Chapter 5
Deep Data Analytics in Structural and Functional Imaging of Nanoscale Materials

Maxim Ziatdinov, Artem Maksov and Sergei V. Kalinin

Abstract Recent advances in scanning probe microscopy and scanning transmission electron microscopy have opened unprecedented opportunities in probing the materials structural parameters and electronic properties in real space on a picometre-scale. At the same time, the ability of modern day microscopes to quickly produce large, high-resolution datasets has created a challenge for rapid physics-guided analysis of data that typically contain several hundreds to several thousand atomic or molecular units per image. Here it is demonstrated how the advanced statistical analysis and machine learning techniques can be used for extracting relevant physical and chemical information from microscope data on multiple functional materials. Specifically, the following three case studies are discussed (i) application of a combination of convolutional neural network and Markov model for analyzing positional and orientational order in molecular self-assembly; (ii) a combination of sliding window fast Fourier transform, Pearson correlation matrix and canonical correlation analysis methods to study the relationships between lattice distortions and electron scattering patterns in graphene; (iii) application of a non-negative matrix factorization with physics-based constraints and Moran's analysis of spatial associations to extracting electronic responses linked to different types of structural domains from multi-modal imaging datasets on iron-based superconductors. The approaches demonstrated here are universal in nature and can be applied to a variety of microscopic measurements on different materials.

M. Ziatdinov (✉) · A. Maksov · S. V. Kalinin
Oak Ridge National Laboratory, Institute for Functional Imaging of Materials, Oak Ridge,
TN 37831, USA
e-mail: ziatdinovma@ornl.gov

M. Ziatdinov · A. Maksov · S. V. Kalinin
Oak Ridge National Laboratory, Center for Nanophase Materials Sciences,
Oak Ridge, TN 37831, USA

A. Maksov
Bredesen Center for Interdisciplinary Research, University of Tennessee, Knoxville,
TN 37996, USA

T. Lookman et al. (eds.), *Materials Discovery and Design*, Springer Series
in Materials Science 280, https://doi.org/10.1007/978-3-319-99465-9_5

5.1 Introduction

According to the established paradigm of structure-property relationship, there is a direct link between materials atomic structure and their optical, mechanical, electronic, and magnetic functionalities [1, 2]. This allows scenarios in which relatively small changes in the material structural and chemical compositions may have a decisive impact on the physical properties of the system. Examples include ultra-high piezoelectric response of relaxor ferroelectrics due to interaction between nanopolar domains and acouwestic phonon mode [3], filamentary superconductivity associated with nonuniform distribution of Pr dopants in iron arsenides [4], high critical-current density due to clustering of oxygen vacancies in cuprates [5–7], reduced mobility of Dirac electrons in graphene transistor devices due to formation of charge nano-puddles [8, 9], fluctuating superconducting state above a transition temperature (Tc) in high-Tc cuprates associated with emergence of nanometre-sized electron pairing regions [10], and emergence of glassy mixed-phases state in manganites linked to a quenched chemical disorder [11].

The advances in scanning transmission electron and scanning probe microscopies (STEM and SPM) have opened an unprecedented path towards simultaneously probing the material structural parameters (e.g. bond lengths) and its functional properties (e.g. electronic polarization or superconducting gap) in real space with a nanometer precision, making them the perfect tools for studying nanoscale inhomogeneities and their role in bulk crystalline behavior [12, 13]. Examples in SPM include direct imaging of chemical bonds in molecules [14], visualizing atomic collapse in artificial nuclei on graphene [15], and inferring mechanisms behind fundamental physical phenomena, such as high-Tc superconductivity, from single atom defect induced scattering patterns [6]. Meanwhile, STEM experiments can produce picometer-resolved images of ferroelectric polarization [16, 17], octahedral tilts [18], and chemical expansion strains [19]. Furthermore, combination of STEM and SPM with different spectroscopic techniques, such as optical and Raman spectroscopy, electron energy loss spectroscopy and mass spectroscopy have led to a rise of new multi-modal imaging capabilities that now allow a simultaneous capturing of materials structural, electronic, chemical, and optical properties at the nano and meso-scales. Such experimental capabilities allow, in principle, constructing combinatorial libraries of lattice configurations and functionalities at the single-defect level. This, however, requires first a development of methods for extracting all the experimentally accessible (spatially-dependent) information on structure and function variables and for cross-correlating the information from different "channels" in physically-meaningful and statistically-meaningful ways.

We illustrate several frameworks based on machine learning and multivariate analysis that allow automated and highly accurate extraction and mapping of different structural and functional descriptors from experimental datasets as well as studying their local correlations. The approach for a two-channel microscopic imaging experiment is schematically outlined in Fig. 5.1. It starts with recording 'structure' and 'function' information over the same sample area via two different acquisition

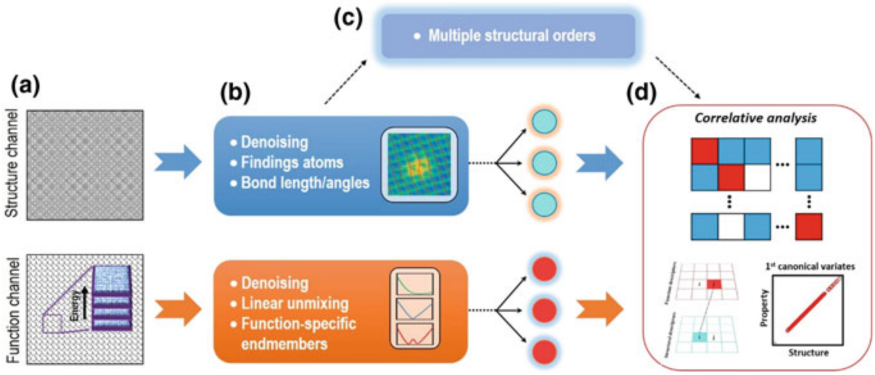

Fig. 5.1 Schematic workflow for structure-property relationships analysis. **a** 2-channel (‘structure’ and ‘function’) data acquisition. **b** Processing data from both channels to extract relevant structure and function descriptors. **d** Mining the combinatorial library of lattice configurations and functionalities. For systems with multiple structural orders one can apply correlative analysis ‘toolbox’ directly to the processed structural data (**c**–**d**)

channels (Fig. 5.1a). In this case, the first channel corresponds to 2D images in which Z is a ‘structural’ variable used to calculate lattice parameters, such as inter-atomic (or atomic columns) distances and apparent heights. The second channel represents 3D dataset in which G is a ‘function’ variable, for example, differential conductance or electron energy loss. After performing an image alignment such that, the data from both channels is cleaned from spurious noise features and outliers in a way that minimizes the information loss (e.g., using principal component analysis). The next step is constructing structural and functional descriptors. For structure channel, one may adapt various pattern recognition techniques from a field of computer vision, such as sliding window Fast Fourier Transform, deep neural networks and Markov random field. For function channel, blind source un-mixing/decomposition methods such as Bayesian linear unmixing and non-negative matrix factorization performed on hyperspectral “functional” data can generally provide a physically meaningful separation of spectral information when multiple ‘phases’ are present in the dataset (Fig. 5.1b, c). Once completed, one proceeds to performing direct data mining of structure-property relationships from correlative analysis of the derived structural and functional descriptors (Fig. 5.1d). The correlation analysis ‘toolbox’ typically includes methods such as Pearson correlation matrix, global and local Moran’s correlative analysis, and linear and kernel canonical correlation analysis. Note well that for systems with multiple order parameters and/or systems where both structural and electronic information can be effectively extracted from a single image, the correlation analysis can be performed directly on variables extracted from the structure channel.

In the following, we analyze structure-property relationship on different molecular and solid state systems using data obtained from constant-current mode and spectroscopic mode of scanning tunneling microscope [20]. The STM topographic images obtained in a constant-current mode represent a 2-dimensional dataset where $Z(\mathbf{R})$ is a convolution of height variations and electronic density of states in each $\boldsymbol{R}(X, Y)$ point (pixel) on the surface. The spectroscopic mode of STM (usually referred to as STS) produces a 3-dimensional set of data where the value of differential conductance $G(\boldsymbol{R}', V)$ is proportional to local density of states at specific energy $E = eV$ at each $\boldsymbol{R}(X', Y')$ point on the surface. For all cases studied here $\boldsymbol{R}(X', Y') = \boldsymbol{R}(X, Y)$. The necessary mathematical frameworks will be introduced separately for each case study.

5.2 Case Study 1. Interplay Between Different Structural Order Parameters in Molecular Self-assembly

5.2.1 Model System and Problem Overview

To demonstrate an application of advanced data science tools to molecular resolved STM images, a self-assembly of $C_{21}H_{12}$ molecules [21, 22] is chosen as a model system (Fig. 5.2a). Each individual molecular unit in the self-assembly can be viewed as a fragment bowl of buckminsterfullerene (hereafter, buckybowl). A buckybowl in the self-assembly can reside in two different structural conformations (bowl-up and bowl-down) as well as in multiple lateral orientations with respect to the substrate. In the absence of external perturbation and/or substrate disorder the molecular monolayer forms a long-range superperiodic pattern, in which each bowl-down state is surrounded by six bowl-up states. In the following, this superstructure is referred to as **2U1D**, where **U** and **D** stand for bowl-up and bowl-down states, respectively. At the low tip-sample separation distances in the constant current STM experiment (typically achieved at sample bias $U_s \lesssim 0.1$ V) it is usually possible to induce a switching between different molecular degrees of freedom via mechanochemistry effects, whereas at large separation distances (at $U_s \gtrsim 1$ V) the switching events, particularly those involving structural changes, are minimized [22]. Thus one can interpret the scans at low and high bias voltages as "writing" (albeit randomly) and "reading" molecular patterns, respectively. The representative STM image of buckybowl self-assembly is shown in Fig. 5.2c. The STM data used as an input in the current analysis was acquired in the reading regime; prior to acquisition of the image of interest, several STM scans were performed over the same area at the lower tip-surface distances (switching regime) producing additional "excitations", that is, enhancing a disorder, in the initial molecular structure. A global 2-dimensional Fast Fourier Transform (2D FFT) obtained from image in Fig. 5.2c shows a strong suppression of peaks associated with **2U1D** structure (compared to peaks in the outer hexagon associated with positional order in molecular lattice) indicating a presence of disorder in the molecular film. In the following, an approach based on a synergy of ab-initio simulations, Markov random field model and convolutional neural network

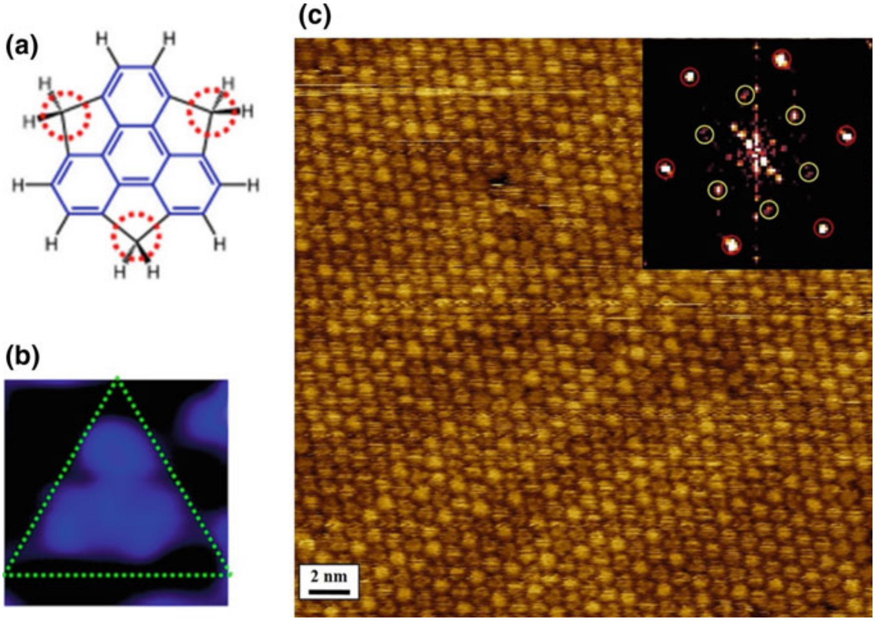

Fig. 5.2 Self-assembly of sumanene molecules (buckybowls) on gold substrate. **a** Chemical structure of sumanene. **b** Experimental STM image of individual buckybowl. Adapted with permission from [22]. Copyright 2018 American Chemical Society.**c** Large-scale STM image over field of view with approximately 1000 molecules. The inset shows FFT transform of data in (**c**). The yellow circles denote FFT spots associated with a formation of **2U1D** superlattice. Adapted from [23]

is introduced for "reading out" complex molecular patterns of buckybowls on gold substrate from molecule-resolved STM images [23].

5.2.2 How to Find Positions of All Molecules in the Image?

The first crucial step in analyzing the STM data on complex surface molecular structures is the identification and extraction of positions of all molecules for each image. Simple visual examination of STM image in Fig. 5.2c suggests that it contains up to about 1000 individual molecules. The normalized cross-correlation is performed to obtain correlation surfaces defined as

$$\gamma(\mathrm{u, v}) = \frac{\sum_{x,y}[f(x,y) - \overline{f}_{u,v}][t(x-u, y-v) - \overline{t}]}{\{\sum_{x,y}[f(x,y) - \overline{f}_{u,v}]^2 \sum_{x,y}[t(x-u, y-v) - \overline{t}]^2\}^{0.5}} \tag{5.1}$$

where f is the original image, t is the template, $\overline{f}_{u,v}$ is the mean of $f(x,y)$ in the region under the template, $\overline{t}$ is the mean of the template. The bowl-up DFTsimulated STM image is chosen as a template, which produced the highest accuracy in determination the positions of molecular centers. The uniform threshold is applied to the generated correlation surface γ, with cutoff set to 0.35, in order to maximize the number of extracted molecules. This results in a binary image, for which the connected components are identified and their centers are assigned as centers of the corresponding molecules. The apparent height I_m of each molecule, which represents a convolution of an actual geometric height and local density of electronic states, is calculated as $I_m = \sum_{x=1}^{15}\sum_{y=1}^{15} i_{x,y}$ where $i_{x,y}$ is the intensity of pixel at position *x, y* in the extracted image patch for molecule *m*. The summation is performed for 15 × 15 pixel patches around the center of each molecule. To remove outliers due to possible contaminations on a surface which may not directly associate with molecules, a maximum intensity value defined as $I_{max} = mean(I) + 3 * std(I)$ is introduced such that all intensities that exceed the maximum value are scaled back set to I_{max}.

Once all positions and intensities are identified a principal component analysis is performed on the stack of images of individual molecules. The aim of the principal component analysis (PCA) can be interpreted as finding a lower dimensional representation of data with a minimum loss of important (relevant) information [24]. Specifically, in PCA one performs an orthogonal linear transformation that maps the data into a new coordinate system such that the greatest variance comes to lie on the first coordinate called the first principal component, the second greatest variance on the second coordinate, and so forth. Hence, the most relevant information (including information on the orientation/rotation of molecules) can be represented by a small number of principal components with the largest variance, whereas the rest of the (low-variance) components correspond to 'noise'. The PCA analysis suggests that suggests that a likely number of rotational classes needed to be considered for this dataset is four.

5.2.3 Identifying Molecular Structural Degrees of Freedom via Computer Vision

Convolutional neural networks. The identification of molecular "shapes" (different orientation with respect to substrate) is performed using a technique from a field of computer vision known as convolutional neural networks. Convolutional neural networks (cNN) represent one of the key examples of a successful application of neuroscientific principles to the field of machine leaning. The cNNs are used for processing data which is characterized by a known, grid-like topology such as 2-dimensional grid of pixels obtained in the STM constant current experiments [25]. The architecture of the convolutional network used in the current work is shown in Fig. 5.3a and it includes convolutional layers, pooling layers, as well as a fully connected "dense" layer. The convolution layer is formed by running *learnable* kernels ('filters') of the selected size over the input image (or image in the previous layer).

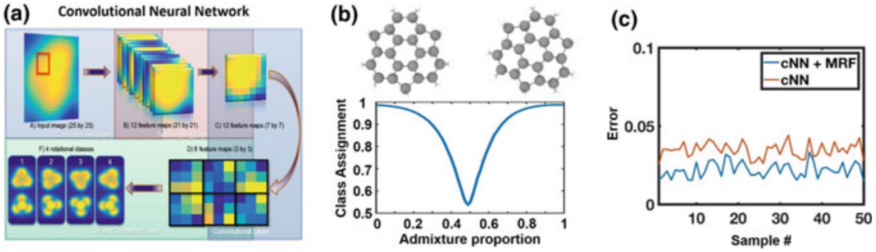

Fig. 5.3 Deep learning of molecular features. **a** Schematic graph of convolutional neural network (cNN) architecture for determining of molecular lateral degrees of freedom on the substrate. **b** Role of dynamical averaging (admixture of a different rotational class) in probability of the correct class assignment. **c** Error rate for cNN only and for cNN refined with Markov random field model. Adapted from [23]

The pooling layers produce downsampled versions of the input maps. The *i*-th feature map in layer *l*, denoted as V_i^l can be expressed as [26]

$$V_i^l = \sum_{i \in M_i} V_j^{(l-1)} * K_{i,j}^l + B_i^l \tag{5.2}$$

Here *K* is a kernel connecting the *i*-th feature map in layer *l* and the *j*-th feature map layer $(l-1)$, B_i^l describes the bias, and M_i corresponds to a selection of input maps. The output Z_i^l is a fully connected ("dense") layer that takes as input the "flattened" feature maps of the layer below it:

$$Z_i^l = \sum_{i \in M_i} \sum_{m \in M_i} \sum_{n \in M_i} (V_j^{(l-1)})_{m,n} W_{i,j,m,n}^l \tag{5.3}$$

where $W_{i,j,m,n}^l$ connects *i*-th unit at position *m, n* in the feature map of layer $(l-1)$ to the *j*-th unit in layer *l*. The cNN is trained on a set of synthetic STM images (25,000 samples) obtained from DFT simulations of different rotational classes.

Markov random field. The unique aspect of the present approach is that the cNN is followed by Markov random field model [27] which takes into account probabilities of neighboring molecules to be in the same lateral orientation on the substrate. This allows us to "refine" the results learned by neural network in a fashion that takes into account physics of the problem. The MRF model makes use of an undirected graph $G = (V, E)$, in which the nodes V are associated with random variables $(X_v)_{v \in V}$, and E is a set of edges joining pairs of nodes. The underlying assumption of Markov property is that each random variable depends on other random variables only through its neighbors:

$$X_v \perp X_{V \backslash v \cup N(v)} | X_{N(v)}, \tag{5.4}$$

for $N(v) =$ neighbors of v. Importantly, the explicit Markov structure implicitly carries longer-range dependencies. These priors are directly linked to the underlying

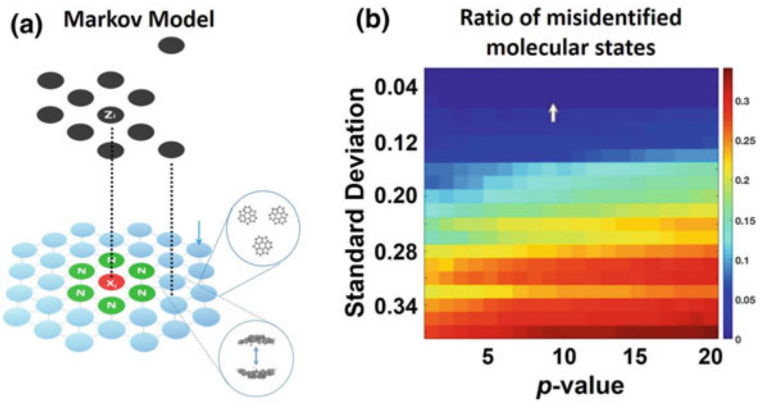

Fig. 5.4 Molecular self-assembly as Markov random field model (MRF). **a** Graphical Markov model structure used for analysis of a molecular self-assembly. **b** Error rate as a function of standard deviation of normalized STM intensity distributions and an optimization parameter (*p*-value). The arrow shows the value of these parameters for the analysis of the synthetic data. Adapted from [23]

physics of the system, that is, the presence of short-range interactions in molecular assembly which are now explicitly taken into account during image analysis. The experimental STM data on buckybowls is mapped on to a graph such that each molecule is represented as a node, and edges are connections to each molecule's nearest neighbors (Fig. 5.4a). The posterior distribution of an MRF can be factorized over individual molecules such that

$$P(x|z) = \frac{1}{Z} \prod_{<ij>} \Psi_{ij}(x_i, x_j) \prod_{i} \Psi_i(x_i, z_i) \tag{5.5}$$

where Z is the partition function, and $\Psi_i(x_i, z_i)$ and $\Psi_{ij}(x_i, x_j)$ are unary and pairwise potentials, respectively. These potentials are defined based on the knowledge about physical and chemical processes in the molecular system, such as a subtle interplay between a difference in adsorption energy for **U** and **D** molecules, molecular interactions different molecular configurations, and imperfection of the substrate. Finding an exact solution to MRF model is intractable in such a case as it would require examining all 2^n combinations of state assignments, where n is the number of molecules, that is, about 1000 for examined images. However, one can obtain a close approximate solution by using a max-product loopy belief propagation method [28], which is a message-passing algorithm for performing inference on MRF graphs, with unary and pairwise potentials as an input. Briefly, from initial configuration, nodes propagate message containing their beliefs about state of the neighboring nodes given all other neighboring nodes messages. This results in an iterative algorithm. All messages start at 1, and are further updated as max-product of potentials and incoming messages:

$$msg(x_j)_{i\rightarrow j} = max_l[\sum_{x_i} \Psi_{ij}(x_i, x_j)\Psi_i(x_i, z_i) * \prod_{k=neighbors\ of\ i\neq j} msg(X)_{k\rightarrow i}] \tag{5.6}$$

At each iteration belief is calculated for each node and the state with highest belief is selected, until message update converges:

$$Belief(x_i) = \Psi_i(x_i, z_i) * \prod_{j=neighbors\ of\ i} msg(x_i)_{j\rightarrow i} \tag{5.7}$$

According to theoretical modeling, it is unlikely that two neighboring molecules can have the same rotational state [29]. Therefore assign probability of each class to have a neighbor of its own class is considered to be 1% and probabilities to have a neighbor of other 3 rotational classes is considered to be 33%. Finally, the decoding using loopy belief propagation is performed in order to acquire a more precise solution. Note well that by tuning a graph structure and/or form of the potentials one can easily apply Markov random field approach to other molecular order parameters or even different molecular architectures. Indeed, one can also apply MRF to decoding different conformational states of molecules (note that an application of the cNN to a problem of determining different conformational states typically returns relatively poor results). For MRF modelling of bowl-up and bowl-down states, the unary potentials $\Psi_i(x_i, z_i)$ over molecular states are assigned based on the proximity of a particular molecule's intensity in the STM image to the threshold value between the states T. The node probabilities are calculated as two logistic functions:

$$\Psi_i(x_i = 1, z_i = I_i) = \frac{1}{1 + Exp[S * (T - I_i)]} \tag{5.8a}$$

$$\Psi_i(x_i = 2, z_i = I_i) = 1 - \Psi_i(x_i = 1, z_i = I_i) \tag{5.8b}$$

where $I_i \in [0, 1]$ is the intensity of a given molecule i, and S is a parameter that controls the growth rate of the logistic function. The logistic functions allow us to assign molecular intensities sufficiently far from the threshold as belonging to their corresponding class with probability of $\sim$1, while also providing more flexibility in the region around the threshold value itself. Next, the pairwise potentials $\Psi_{ij}(x_i, x_j)$ for the molecular system are determined. The optimal **2U1D** configuration proposed above is characterized by six **U** molecules surrounding one **D** molecule, such that **D** molecule is never allowed to have the nearest neighbor in the same bowl conformation. As we are interested in the distortion of an ideal **2U1D** structure (six bowl-up molecules surrounding one bowl-down molecule), a disorder parameter p is introduced such that a probability of **D** and **U** molecules having their neighbor in the same conformational state becomes p and $1 - p$, respectively.

Testing on synthetic data. Prior to analyzing real experimental data, a validity of the described approach is tested on synthetic dataset(s). Specifically, the DFT-based calculations of the STM signal associated with an individual molecule for each configuration are combined with Markov Chain Monte Carlo sampler to generate synthetic images of molecular self-assembly containing a large number ($\succcurlyeq$1000) of molecules.

Additionally, the synthesized data is "distorted" by addition of blurring associated with a convolution with the STM tip probe function, Poisson noise associated with tunnelling statistics, and dynamical averaging due to potential admixture of another azimuthal rotational state to a given structural configuration. Since the exact distribution of molecular states in synthetic data is known for each sample, one can evaluate an error rate for this method. It was found that the proposed approach results in a remarkably accurate identification of different molecular conformational and rotational states in scenarios where the distribution of the STM intensities in the synthetic data closely resembles the typical experimental data. The MRF approach allowed to identify accurately distributions of bowl-up and bowl-down configurations in the large scale synthetic STM images, even when no estimations regarding the *p*-value is available apriori (Fig. 5.4b), while its addition to cNN helped to improve the decoding results by reducing number of misclassified states (Fig. 5.3c). It was also found that the cNN framework allows to obtain a reliable classification of molecules rotational states even in the presence of relatively strong dynamical averaging between proximate rotational states of the molecule (Fig. 5.3b) which is relatively common in the STM experiments [30, 31].

5.2.4 Application to Real Experimental Data: From Imaging to Physics and Chemistry

Having confirmed that the introduced approach works on synthetic data we proceed to analysis of real experimental data. The results if full decoding of rotational (via cNN+MRF) and conformational (via MRF) states are presented in Fig. 5.5. Once a full decoding is performed, it becomes possible to explore a nature of disorder in the molecular self-assembly by searching for *local* correlations between different molecular degrees of freedom. Of the specific interest is a potential interplay between molecule bowl inversion and azimuthal rotation of the neighboring molecules. To obtain such an insight, method based on calculation the so-called Moran's I is adopted that can measure a spatial association between the distributions of two variables at nearby locations on the *lattice* [32]. The 'correlation coefficient' for global Moran's I is given by

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(X_i - \overline{X})(Y_j - \overline{Y})}{\sum_i (Y_i - \overline{Y})^2} \tag{5.9}$$

where N is the number of spatial units, X and Y are variables, $\overline{X}$ and $\overline{Y}$ are corresponding means, and w is the weight matrix defining neighbor interactions. It is worth noting that the presence of the spatial weight matrix in the definition of Moran's I allows us to impose constrains on the number of neighbors to be considered. For highly inhomogeneous system, one may use the so-called local indicators of spatial association which can evaluate the correlation between two orders at the neighboring points on the lattice for each individual coordination sphere. This is achieved through

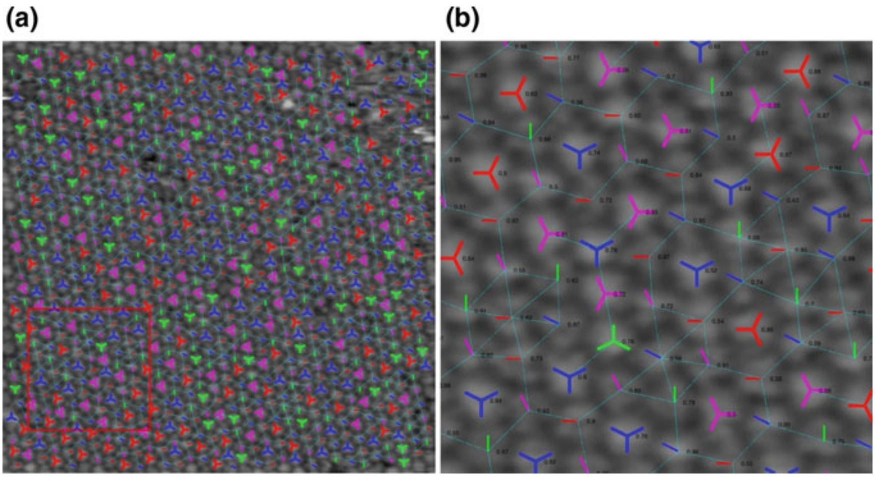

Fig. 5.5 Application of the current method to experimental data of buckybowls on gold (111). Decoding of rotational states (cNN+MRF) and bowl-up/down states (MRF, $p = 7$) for the experimental image from Fig. 5.1c. **b** Zoomed-in area from *red rectangle* in **a** where numbers denote an accuracy of state determination. Adapted from [23]

calculating *local* bivariate Moran's I for each spatial unit such as

$$I_{xy} = \frac{\sum_i \sum_{j \neq i} w_{ij} x_i y_i}{W} \tag{5.10}$$

where x and y are standardized to zero mean and variance of 1.

The results for spatial correlation between bowl-up/down configuration and different rotational classes for the first 'coordination sphere' is shown in Fig. 5.6a where

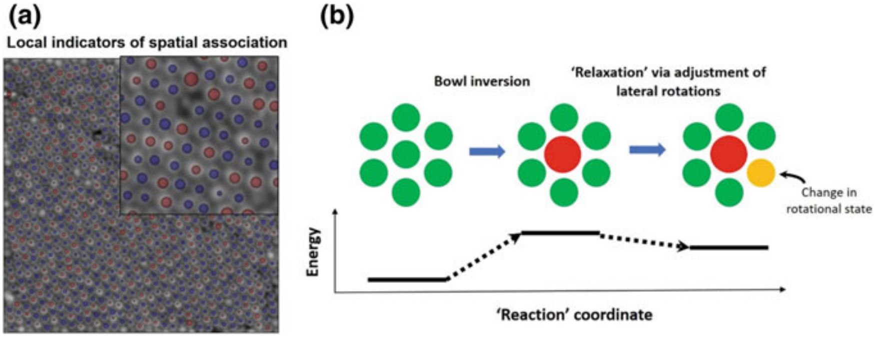
(a)
Local indicators of spatial association
(b)
Bowl inversion
'Relaxation' via adjustment of lateral rotations
Change in rotational state
Energy
'Reaction' coordinate

Fig. 5.6 From imaging to physics. **a** Local indicators of spatial associations based on the Moran's I calculated for the first coordination "sphere". **b** Proposed reaction mechanism involving change in molecular rotational state(s) after bowl inversion. Adapted from [23]

a different size of circles reflects different values of the Moran's I across a field of view. Generally, the map in Fig. 5.6a implies a spatial variation in coupling between the two associated order parameters, which could also be sensitive to presence of defects. The average value of Moran's I for the first 'coordination sphere' is 0.310, whereas the average value for correlation of rotational classes with bowl-up and bowl-down molecular conformations are 0.246 and 0.426 respectively. This result can be interpreted as that a bowl-up-to-bowl-down inversion of a molecule that creates an 'additional' molecule in the **D** state requires a larger change in a rotational state of the neighboring molecules in order to compensate for a formation of energetically unfavorable, "extra" bowl-down state (as compared to a reversed, bowl-down-to-bowl-up inversion). Based on these findings, it is possible to propose a two-stage "reaction" mechanism, where in the first stage an excitation of a new bowl-down state elevates the energy of the system, which is then relaxed in the second stage of the proposed reaction through adjustment of rotational states of the nearby molecule(s). The latter is associated with the obtained values of Moran's I. The crude value for energy difference between different rotational states induced by bowl inversion, and calculated by estimating Boltzmann factor directly from the ratio of two different correlation values, is ≈0.015 eV.

Unlike previous studies which only considered a bowl inversion process for an *isolated* single molecule, the presented analysis based on synergy of convolutional neural networks, Markov random field model and ab-initio simulations allowed to obtain a deeper knowledge of local interactions that accompany a switching of conformational state of neighboring molecules in the self-assembled layer. This new advanced understanding of local degrees of freedom in the molecular adlayer could lead to a controllable formation of various molecular architectures on surfaces which in turn could result in a realization of multi-level information storage molecular device or systems for molecular level mechanical transduction. As far as future directions of applying machine learning and pattern recognition towards molecular structures are concerned, it should be noted that the physical priors used for input in cNN and MRF could be also in principle extracted from state-of-the-art ab-initio analysis and molecular dynamics (MD) simulations. This could potentially provide more accurate decoding results. In addition, a choice of the optimization parameter in MRF analysis could be optimized in future using a statistical distance approach [33]. Finally, we envision an adaption of deep learning technique called domain-adversarial neural networks [35] which allows to alter theoretically predicted classes based on the observed data. The underlying idea of this approach is that the theoretical and experimental datasets are similar yet different in such a way that traditional neural networks may not capture correct features just from the labeled data.

5.3 Case Study 2. Role of Lattice Strain in Formation of Electron Scattering Patterns in Graphene

5.3.1 Model System and Problem Overview

Graphene, a two-dimensional honeycomb lattice of sp^2-carbon atoms, has attracted enormous research interest mostly due to its unique electronic properties, such as anomalous quantum Hall effect and Klein tunneling, which are a consequence of massless Dirac fermions with linear energy dispersion in the electronic band structure. Presence of a disorder in graphene lattice, such as substitutional dopants, vacancies and adatoms, as well as nanoscale variations in bond lengths (due to in-plane and out-of-plane surface deformations), can have a major impact on the material electronic (and magnetic) structure. Below we describe the study on a relationship between nanoscale modulations of lattice strain and parameters of electron scattering induced by point defects in graphene [34]. This study was performed by applying a combination of sliding window fast Fourier transform, Pearson correlation matrix and canonical correlation analysis to low-bias atomically-resolved scanning probe microscopy images of graphene.

Two graphenic systems on different substrates with different types of defects were chosen. The first system is a topmost graphene layer of graphite peppered with hydrogen-passivated single atomic vacancies (hereafter denoted as G_H) [36]. The second system is a monolayer graphene of reduced graphene oxide on gold (111) substrate (hereafter G_O) covered with oxygen-passivated atomic defects and oxygen functional groups [37]. The representative scanning probe microscopy images for G_H and G_O samples are shown in Fig. 5.7a and b, respectively. Both images were obtained in a low-bias regime ($U_s \leq 0.1$ V) where the current is proportional to the density of states at the Fermi level. The global 2D FFTs for data in Fig. 5.7 a, b shows (see insets) similar reciprocal space patterns for both systems characterized by the two hexagons rotated by 30° with respect to each other, with their lattice constants differ by a factor of $\approx \sqrt{3}$. The outer and inner hexagon is associated with lattice structure and electronic density of states, respectively. Specifically, a formation of the inner hexagon in undistorted graphene is explained as due to the constructive interference between incident and backscattered states from the electron valleys at opposite corner points of the hexagonal Brillouin zone [38, 39]. Owing to the symmetry of graphene lattice, there are three backscattering channels. For point defect that do not preserve the symmetry of graphene lattice as well as in graphene with distorted lattice the scattering probability may be different for each of the three channels. Indeed, it is possible to observe experimentally (in a real space) a fine structure of the electronic superlattice around the defects characterized by the alternation of intensities of the FFT spots in the inner hexagon (see Fig. 5.7a, d and e). The precise origin of such a modulation in graphene electronic superlattice is not yet well understood.

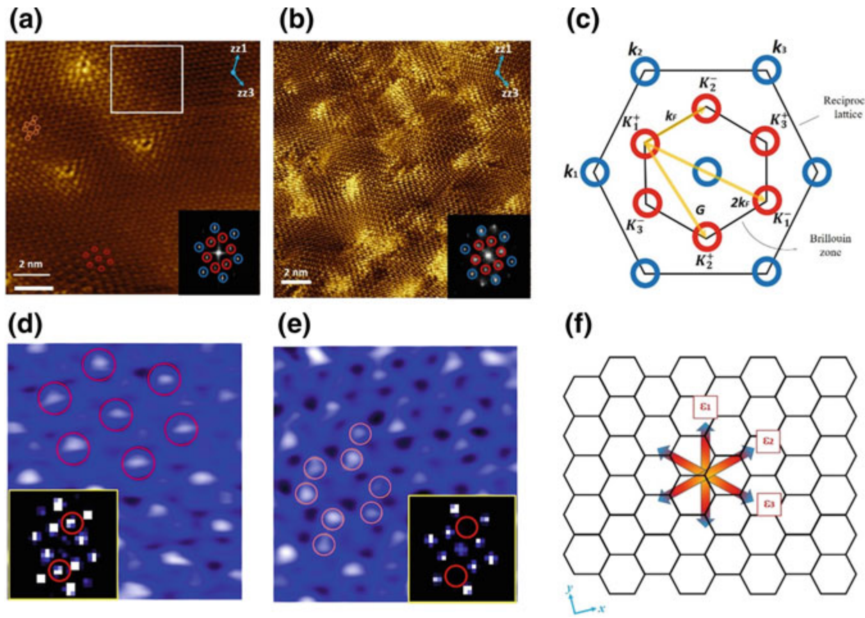

Fig. 5.7 Imaging lattice and electronic structure in graphenic samples. **a** STM image of the top graphene layer of graphite with hydrogen-passivated monoatomic vacancy. $U_s = 100$ mV, $I_{setpoint} = 0.9$ nA. The sliding window used for our analysis is overlaid with the image. **b** Low-bias (2 mV) current-mapping c-AFM image of reduced graphene oxide on gold (111) substrate. The 2D FFT data for both images is shown in the insets. **c** Schematics of graphene electron scattering in the reciprocal space. **d** Hexagonal superperiodic lattice and its 2D FFT. **e** Staggered-dimer-like electronic superlattice and it 2D FFT. Both superlattices are also marked in (**a**). **f** Schematic depiction of 3 different strain components in real space used in our analysis. © IOP Publishing. Reproduced from Ziatdinov et al. [34] with permission. All rights reserved

5.3.2 How to Extract Structural and Electronic Degrees of Freedom Directly from an Image?

Sliding FFT. The goal is to analyze a structure-property relationship in the two graphene systems by studying the correlation between local lattice distortions associated with k_l peaks and electronic features associated with K_e peaks (see Fig. 5.7c). First, a square window of size (w_x, w_y) is created and being shifted across the input image (T_x, T_y) in series of steps x_s and y_s such that the entire image is scanned. At each step, the 2D FFT is computed for the image portion that lies within the window [40]. Hanning window is used to minimize edge effects, as well as a 2 zoom combined and a 2× interpolation function for higher pixel density during the each step of this sliding FFT procedure. The amplitudes and coordinates of the selected peaks are extracted from each 2D FFT image by fitting them with 2D Gaussian distribution, defined as

$$G(q_x, q_y) = A\exp[-(\frac{(q_x - q_x^0)^2 + (q_y - q_y^0)^2}{2\sigma^2})] \tag{5.11}$$

Here A is the peak amplitude, (q_x, q_y) are the Cartesian coordinates of the peak position, and σ is the standard deviation. The unique aspect of graphene is that charge density oscillations are commensurate with the underlying atomic lattice. Therefore, the sliding FFT maps can be used to extract information on both electronic and structural properties of the material. Specifically, the values of intensity and coordinates associated with inner hexagon peaks provide information about intensity of electronic scattering and position of Dirac cone. For the outer hexagon, the coordinates of the peaks from local FFT maps give information about the nanoscale strain distribution in the sample.

The Dirac point drift and electron scattering intensities along the i-th channel are computed as ΔK, $\Delta K_i = (K_i - \overline{K}_i)/\overline{K}_i$ and $I_{K_i} = I\ (K_i^+ \rightarrow K_i^-)$, respectively. To derive a strain map, a strain ε_i is defined as a variation of the lattice vector a_i along the i-th direction, that is, $\varepsilon_i = (a_i - \overline{a}_i)/\overline{a}_i$, where $\overline{a}_i$ is the mean value of the lattice vector in the full image (Fig. 5.7f). It is assumed that for the randomly fluctuating strain fields the mean value of the lattice vector is close to the value of lattice constant in the unperturbed lattice. The a_i is calculated for each step of the sliding FFT algorithm using a standard relation between real space and reciprocal space lattices in graphene. The resolution of spatial maps of the derived structural and electronic descriptors is determined by the size of sliding FFT window and the size of step.

5.3.3 Direct Data Mining of Structure and Electronic Degrees of Freedom in Graphene

Pearson and canonical correlation analysis. Once all the structural and electronic variable of interest are extracted, it becomes possible to explore potential correlations between the corresponding descriptors. Specifically, Pearson correlation matrix analysis and canonical correlation analysis are adopted to explore how formation of various electron interference patterns can be affected by nanoscale variations in the lattice strain. The correlation parameter for each pair of variables x and y is defined as a linear Pearson correlation coefficient,

$$r_{xy} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \overline{y})}} \tag{5.12}$$

where $\overline{x}$ is the mean of x, $\overline{y}$ is the mean of y, and N is a number of scalar observations. While Pearson correlation matrix analysis is a useful technique for studies of bivariate correlations, it is useful to adopt a method called canonical correlation analysis (CCA) that allows grouping the variables in each multivariate dataset such that optimal

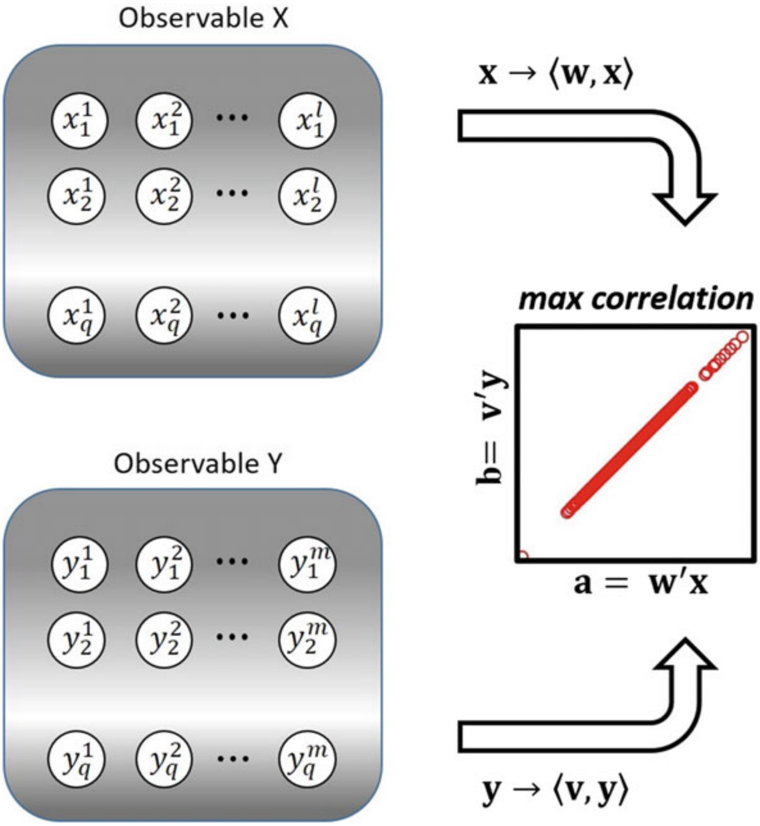

Fig. 5.8 Canonical correlation analysis (CCA). Schematics of CCA workflow. © IOP Publishing. Reproduced from Ziatdinov et al. [34] with permission. All rights reserved

correlation is achieved between two sets [41]. Specifically, CCA solves the problem of finding basis vectors **w** and **v** for two multi-dimensional datasets X and Y such that the correlation between their projections $x \rightarrow \langle w, x \rangle$ and $y \rightarrow \langle v, y \rangle$ onto these basis vectors is maximized. The canonical correlation coefficient ρ is expressed as

$$\rho = max_{w,v} \frac{w' C_{xy} v}{\sqrt{w' C_{xx} w v' C_{yy} v}} \tag{5.13}$$

where $C_{xx,} C_{yy}$ are auto-covariance matrices, and $C_{xy,} C_{yx}$ are cross-covariance matrices of **x** and **y**. The projections $a = w'x$ and $b = v'y$ represent the first pair of canonical variates (Fig. 5.8).

Application to experimental data. The results of correlation matrix and canonical correlation analysis for G_H sample are summarized are summarized in Fig. 5.9a and b, respectively. The canonical correlation coefficient is 0.62 and the associated canonical scores are given by

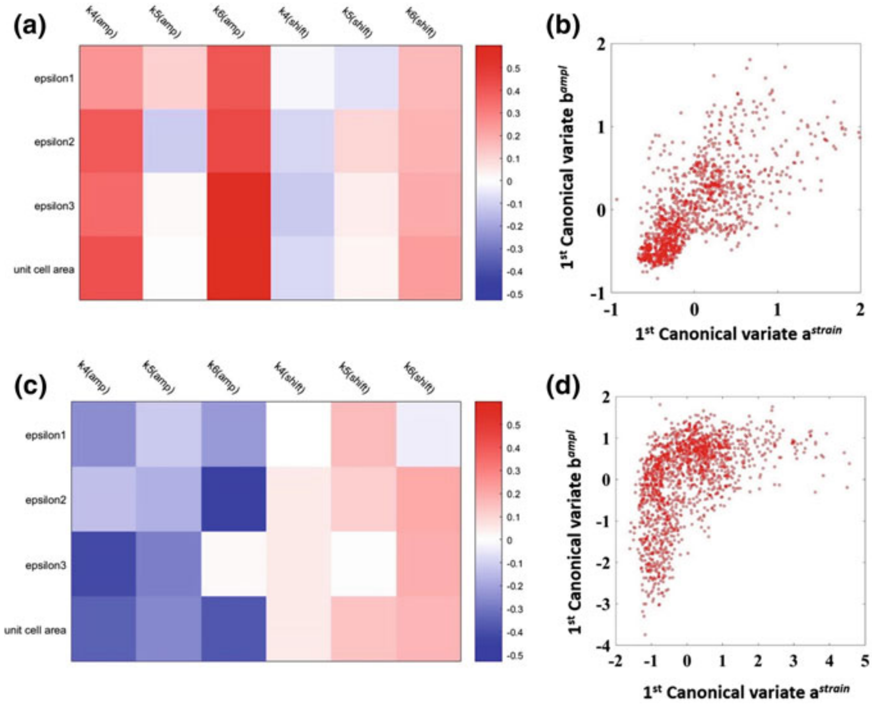

Fig. 5.9 Correlative analysis of graphene structural and electronic degrees of freedom. **a**–**b** Pairwise Pearson correlation matrix (**a**) and plot of the canonical variable scores for the correlation between strain components and scattering intensity for the G_H sample. **c**–**d** Same for G_O sample. © IOP Publishing. Reproduced from Ziatdinov et al. [34] with permission. All rights reserved

$$a_i^{strain} = 0.37(\varepsilon_1)_i + 0.50(\varepsilon_2)_i + 0.36(\varepsilon_3)_i \tag{5.14a}$$

$$b_t^{ampl} = 0.39(I_{k1})_i - 0.33(I_{K2})_i + 0.80(I_{K3})_i \tag{5.14b}$$

where the magnitudes of the coefficients before the variables give the optimal contributions of the individual variables to the corresponding canonical variate. Here the scattering intensities associated with two channel I_{K1} and I_{K2} show a non-negligible positive correlation with strain components in both Pearson correlation matrix and the canonical scores. A dependence of electron scattering intensity on lattice strain for G_H sample can be in principle understood within nearest-neighbor tight-binding model. Specifically, the tight-binding Hamiltonian for graphene monolayer is expressed as [42]

$$H = -\gamma \sum_{\langle i,j \rangle} (a_i^{\dagger} b_j + h.c.) \tag{5.15}$$

where γ is the nearest neighbor hopping parameter, operators $a_i^\dagger\,(b_i^\dagger)$ and $a_i\,(b_i)$ create and annihilate an electron, respectively, at two graphene sublattices, and h.c. stands for the Hermitian conjugate. The density of states $D(E)$ in monolayer graphene is given by

$$D(E) = \frac{|E|}{\pi\sqrt{3}\gamma^2} \tag{5.16}$$

Further, the dependence of the hopping parameter on the bond length can be described in terms of the exponential decay model [43, 44],

$$\gamma \cong \gamma_0 exp(-\tau\varepsilon) \tag{5.17}$$

where τ is typically assigned values between 3 and 4. It follows from (5.16) and (5.17) that the positive correlation between the strain components and the scattering amplitudes in channels I_{K1} and I_{K3} can be explained by enhancement of the density of electronic states available for scattering with increasing the bond length. This also agrees with the first-principles calculations that demonstrated an emergence of new peaks in the density of states near the Fermi level with increasing the bond length [45]. Interestingly, a response of channel I_{K2} to the variations in strain is clearly different from that of channels I_{K1} and I_{K3}. The altered behavior of structure-property relationship for I_{K2} channel becomes even clearer by looking at canonical variates in (5.14) that show a negative sign of a coefficient in front of I_{K2}. Such altered behavior in one of the scattering channels may lead to the formation of observed fine structure of electronic superlattice, namely, coexistence of staggered dimer-like and hexagonal superlattices.

Unlike the G_H sample, the oxidized graphene layer G_O shows a negative correlation between lattice strain and scattering intensities for all the scattering channels (Fig. 5.9c and d). The CCA canonical variates for GO sample are

$$a_i^{strain} = 0.31(\varepsilon_1)_i + 0.73(\varepsilon_2)_i + 0.32(\varepsilon_3)_i \tag{5.18a}$$

$$b_t^{ampl} = -0.37(I_{k1})_i - 0.41(I_{K2})_i + 0.80(I_{K3})_i \tag{5.18b}$$

with CCA coefficient equal to 0.50. This indicates a presence of *apparent* lattice contraction in the 2D-projected SPM images caused by out-of-plane "rippling" of graphene lattice in the presence of oxygen functional groups on the surface. In addition to out-of-lane surface deformations [46, 47], the attached oxygen functional groups also cause an expansion of the lattice constant in their vicinity [47, 48] which, in this case, is hidden from our view "under" the rippled regions in the image. Similar to the analysis for G_H sample, the correlation between scattering intensity and lattice stain can be explained based on the nearest neighbor tight binding model, where an increased lattice constant under the curved regions leads to enhanced density of electronic states available for scattering. Interestingly, the ε_2 strain component and the scattering intensity in I_{K3} channel display the strongest contribution to their respective canonical variates indicating non-uniform strain-scattering relation at the nanoscale and their potential connection to the variations in the electronic superlattice patterns in G_O sample.

We now comment on a character of Dirac point shift. It is worth recalling that for the underformed graphene lattice the positions of electron scattering maxima ("Dirac valleys") are located at the corners of graphene Brillouin zone. Interestingly, however, only relatively small correlation between positions of Dirac point and lattice strain was found in both G_O and G_H systems. Since the position of the Brillouin zone corners in both deformed and non-deformed graphene are given by a direct linear transformation of the reciprocal lattice vectors, these results suggest that in the deformed graphene lattice the locations of electron scattering maxima do not necessarily coincide with the corners of the (new) Brillouin zone.

To summarize this section, we have demonstrated a successful approach for analyzing structure-property relationship at the nanoscale using a combination of sliding window fast Fourier transform, Pearson correlation matrix and canonical correlation analysis. A peculiar connection between variations in coupling between lattice strain components and intensity of electron scattering was found that could explain an emergence of the experimentally observed fine structure in the electronic superlattice. It is worth noting that the analysis demonstrated here was mainly limited to linear structure-property-relationships. One potential way to overcome this limitation would be to use kernelized version of CCA [49] with physics based kernels. For example, one may construct a certain function $F(x, z)$, where z is a physical parameter that determines a non-linearity, so that the resultant kernel $K(x, y) = F'(x, z) * F(y, z)$ will approximate a linear behavior in a limit of very small z, whereas for large values of z it will approximate a non-linear behavior.

5.4 Case Study 3. Correlative Analysis in Multi-mode Imaging of Strongly Correlated Electron Systems

5.4.1 Model System and Problem Overview

In our last case study, a structure-property relationship is analyzed for the case where structural and electronic information are obtained through two separate channels of scanning tunneling microscopy experiment on iron-based strongly correlated electronic system. This type of materials display a rich variety of complex physical phenomena including an unconventional superconductivity [6]. The Au-doped $BaFe_2As_2$ compound was selected which, at the dopant level of $\sim$1%, presides in the spin-density wave (SDW) regime below $T_N \approx 110$ K [50, 51]. At increased concentration of Au-dopants, the magnetic interactions associated with SDW phase become suppressed and the system turns into a superconductor ($T_c \approx 4$ K) at $\sim$3% [51]. The interactions present in SDW regime may thus provide important clues about mechanisms behind emergence of superconductivity in FeAs-based systems. Of specific interest is a region of cleaved $Ba(Fe_xAu_{1-x})_2As_2$ surface (Fig. 5.10) that

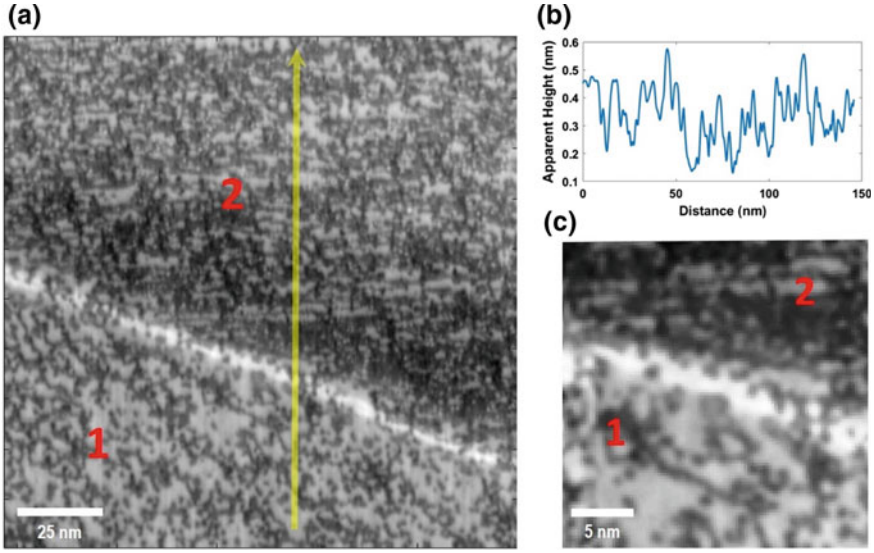

Fig. 5.10 Scanning tunneling microscopy data on Au-doped $\mathrm{BaFe_2}\mathbf{As}_2$. **a** STM topographic image showing domain-like structure where two different (as seemingly appears from the topography) domains are denoted as 1 and 2. **b** Topographic profile along yellow line in (**a**). **c** Smaller topographic area of a 2-domain-like structure that was used for scanning tunneling spectroscopy (STS measurements)

seemingly shows a presence of two different domains-like structures (marked 1 and 2 in Fig. 5.10a) separated by a bright linear topographic feature. Manual inspection of conductance maps at several different values of energy from such region demonstrates a spatially inhomogeneous electronic structure across the FOV, as well as potentially different *dominant* forms of electronic behavior in domain 1 and domain 2, but does not allow an accurate mapping of these electronic behaviors.

5.4.2 How to Obtain Physically Meaningful Endmembers from Hyperspectral Tunneling Conductance Data?

To gain a deeper insight into the types and spatial distribution of different electronic behaviors in this 2-domain-like structure, the non-negative matrix factorization (NMF) method is applied to a scanning tunneling spectroscopy (STS) dataset of dimensions 100 × 100 × 400 pixels recorded over a portion of the structure of interest (Fig. 5.10c). NMF solves the problem of decomposing the input data represented by matrix X of size $m \times n$, where m is the number of features (m = 512 for this dataset) and n is the number of samples (n = 10,000 for this dataset), into two non-negative factors W and H such that $X \approx WH$ [52]. The k columns of W are interpreted as source signals (endmembers) whereas H defines the loading maps (abundance). Due to the non-negativity constraint, NMF can be applied to problems involving finding $k \ll min(m, n)$ physically-meaningful source signals (i.e. physically-defined phases) from the input data, such that all the data can be explained as a mixture of the k basic phases. The NMF can be formally defined as a constrained optimization problem, which can be written, according to Li and Ngom, in a general form as [53]

$$min_{W,H} f(W, H) = \frac{1}{2}\|X - WH\|_F^2 + \\ + \sum_{i=1}^{k}(\alpha_1 \|w_i\|_1 + \frac{\alpha_2}{2}\|w_i\|_2^2) + \sum_{i=1}^{k}(\lambda_1 \|h_i\|_1 + \frac{\lambda_2}{2}\|h_i\|_2^2) \tag{5.19}$$

subject to $W \geq 0,\ H \geq 0$ and where $\|\bullet\|_F$ is the Frobenius norm, w_i and h_i are the i-th columns of W and H, respectively, α_1 and α_2 are regularization parameter for sparsity and smoothness, respectively, for the endmembers domain, while λ_1 and λ_2 control sparsity and smoothness, respectively, for the loading maps (abundancies) domain.

The results on NMF based decomposition into 3 components are shown in Fig. 5.11 (no new information was obtained by increasing a number of compo-

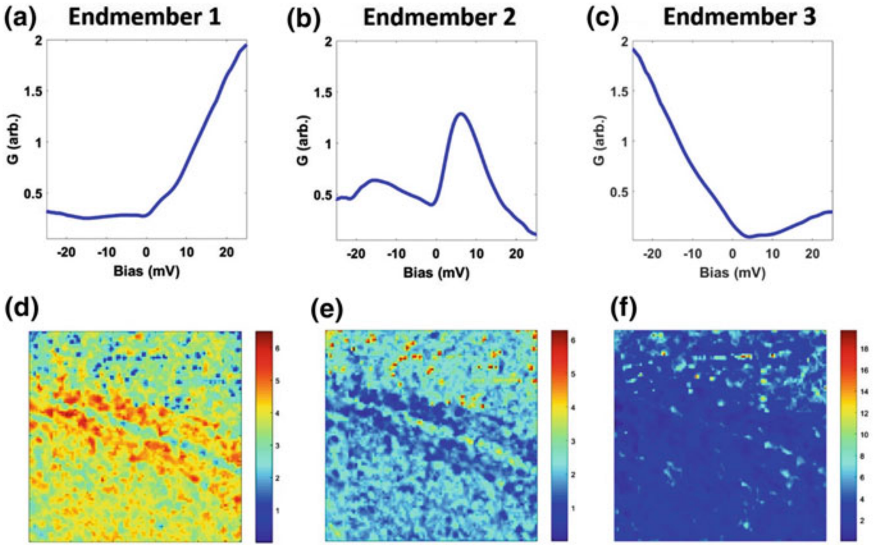

Fig. 5.11 Extraction of electronic descriptors from STS dataset on Au-doped $BaFe_2As_2$. **a**–**c** NMF decomposed spectral endmembers. **d**–**f** corresponding loading maps (the same region as shown Fig. 5.10c)

nents). Spatial weight of endmember 1 is mainly concentrated within the domain 1 (Fig. 5.11a, d). The corresponding spectral curve shows a reduced density of states at the negative energies that agrees with the theoretical and angle-resolved photoemission spectroscopy evidence for partial gap opening just below the Fermi level in the SDW regime. The endmember 2 spectral curve shows a well-defined asymmetric double-peak structure (Fig. 5.11b). Analysis of loading maps for this component (Fig. 5.11e) reveals that this type of electronic behavior is constrained to point-like features on the surface. Furthermore, these features are predominantly located in the domain 2. Therefore, they are associated with a presence of dopant states. Interestingly, the asymmetric double peak structure observed in the endmember 2 is in a good qualitative agreement with non-magnetic dopant-induced double resonance peak model in SDW phase. Analysis of loading maps for the endmember 3 suggests that it may also originate from some form of localized disorder (Fig. 5.11f). These point-like defect states are located mainly in the domain 2 although there is a diluted concentration of defect in the domain 1 as well. While there is no well-defined peak in the density of states associated with this type of defect in the low energy range of interest (Fig. 5.11c), an alternation of the local density of states around the Fermi level was still observed as compared to SDW phase (endmember 1). It is therefore concluded that endmember 2 and endmember 3 describe two distinct types of point defect/dopants that have different structural and/or chemical origin. Thus, the characteristic difference between two domain-like structures 1 and 2 is that there is a significant accumulation of point "impurities"/dopants in only one of those domains. This effectively can be interpreted a peculiar transition between "heavily-doped" and "lightly-doped" regions on the surface.

Correlative analysis of surface geometry and electronic structure. We next proceed to correlative analysis of STM topographic data and loading maps of NMF electronic components. Since no atomic lattice was resolved for this surface region, a correlative analysis is carried out in a pixel-by-pixel fashion. The global Moran's I analysis for the NMF components 1, 2, and 3 and topography returns the values of −0.472, 0.351, and −0.282, respectively. In order to derive physics from such type of structure-property cross-correlation analysis it is crucial to be able to visualize directly those regions on the surface that show higher/lower correlation values. For this purpose, the local indicators of spatial associations described earlier for analysis of correlation between different molecular orders are employed. In addition the results of local Moran's analysis can be mapped on to quadrants resulting into what is known as Moran's Q maps. The local Moran's I and Moran's Q maps are shown in Fig. 5.12. The analysis of Moran's I correlation maps for the endmember 1 (SDW) and endmember 2 (localized defect state) captures a well-defined point-like regions of positive and negative correlation, respectively, which indicates a relatively large number of impurities (characterized by localized states) residing in local dips of the topographic map (Fig. 5.12a, b). The correlative analysis also offers a unique chance to get an insight into 'coupling' of different electronic orders to the boundary between domain 1 and domain 2 (linear bright topographic feature in Fig. 5.10a, c). Particularly, a peculiar depletion of SDW phase along the domains boundary was found that

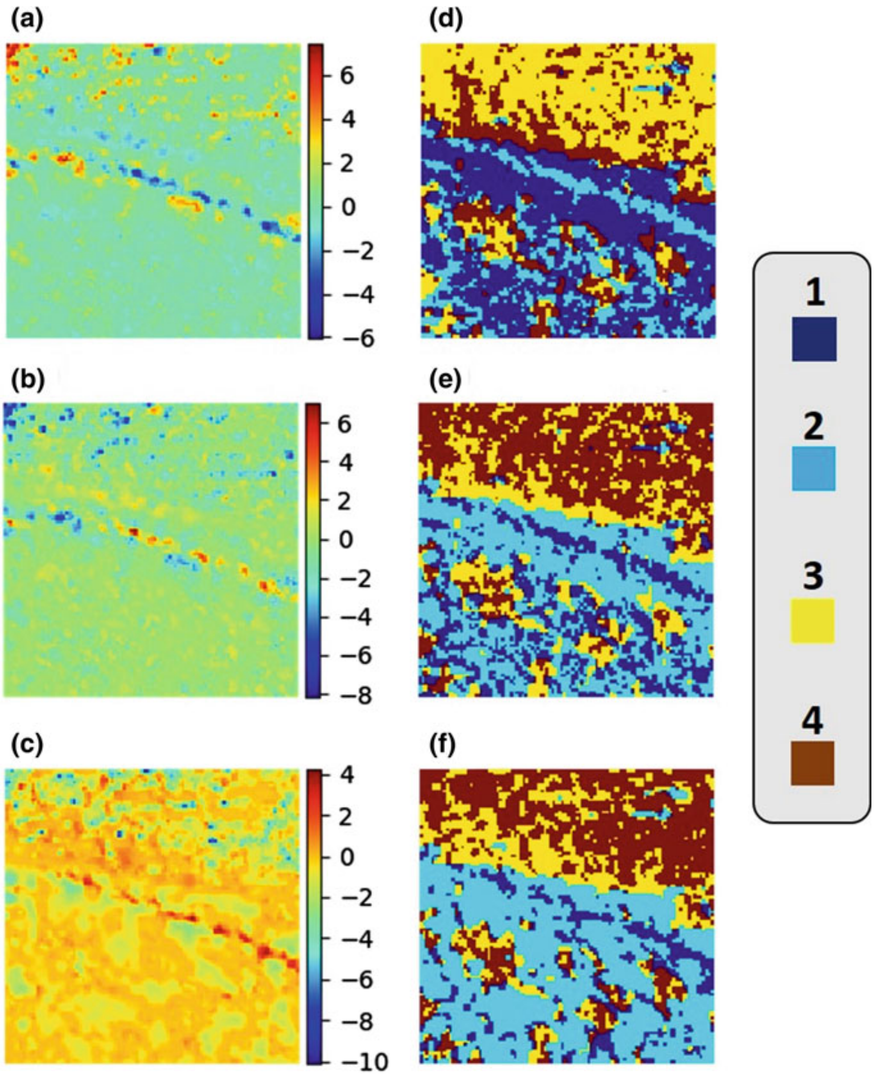

Fig. 5.12 Local indicators of spatial association. Local bivariate Moran's I and Moran's Q (quadrants) calculated for relationship between topographic data (apparent height) and endmember 1 (**a**, **d**); endmember 2 (**b**, **e**); endmember 3 (**c**, **f**). Quadrants legend: Q = 1—positive correlation between high x and high neighboring y's; Q = 2—negative, low x and high neighboring y's; Q = 3—positive, low x and low neighboring y's; Q = 4—negative, high x and low neighboring y's

is clearly evident from appearance of a well-defined linear Q = 2 feature in Moran's Q maps (Fig. 5.12d), that is, a region in which low local values of SDW component correspond to high local values of apparent height (topography). Meanwhile, a presence of Q = 1 features in Fig. 5.12e, f indicates an aggregation of localized states associated with both types of structural/chemical disorder (i.e., NMF components 2 and 3) along the extended regions of domain boundary. These chain-like formations of defects potentially suggest an existence of different conduction mechanism along the quasi-1D domain boundary.

To summarize this last section, we have developed a framework for an automated analysis of multimodal imaging data, and illustrated our approach on scanning tunneling microscopy/spectroscopy datasets from iron-based strongly correlated electronic systems. A peculiar domain-like structure characterized by presence/absence of significant dopants accumulation in different domains and non-trivial depletion of spin density wave state along the domain boundary were discovered. Furthermore, the analysis showed an interesting aggregation of impurities along the certain extended regions of the boundary implying a potential for realizing a special type of domain boundary conductivity under certain conditions. Going forward, we foresee an application of the outlined approach to analysis of different modes of electron-boson interaction in high-T_c superconductors as well in other strongly correlated materials of interest. Finally, we emphasize that this approach is universal, and can be easily applied to other forms of multimodal imaging techniques, such as STEM-EELS [54] or multimodal X-ray imaging techniques [55].

5.5 Overall Conclusion and Outlook

Overall, the incorporation of the advanced data analytics and machine learning approaches in functional and structural imaging coupled with computational-based simulations could lead to breakthroughs in the rate and quality of materials discoveries. The use of these approaches would enable full information retrieval and exploration of structure-property relationship in structural and functional imaging on atomic level in an automated fashion. This, in turn, would allow a creation of libraries of atomic configurations and associated properties. This information can be then directly linked to theoretical simulations to enable effective exploration of material behaviors and properties. Furthermore, knowledge of extant defect configurations in solids can significantly narrow the range of atomic configurations to be probed from the first-principles, thus potentially solving an issue with exponential growth of number of possible configurations with system size. These approaches can further be used to build experimental databases across imaging facilities nationwide (as well as worldwide), establish links to X-ray, neutron and other structural databases, and enable immediate in-line interpretation of information flows from microscopes, X-Ray and neutron facilities and simulations.

Acknowledgements This research was sponsored by the Division of Materials Sciences and Engineering, Office of Science, Basic Energy Sciences, US Department of Energy (MZ and SVK). Part of research was conducted at the Center for Nanophase Materials Sciences, which is a DOE Office of Science User Facility.

References

1. T. Le, V.C. Epa, F.R. Burden, D.A. Winkler, Chem. Rev. **112**(5), 2889–2919 (2012)
2. O. Isayev, D. Fourches, E.N. Muratov, C. Oses, K. Rasch, A. Tropsha, S. Curtarolo, Chem. Mater. **27**(3), 735–743 (2015)
3. G. Xu, J. Wen, C. Stock, P.M. Gehring, Nat. Mater. **7**(7), 562–566 (2008)
4. K. Gofryk, M. Pan, C. Cantoni, B. Saparov, J.E. Mitchell, A.S. Sefat, Phys. Rev. Lett. **112**(4), 047005 (2014)
5. O.M. Auslaender, L. Luan, E.W.J. Straver, J.E. Hoffman, N.C. Koshnick, E. Zeldov, D.A. Bonn, R. Liang, W.N. Hardy, K.A. Moler, Nat. Phys. **5**(1), 35–39 (2009)
6. I. Zeljkovic, J.E. Hoffman, Phys. Chem. Chem. Phys. **15**(32), 13462–13478 (2013)
7. M. Daeumling, J.M. Seuntjens, D.C. Larbalestier, Nature **346**(6282), 332–335 (1990)
8. Y. Zhang, V.W. Brar, C. Girit, A. Zettl, M.F. Crommie, Nat. Phys. **5**(10), 722–726 (2009)
9. J. Martin, N. Akerman, G. Ulbricht, T. Lohmann, J.H. Smet, K. von Klitzing, A. Yacoby, Nat. Phys. **4**(2), 144–148 (2008)
10. K.K. Gomes, A.N. Pasupathy, A. Pushp, S. Ono, Y. Ando, A. Yazdani, Nature **447**(7144), 569–572 (2007)
11. E. Dagotto, Science **309**(5732), 257 (2005)
12. S.V. Kalinin, S.J. Pennycook, Nature **515** (2014)
13. S.V. Kalinin, B.G. Sumpter, R.K. Archibald, Nat. Mater. **14**(10), 973–980 (2015)
14. D.G. de Oteyza, P. Gorman, Y.-C. Chen, S. Wickenburg, A. Riss, D.J. Mowbray, G. Etkin, Z. Pedramrazi, H.-Z. Tsai, A. Rubio, M.F. Crommie, F.R. Fischer, Science (2013)
15. Y. Wang, D. Wong, A.V. Shytov, V.W. Brar, S. Choi, Q. Wu, H.-Z. Tsai, W. Regan, A. Zettl, R.K. Kawakami, S.G. Louie, L.S. Levitov, M.F. Crommie, Science (2013)
16. C.-L. Jia, S.-B. Mi, K. Urban, I. Vrejoiu, M. Alexe, D. Hesse, Nat. Mater. **7**(1), 57–61 (2008)
17. H.J. Chang, S.V. Kalinin, A.N. Morozovska, M. Huijben, Y.-H. Chu, P. Yu, R. Ramesh, E.A. Eliseev, G.S. Svechnikov, S.J. Pennycook, A.Y. Borisevich, Adv. Mater. **23**(21), 2474–2479 (2011)
18. A. Borisevich, O.S. Ovchinnikov, H.J. Chang, M.P. Oxley, P. Yu, J. Seidel, E.A. Eliseev, A.N. Morozovska, R. Ramesh, S.J. Pennycook, S.V. Kalinin, ACS Nano **4**(10), 6071–6079 (2010)
19. Y.-M. Kim, J. He, M.D. Biegalski, H. Ambaye, V. Lauter, H.M. Christen, S.T. Pantelides, S.J. Pennycook, S.V. Kalinin, A.Y. Borisevich, Nat. Mater. **11**(10), 888–894 (2012)
20. W.J. Kaiser (ed.), *Scanning Tunneling Microscopy* (Academic Press, San Diego, 1993), p. ii
21. H. Sakurai, T. Daiko, T. Hirao, Science **301**(5641), 1878 (2003)
22. S. Fujii, M. Ziatdinov, S. Higashibayashi, H. Sakurai, M. Kiguchi, J. Am. Chem. Soc. **138**(37), 12142–12149 (2016)
23. M. Ziatdinov, A. Maksov, S.V. Kalinin, npj Computational Materials **3**, 31 (2017)
24. S. Jesse, S.V. Kalinin, Nanotechnology **20**(8), 085714 (2009)
25. I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning* (MIT Press, 2016)
26. D. Stutz, *Seminar Report* (RWTH Aachen University, 2014)
27. G.R. Cross, A.K. Jain, IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-5**(1), 25–39 (1983)
28. M. Schmidt, http://www.cs.ubc.ca/~schmidtm/Software/UGM.html (2007)
29. R. Jaafar, C.A. Pignedoli, G. Bussi, K. Aït-Mansour, O. Groening, T. Amaya, T. Hirao, R. Fasel, P. Ruffieux, J. Am. Chem. Soc. **136**(39), 13666–13671 (2014)
30. H. Amara, S. Latil, V. Meunier, P. Lambin, J.C. Charlier, Phys. Rev. B **76**(11), 115423 (2007)
31. A.A. El-Barbary, R.H. Telling, C.P. Ewels, M.I. Heggie, P.R. Briddon, Phys. Rev. B **68**(14), 144107 (2003)
32. L. Anselin, Geogr. Anal. **27**(2), 93–115 (1995)
33. L. Vlcek, A.A. Chialvo, J. Chem. Phys. **143**(14), 144110 (2015)
34. M. Ziatdinov et al., Nanotechnology **27**, 495703 (2016)
35. Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, *ArXiv e-prints*, vol. 1505 (2015)
36. M. Ziatdinov, S. Fujii, K. Kusakabe, M. Kiguchi, T. Mori, T. Enoki, Phys. Rev. B **89**(15), 155405 (2014)

37. S. Fujii, T. Enoki, ACS Nano **7**(12), 11190–11199 (2013)
38. P. Ruffieux, M. Melle-Franco, O. Gröning, M. Bielmann, F. Zerbetto, P. Gröning, Phys. Rev. B **71**(15), 153403 (2005)
39. K.-I. Sakai, K. Takai, K.-I. Fukui, T. Nakanishi, T. Enoki, Phys. Rev. B **81**(23), 235417 (2010)
40. R.K. Vasudevan, A. Belianinov, A.G. Gianfrancesco, A.P. Baddorf, A. Tselev, S.V. Kalinin, S. Jesse, Appl. Phys. Lett. **106**(9), 091601 (2015)
41. W.J. Krzanowski, *Principles of Multivariate Analysis: A User's Perspective* (Oxford University Press, Inc., 1988)
42. P.R. Wallace, Phys. Rev. **71**(9), 622–634 (1947)
43. V.M. Pereira, A.H. Castro Neto, N.M.R. Peres, Phys. Rev. B **80**(4), 045401 (2009)
44. R.M. Ribeiro, M.P. Vitor, N.M.R. Peres, P.R. Briddon, A.H.C. Neto, New J. Phys. **11**(11), 115002 (2009)
45. V.J. Surya, K. Iyakutti, H. Mizuseki, Y. Kawazoe, Comput. Mater. Sci. **65**, 144–148 (2012)
46. S. Fujii, T. Enoki, J. Am. Chem. Soc. **132**(29), 10034–10041 (2010)
47. V.V. Shunaev, O.E. Glukhova, J. Phys. Chem. C **120**(7), 4145–4149 (2016)
48. J. Ito, J. Nakamura, A. Natori, J. Appl. Phys. **103**(11), 113712 (2008)
49. K. Fukumizu, F.R. Bach, A. Gretton, J. Mach. Learn. Res. **8**, 361–383 (2007)
50. M. Ziatdinov, A. Maksov, L. Li, A.S. Sefat, P. Maksymovych, S.V. Kalinin, Nanotechnology **27**(47), 475706 (2016)
51. L. Li, H. Cao, M.A. McGuire, J.S. Kim, G.R. Stewart, A.S. Sefat, Phys. Rev. B **92**(9), 094504 (2015)
52. D.D. Lee, H.S. Seung, Nature **401**(6755), 788–791 (1999)
53. Y. Li, A. Ngom, Source Code Biol. Med. **8**(1), 10 (2013)
54. M. Varela, J. Gazquez, S.J. Pennycook, MRS Bull. **37**(1), 29–35 (2012)
55. O. Bunk, M. Bech, T.H. Jensen, R. Feidenhans'l, T. Binderup, A. Menzel, F. Pfeiffer, New J. Phys. **11**(12), 123016 (2009)