT combined with multivariate analysis [42] would allow hidden scattering patterns to be uncovered. Finally, we note that the presence of 1D charge modulation at the surface did not allow us to measure the atomic lattice constant in the regions close to the defects that showed peculiar changes in electronic behavior within the SDW phase. We do expect, however, that for the systems in which the atomic lattice can be resolved (i.e., no surface "reconstructions" occur), one can perform a direct data mining to correlate minute variations in atomic positions with the changes in spectral characteristics, such as the magnitude of SDW and/or SC gaps. As the ever-increasing amount of STM/S data on strongly correlated systems makes the individual inspection of datasets highly impractical and, in many cases, nearly impossible, the approach outlined here present an ideal tool for an accurate mapping of locally inhomogeneous electronic structure on the surfaces in an automated fashion of a full information extraction.

Acknowledgment:

This work was supported by the U.S. Department of Energy (DOE), Office of Science, Basic Energy Sciences (BES), Materials Science and Engineering Division. Research was conducted at the Center for Nanophase Materials Sciences, which is a DOE Office of Science User Facility.

References:

[1] J. Martin, N. Akerman, G. Ulbricht, T. Lohmann, J. H. Smet, K. von Klitzing, and A. Yacoby, Nat. Phys. **4**, 144 (2008).

[2] Y. Zhang, V. W. Brar, C. Girit, A. Zettl, and M. F. Crommie, Nat. Phys. **5**, 722 (2009).

[3] G. Xu, J. Wen, C. Stock, and P.M. Gehring, Nat. Mat. **7**, 562 (2008).

[4] K. Gofryk, M. Pan, C. Cantoni, B. Saparov, J.E. Mitchell, A.S. Sefat, Phys. Rev. Lett. **112**, 047005 (2014).

[5] K. Gomes, A. Pasupathy, A. Pushp, S. Ono, Y. Ando, and A. Yazdani, Nature (London) **447**, 569 (2007).

[6] Scanning Tunneling Microscopy, Edited by: Joseph A. Stroscio, William J. Kaiser (Academic Press, San Diego, 1993).

[7] B. Kiraly, A. J. Mannix, M. C. Hersam, and N. P. Guisinger, Chem. Mater. **27**, 6085 (2015).

[8] M. M. Ugeda, I. Brihuega, F. Guinea, and J. M. Gómez-Rodríguez, Phys. Rev Lett. **104**, 096804 (2010).

[9] C. Howald, P. Fournier, and A. Kapitulnik, Phys. Rev. B **64**, 100504(R) (2001).

[10] A. N. Pasupathy, A. Pushp, K. K. Gomes, C. V. Parker, J. Wen, Z. Xu, G. Gu, S. Ono, Y. Ando, A. Yazdani, Science **320**, 196 (2008).

[11] H. Beidenkopf, P. Roushan, J. Seo, L. Gorman, I. Drozdov, Y. S. Hor, R. J. Cava, A. Yazdani, Nat. Phys. **7**, 939 (2011).

[12] J. E. Hoffman, Rep. Prog. Phys. **74**, 124513 (2011).

[13] Ø. Fischer, M. Kugler, I. Maggio-Aprile, C. Berthod, and C. Renner, Rev. Mod. Phys. **79**, 353 (2007).

[14] K. Zhao, Y.-F. Lv, S.-H. Ji, X. Ma, X. Chen and Q.-K. Xue, J. Phys.: Condens. Matter **26,** 394003 (2014).

[15] G. R. Stewart, Rev. Mod. Phys. **83**, 1589 (2011).

[16] L Li, H. Cao, M. A. McGuire, J. S. Kim, G. R. Stewart, and A. S. Sefat, Phys. Rev. B **92**, 094504 (2015).

[17] E. P. Rosenthal, E. F. Andrade, C. J. Arguello, R. M. Fernandes, L. Y. Xing, X. C. Wang, C. Q. Jin, A. J. Millis, and A. N. Pasupathy, Nat. Phys. **10**, 225 (2014).

[18] T.-M. Chuang, M. P. Allan, J. Lee, Y. Xie, N. Ni, S. L. Bud'ko, G. S. Boebinger, P. C. Canfield, J. C. Davis, Science **327**, 181 (2010).

[19] V. Z. Kresin, Y. N. Ovchinnikov, S. A. Wolf, Phys. Rep. **431**, 231 (2006).

[20] See Supplemental Material at [*URL will be inserted by publisher*] for bulk magnetic susceptibility and resistivity measurements, additional STM data at 4 K and 77 K, and the details of multivariate statistical data analysis.

[21] S.V. Kalinin, B.G. Sumpter, R.K. Archibald, Nat. Mater. **14**, 973-980 (2015).

[22] F. C. Niestemski, V. B. Nascimento, B. Hu, E. W. Plummer, J. Gillett, S. E. Sebastian, Z. Wang, and V. Madhavan, arXiv:0906.2761.

[23] W. Lin, Q. Li, B. C Sales, S. Jesse, A. S Sefat, S. V Kalinin, M. Pan, ACS Nano **7**, 2634 (2013).

[24] S. Haykin, Neural Networks: A Comprehensive Foundation (Prentice Hall, 1999).

[25] E. Strelcov, A. Belianinov, Y.-H. Hsieh, S. Jesse, A. P. Baddorf, Y.-H. Chu, S. V. Kalinin, ACS Nano **8**, 6449 (2014).

[26] S. Jesse, S. V. Kalinin, Nanotechnology **20**, 085714 (2009).

[27] N. Bonnet, J. Microsc. (Oxford, U.K.) **190**, 2 (1998).

[28] A. Belianinov et al., Adv. Struct. Chem. Imag. **1**, 1 (2015).

[29] M. Yi, et al., Nat. Commun. **5**, 3711 (2014).

[30] S. Tan, et al., Nat. Mat. **12**, 634 (2013).

[31] N. Dobigeon, S. Moussaoui, M. Coulon, J.-Y. Tourneret, A. O. Hero, IEEE Trans. Signal Proces. **57**, 4355 (2009).

[32] E. Strelcov, A. Belianinov, Y.-H. Hsieh, Y.-H. Chu, S. V. Kalinin, Nano Lett. **15**, 6650 (2015).

[33] V. B. Nascimento, et al., Phys. Rev. Lett. **103** 076104 (2009).

[34] S. Eley, S. Gopalakrishnan, P. M. Goldbart, and N. Mason, Nat. Phys. **8**, 59 (2012).

[35] Manuscript under preparation

[36] Y.-M. Xu, P. Richard, K. Nakayama, T. Kawahara, Y. Sekiba, T. Qian, M. Neupane, S. Souma, T. Sato, T. Takahashi, H.-Q. Luo, H.-H. Wen, G.-F. Chen, N.-L. Wang, Z. Wang, Z. Fang, X. Dai, and H. Ding, Nat. Commun. **2**, 392 (2011).

[37] T. Zhou, H. Huang, Y. Gao, J.-X. Zhu, and C. S. Ting, Phys. Rev. B **83**, 214502 (2011).

[38] K. Ahilan, F. Ning, T. Imai, A. Sefat, R. Jin, M. McGuire, B. Sales, and D. Mandrus, Phys. Rev. B **78**, 100501 (2008).

[39] T. Mertelj, V. Kabanov, C. Gadermaier, N. Zhigadlo, S. Katrych, J. Karpinski, and D. Mihailovic, Phys. Rev. Lett. **102**, 117002 (2009).

[40] X. Zhou et al., Phys. Rev. Lett. **109**, 037002 (2012).

[41] J. E. Hoffman, PhD Thesis (UC Berkeley, 2003).

[42] R. Vasudevan et al., Appl. Phys. Lett. **106**, 091601 (2015).

FIGURES

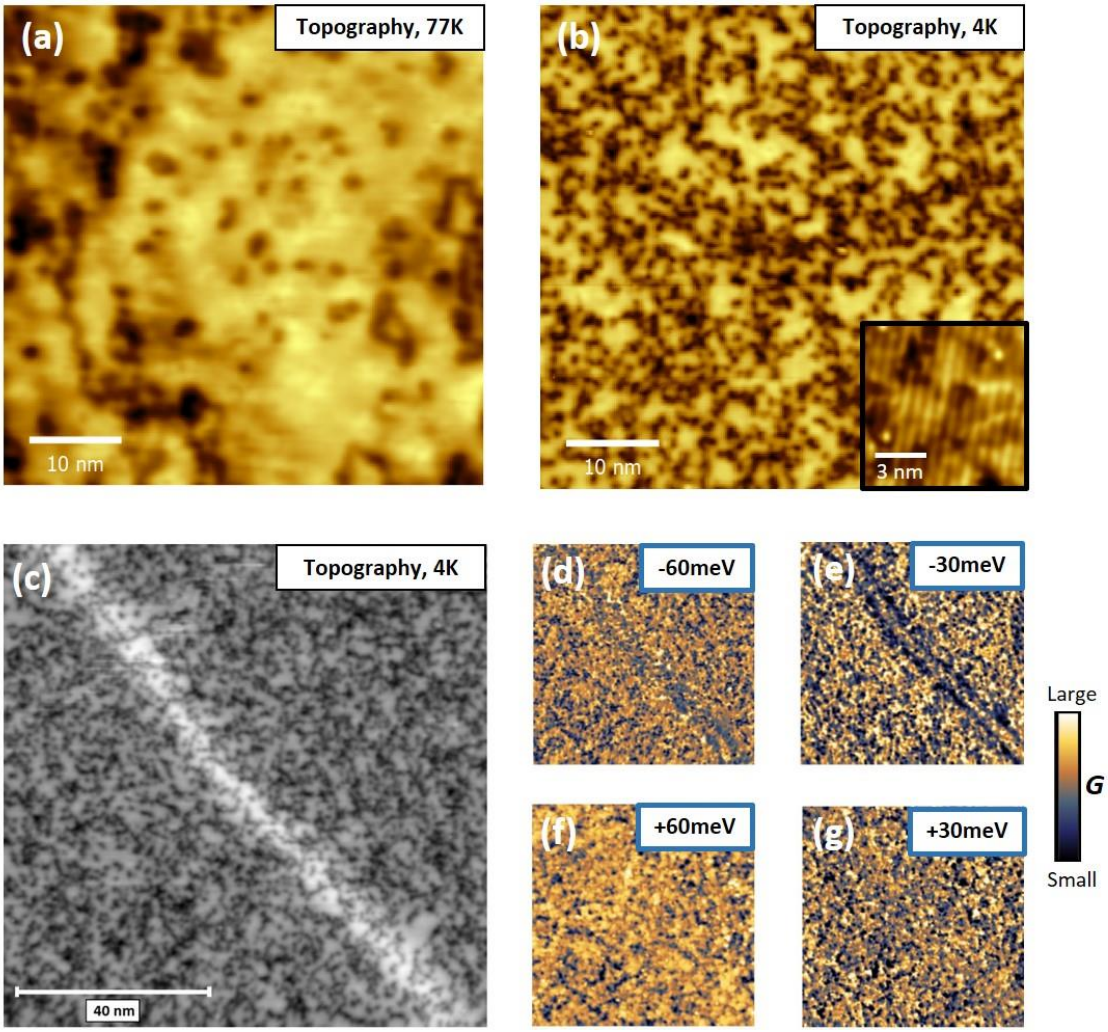

FIGURE 1. Data of 1% Au-doped $BaFe_2As_2$. (a,b) Representative experimental STM topographic images at 77K (a) and 4K (b). Tunneling conditions U=-60mV, I_s=100pA. Inset in (b) shows a zoomed-in area of the surface which displays stripe-like features. (c) STM topographic image of quasi-1D defect at 4K. U=-110mV, I_s=150pA (d-g) Conductance maps G (r, E=eV) in the same area as in (c) at several selected energies.

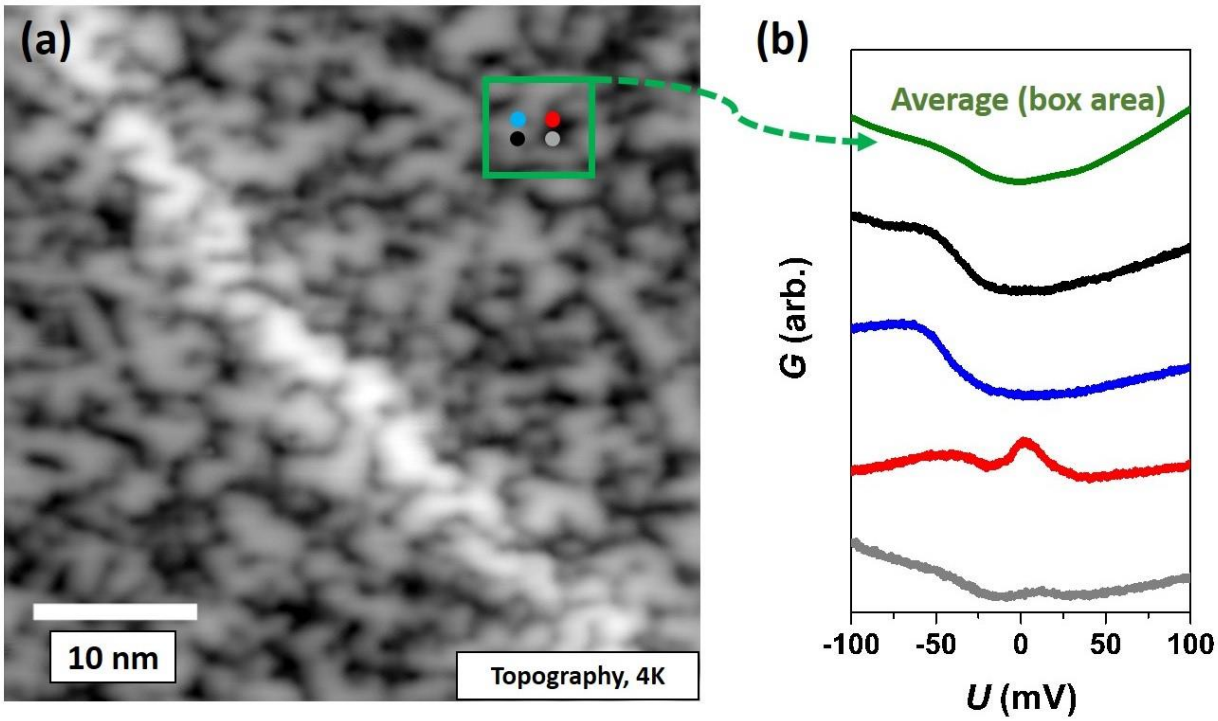

FIGURE 2. Data of 1% Au-doped $BaFe_2As_2$. (a) STM topographic image of the area on which STS grid measurements are performed. (b) Green STS curve is averaged over the 182 individual STS spectra inside the box in (a). Gray, red, blue, and black curve are single-point STS spectra recorded at the corresponding locations (denoted by colored dots) in (a). Note that gray and red curves were extracted from locations displaying similar topographic features.

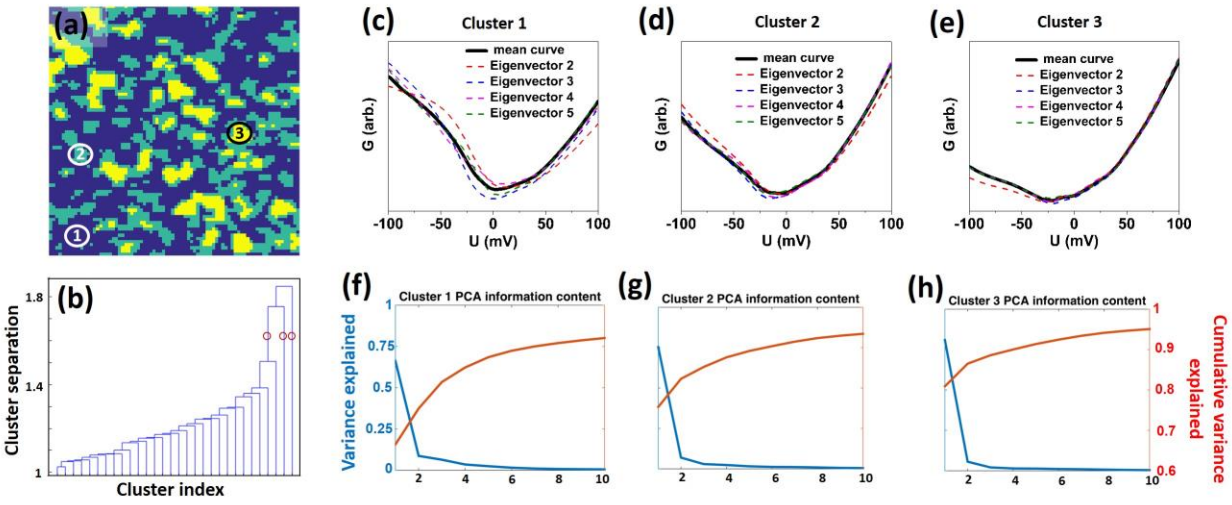

FIGURE 3. (a) *k*-means cluster algorithm resultant map with 3 clusters specified in the image. The surface area is identical to the one in Fig. 2(a). (b) Dendogram plot of hierarchical binary cluster tree (circles illustrate the optimal number of clusters). (c-e) Mean STS curves for each of 3 clusters shown in map (a) are plotted with black solid line. The PCA-derived deviation from mean curve within each cluster is shown by dotted color lines. (f-h) PCA scree plots showing variance within each of 3 clusters.

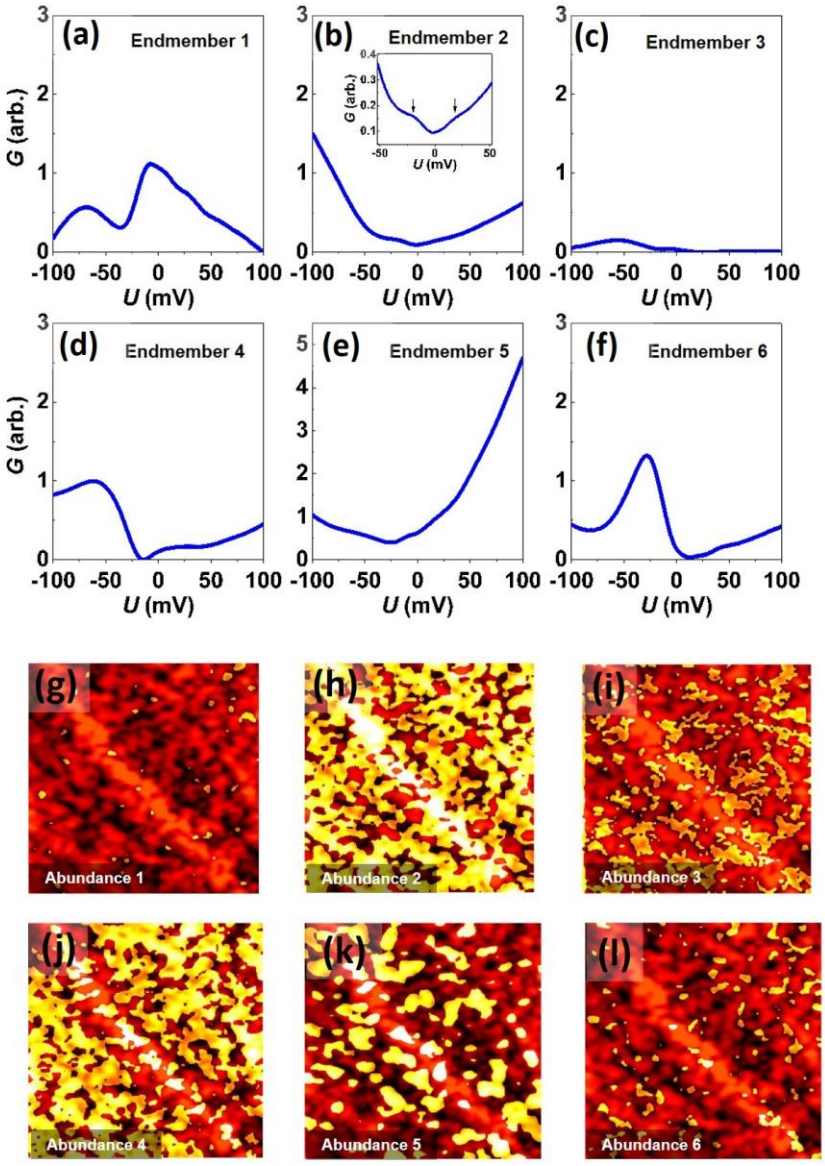

FIGURE 4. (a-f) 6 Bayesian endmembers (see text for details). Inset in (b) zooms in the spectral gap features at the Fermi level. (g-j) Corresponding abundance maps (yellow) overlaid on the topographic image (red). The intensity contributions in the abundance maps below 0.2 were cut off for a better visualization (see SM for a full (0,1) intensity maps).

Deep data mining in a real space: Separation of intertwined electronic responses in a lightly-doped $BaFe_2As_2$

Supplemental Material

Maxim Ziatdinov[1,2], Artem Maksov[1,3], Li Li[4], Athena S. Sefat[4],

Petro Maksymovych[1,2], and Sergei V. Kalinin[1,2,3]

[1]*Center for Nanophase Materials Sciences, Oak Ridge National Laboratory,*

Oak Ridge, TN, 37831

[2] *ORNL Institute for Functional Imaging of Materials, Oak Ridge National Laboratory,*

Oak Ridge, TN, 37831

[3]*Bredesen Center for Interdisciplinary Research, University of Tennessee,*

Knoxville, TN 37996

[4]*Material Science & Technology Division, Oak Ridge National Laboratory,*

Oak Ridge, TN, 37831

A.Experimental

1. Sample and tip preparations

Single crystals of lightly Au-doped $BaFe_2As_2$ (x= 0.009, ~1%) were grown out of self-flux using a high-temperature solution-growth technique [S1]. Temperature-dependent magnetic susceptibility, χ, decreases with decreasing temperature and drops abruptly below T_N = 110 K (Fig. A1), which also overlaps with the structural transition, T_s [S1]. Upon further cooling, χ increases in magnitude, which may be associated with additional magnetic contribution. Our neutron data (unpublished) confirms that the intensity of a wave vector relevant to SDW does not change in the corresponding temperature range, suggesting that the additional contribution to susceptibility comes from a different origin than simply enhanced SDW AF order. Resistance, *R*, diminishes with decreasing temperature from room temperature (R_{300K} ~ mOhm.cm), rising below ~ 140 K, and shows slight upward trend below 110 K ≈ $T_N = T_s$, and a couple of smaller features below (inset of Fig. A1) [S1].

STM/S measurements presented in the manuscript were carried using a Joule-Thomson scanning tunneling microscope (JT-STM, Specs, Berlin). Tungsten (W) STM tip was prepared by gentle field emission at a clean Ag(111) sample. The samples were cleaved *in situ* in the STM machine chamber at approximately 110 K. STS measurements were performed using a standard lock-in amplifier techniques, with a bias modulation between 2 mV and 5 mV.

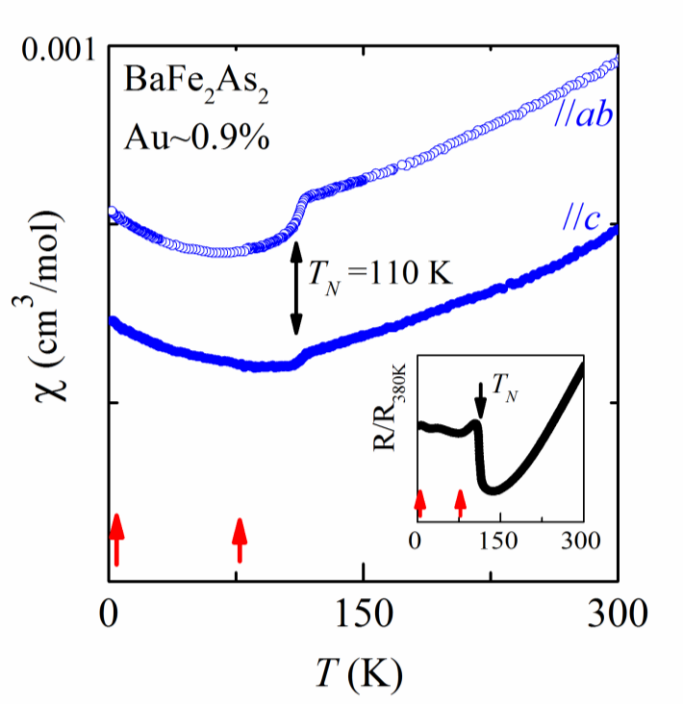

Fig. A1. Temperature-dependent magnetic susceptibility and resistance (inset) results of $BaFe_2As_2$, lightly doped with ~1 % gold (x=0.009). χ is measured along the two crystallographic axes; T_N is inferred from data; red arrows indicate STM/S temperature points at 4 K and 77 K.

2. STM at 77 K and at 4 K.

Our STM observations at 77 K on a cleaved surface shows a square-like lattice (Fig. A2) with a unit cell of (0.56 nm × 0.56 nm). This agrees well with measurements on $BaFe_2As_2$ reported in [S2]. We note that we were not able to observe the same square lattice at 4 K at nearly the same microscopic area on the sample surface. Instead, a well-defined 1D modulation of a charge density was typically observed at 4 K (Fig. A3).

During the STM measurements at 77 K, we were usually able to observe 2 types of bright protrusions (marked by red and orange ovals in Fig. A2), which we relate to specific point defects in the crystallographic lattice. We assume that these defects are preserved upon cooling down to 4 K, although we were not able to get a clear STM image of these defects due to large inhomogeneity and 1D modulations of charge density in the 4 K topographic images.

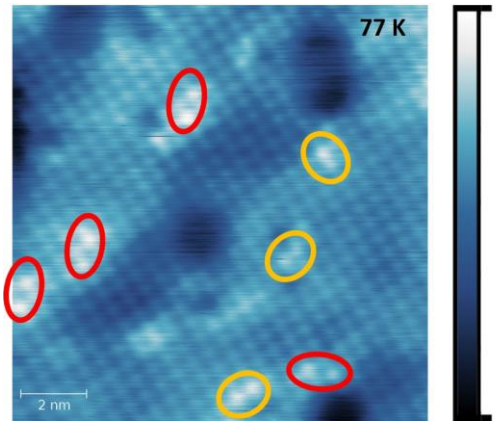

Fig. A2. STM topographic image recorded at T=77 K for 1% Au-doped $BaFe_2As_2$. Tunneling conditions U=100 mV, I_s=1 nA. Two common types of point defects are marked with red and orange ovals.

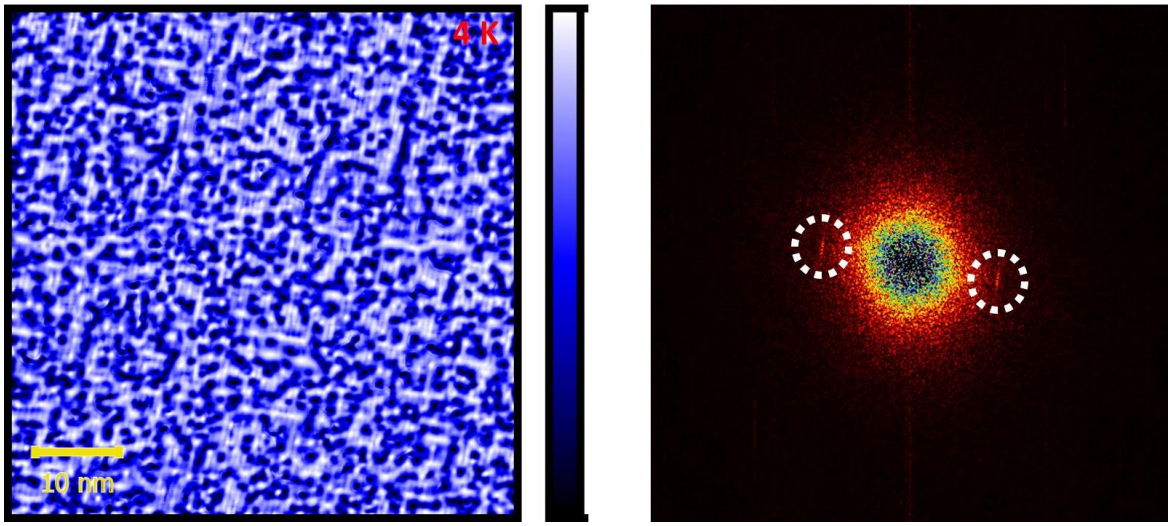

Fig. A3. (Left) STM topographic image recorded at T=4 K for 1% Au-doped $BaFe_2As_2$. Tunneling conditions U=-50 mV, I_s=200 pA. (Right) FFT on the topographic image on the left. Two spots corresponding to 1D modulation of a charge density are denoted by dotted circles.

B. Data analysis

1. Principle Component Analysis (PCA)

Using the results of *k*-means clustering as an input, we acquire separate $G_m(N_m, V)$ subsets, where N_m corresponds to number of observations in the cluster *m*, and *V* corresponds to number of parameters and is the same among all subsets (768 points in the -100 to +100 mV energy range). We perform PCA on each of the subsets to convert them into a product of expansion coefficients a_{ik} and eigenmodes w_k. The eigenvectors are orthogonal and are arranged in descending order by variance [S3]

$$G_i(\omega_j) = a_{ik} w_k(\omega_j) \quad \textbf{(1.1)}$$

As the data is not mean-centered, the first eigenvector is the average STS curve within the selected cluster, and all other eigenvectors are deviations from that average curve sorted in order of explaining maximum amount of variance within remaining data.

Dashed lines in Figure 3(c-e) in the main text demonstrate STS curves reconstructed from PCA by adding corresponding deviations to the average curve. Figure 3(f-h) shows the corresponding scree plots and cumulative variance explained for the first 10 eigenvectors within each cluster.

2. Bayesian Linear Unmixing: General background

STS data $G(X,Y,V)$ represents an example of hyperspectral imagery, where multiple dimensions *V* correspond to spatial distribution defined by (X,Y) coordinates. Our current assumptions is that the STM current can be represented as a linear combination of currents arising from different

sources, which makes this type of dataset a perfect problem for the Bayesian Linear Unmixing (BLU) [S4].

BLU is an advanced statistical method for decomposing a pixel spectrum into a collection of endmembers and corresponding abundances. In our data we have $P = X*Y$ pixels and V energy values. Linear mixing model assumes that spectral curve of a pixel p and overall dataset can be represented as:

$$y_p = \sum_{r=1}^{R} m_r a_{p,r} + n_p \quad \textbf{(2.1)}$$

$$Y = MA + N \quad \textbf{(2.2)}$$

where r is the endmember (R is the total number of endmembers), m_r is the spectral curve of an endmember, $a_{p,r}$ is the abundance of r at pixel p, and n_p is a zero-mean Gaussian noise, $Y = [y_1, \ldots, y_p]$, $M = [m_1, \ldots, m_R]$, $A = [a_1, \ldots, a_P]$, $N = [n_1, \ldots, n_P]$.

Since we did not have concrete predictions for the functional form of the potential sources to use as endmember guesses, we have utilized the unsupervised linear spectral mixture analysis. The model has been imposed with standard non-negativity and full-additivity constraints of the abundance coefficients:

$$\begin{cases} a_{p,r} \geq 0 \\ \sum_{1}^{R} a_{p,r} = 1 \end{cases} \quad \textbf{(2.3)}$$

Additionally, the model is constrained to non-negativity of the endmember spectra, and there is no assumption of the presence of pure pixels.

For the estimation of the prior for the Bayesian model, the data $X = MA$ dimensionality is reduced to K (R - 1 ≤ K ≤ V), by an assumption that without the noise data can be represented by (R – 1)-dimensional convex polytope of R^V, where vertices represent pure endmember spectra m_r. For the next step, PCA projection is obtained, which forms a simplex recovered through N-FINDR [S5]. Using these results, the endmember abundance priors as well as noise variance priors are estimated from the conjugate multivariate Gaussian distribution, where the posterior distribution is calculated based on the endmember independence using Markov chain Monte Carlo (MCMC), which generates asymptotically distributed samples probed via Gibbs sampling strategy. Unmixing was run for 100 MCMC iterations for each attempt.

3. Bayesian Linear Unmixing: choice of number of endmembers

To establish a number of endmembers in BLU analysis, we first used the results of *k*-means and PCA analysis (see also Fig. 3 in the main text). The mean curves associated with clusters 2 and 3 in *k*-means method show a relatively small variance in PCA analysis (see Fig. 3 (d, e, g, h) in the main text) and can be therefore described in terms of a physically defined electronic "phases". Furthermore, we also were able to see persistently the well-defined spectral features associated with these "phases" in BLU analysis as we varied a number of endmembers between R=4 and R=8 (described below). The situation is quite different for cluster 1, which shows relatively large variance in PCA analysis, and can be in principle decomposed into several spectra associated with distinct electronic "phases". The number of these "hidden" spectral behaviors can be estimated from PCA scree plot which suggests that the most relevant information associated with cluster 1 can be expressed by 4 PCA components (Fig. B.1). We therefore use the R=6 endmembers in the BLU analysis of the full dataset, which fits within the physical constraints on the number of distinct electronic responses, which was defined in the main text.

We next proceed to confirming our choice of the total number of endmembers by over- and under-sampling procedure. The proposed optimal number of the endmembers is R=6, as

determined earlier. Here, we demonstrate scenarios, in which the full dataset is BLU-unmixed into *R*=5 (under-sampling) and *R*=7, 8 (over-sampling) components. The results for endmembers and associated abundance maps are shown in Fig.B2 and Fig. B3, respectively. We will denote the spectral curve associated with a specific endmember for each unmixing result as m_r^R. The spectral features associated with spin-density wave induced gap (m_4^R and m_5^R curves) appear for both *R*=5 (under-sampling) and *R*=7, 8 (over-sampling). These are also the spectral features found earlier in k-means clustering.

On the other hand, it is clear that the under-sampling, *R*=5, leads to incomplete separation of other relevant electronic responses from experimental dataset. Indeed, impurity induced spectral features are not accurately revealed in *R*=5 case, in which the m_1^5 endmember exhibits only one broad peak [Fig. B2 (a)]. In addition, the spectral features of a pseudogap-like state are not seen for *R*=5. Comparison of abundance maps between *R*=5 and *R*=6 scenarios [Fig. B3 (a) and B3 (b)] suggests that this inability to accurately reveal the pseudogap-like state for *R*=5 is due to the partial transfer of the spectral weight from the "impurity phase" into the "pseudogap phase", which blurs over the spectral gap features in the positive energy range in m_2^5 curve. The presence of such transfer is confirmed by the inspection of relative intensities of m_1^5 and m_1^6 curves in the energy range from 0 meV to +25 meV, which shows the intensity of m_1^5 curve is roughly twice smaller than that of m_1^6 curve in this region (intensities are estimated as areas under the curves in the corresponding regions).

We next examine the over-sampling scenario. Both impurity double resonance and pseudogap spectral features are persistently seen for *R*=7 and *R*=8, in a good agreement with the results for *R*=6, as is confirmed from the inspection of m_1^R and m_2^R endmembers in Fig. B2 (b)-(d) and the associated abundance maps in Fig. B3 (b)-(d). We also note that the oversampling splits the m_5^7 curve [Fig. 1(c)] into the two curves represented by m_5^8 and m_6^8 endmembers [Fig. B2(d)]. This can also be clearly seen from corresponding abundance maps in Fig. B3(c) and B3(d). We characterize this split as an emergence of a pseudo-component which does not have a direct physical meaning and is a result of oversampling. We finally note that the observed fluctuations in the positon of the dip in a pseudogap-like state [Fig. B4] are likely related to a slightly different extraction of a "noise" component (endmember 3) from the full dataset in each case.

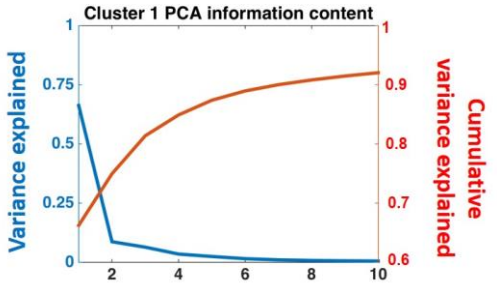

Fig. B1. PCA scree plot describing variance in *k*-means derived cluster 1

(see main text)

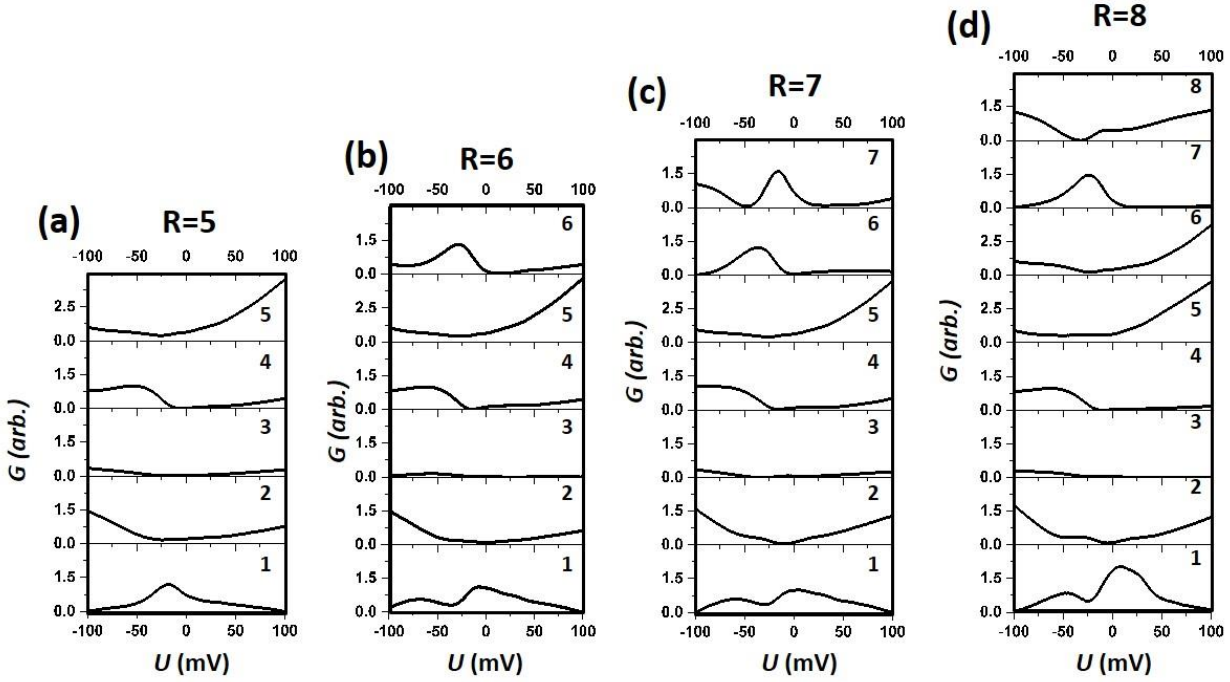

Fig. B2. Bayesian endmembers for different number of total endmembers: *R*=5 (a), *R*=6 (b), *R*=7 (c), *R*=8 (d).

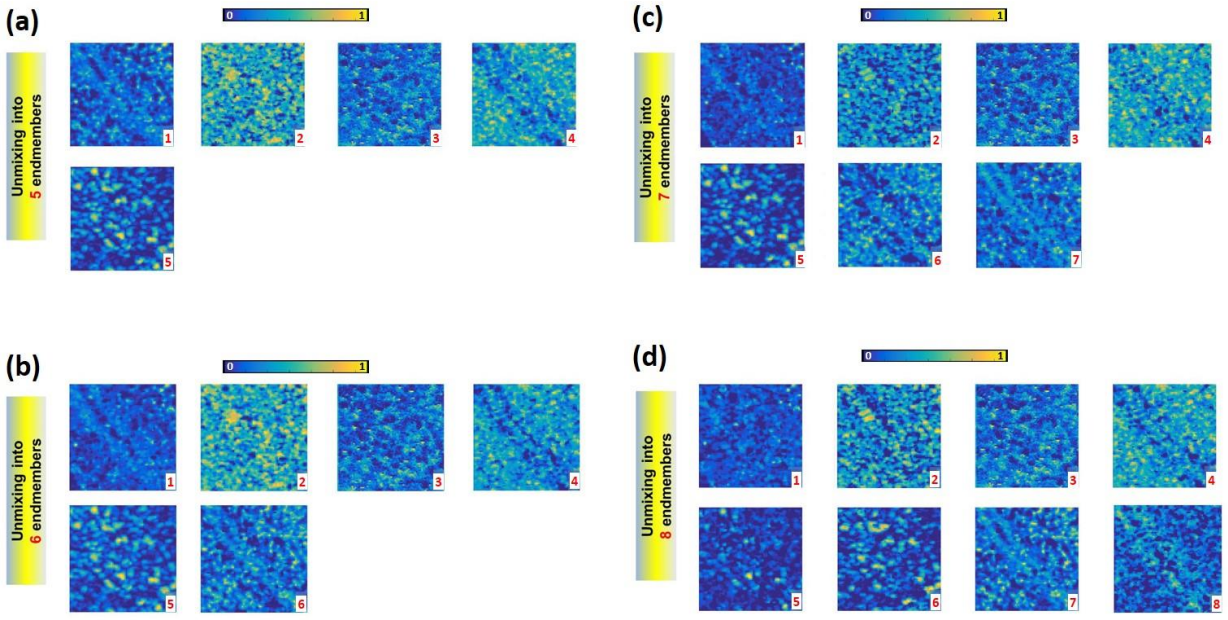

Fig. B3. Abundance maps associated with different number of total Bayesian endmembers in Fig. B2: *R*=5 (a), *R*=6 (b), *R*=7 (c), *R*=8 (d).

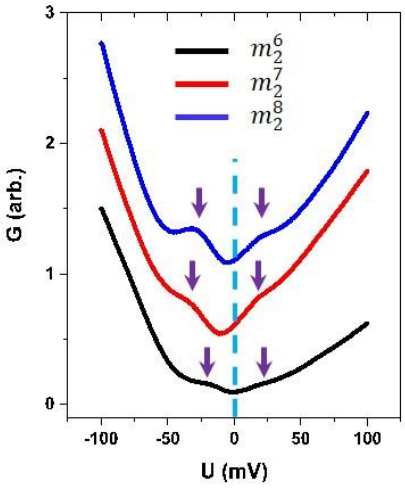

Fig. B4. Endmember associated with a pseudogap-like state for BLU unmixing into *R*=6, *R*=7, *R*=8 components. Arrows and dashed line are guides for eye only.

References:

[S1] L Li, H. Cao, M. A. McGuire, J. S. Kim, G. R. Stewart, and A. S. Sefat, Phys. Rev. B 92, 094504 (2015).

[S2] G. Li, X. He, J. Zhang, R. Jin, A. S. Sefat, M. A. McGuire, D. G. Mandrus, B. C. Sales, and E. W. Plummer, Phys. Rev. B **86**, 060512(R) (2012).

[S3] S. Jesse, S. V. Kalinin, Nanotechnology **20**, 085714 (2009).

[S4] N. Dobigeon, S. Moussaoui, M. Coulon, J.-Y. Tourneret, A. O. Hero, IEEE Trans. Signal Proces. **57**, 4355 (2009).

[S5] M. E. Winter, Proc. SPIE **266**, 264 (1999).