

## Non-negative matrix factorization analysis of spatially-resolved photoemission spectra for epitaxially grown graphene on SiC

Masaki Imamura & Kazutoshi Takahashi

**To cite this article:** Masaki Imamura & Kazutoshi Takahashi (19 Jun 2026): Non-negative matrix factorization analysis of spatially-resolved photoemission spectra for epitaxially grown graphene on SiC, Science and Technology of Advanced Materials: Methods, DOI: [10.1080/27660400.2026.2688747](https://doi.org/10.1080/27660400.2026.2688747)

**To link to this article:** <https://doi.org/10.1080/27660400.2026.2688747>



© 2026 The Author(s). Published by National Institute for Materials Science in partnership with Taylor & Francis Group



[View supplementary material](#)



Accepted author version posted online: 19 Jun 2026.



[Submit your article to this journal](#)



Article views: 4



[View related articles](#)



[View Crossmark data](#)

**Publisher:** Taylor & Francis & The Author(s). Published by National Institute for Materials Science in partnership with Taylor & Francis Group

**Journal:** *Science and Technology of Advanced Materials: Methods*

**DOI:** 10.1080/27660400.2026.2688747

## Non-negative matrix factorization analysis of spatially-resolved photoemission spectra for epitaxially grown graphene on SiC

Masaki Imamura\* and Kazutoshi Takahashi

*Synchrotron Light Application Center, Saga University, 1 Honjo, Saga 840-8502, Japan*

\*e-mail: mimamura@cc.saga-u.ac.jp

**Abstract:** Spatially-resolved ARPES is a powerful tool for probing local electronic structures in low-dimensional materials, but its analysis becomes challenging for spatially inhomogeneous samples due to spectral variations, feature overlap, and minor shifts. Here, we present a practical framework based on non-negative matrix factorization (NMF), which decomposes ARPES spectra into physically interpretable components without relying on prior assumptions. Visualizing the activation matrix as spatial heatmaps reveals latent spectral structures and provides an intuitive map of how individual components are distributed, enabling identification of the domains and local electronic variations. We validate this framework using epitaxial graphene on SiC, demonstrating its ability to quantitatively disentangle spectral features associated with layer thickness, step structures, and growth conditions. This study establishes the NMF-based framework as a scalable and robust tool for managing large-scale datasets and assessing electronic inhomogeneity in low-dimensional materials.

**Keyword:** photoemission spectroscopy, non-negative matrix factorization, graphene, machine learning, data-driven analysis

### 1. Introduction

Angle-resolved photoemission spectroscopy (ARPES) is a powerful tool for investigating the electronic states of condensed materials. The advancement of instrumentation has enabled ARPES to reveal various parameter spaces. Deflection-type electron lenses and momentum microscopes enable efficient mapping of in-plane momentum space, while photon-energy-dependent ARPES provides access to  $k_z$  dispersion. Together, these modern ARPES techniques enable comprehensive electronic-structure analysis across  $k_x$ ,  $k_y$ , and  $k_z$ . Furthermore, measurements incorporating polarization-dependent excitation, spin detector, and pump-probe techniques with ultra-short pulse laser have led to the acquisition of high-dimensional spectral data in spin and time domain [1–4]. Spatially-resolved ARPES, which is the focus of the present study, enables the observation of local electronic structures in materials by employing highly focused synchrotron or laser light as the excitation source, and is commonly referred to as micro- or nano-ARPES[5–9]. Recent advances in instrumentation have enabled the reduction of excitation spot size to the micro- or nanometer scale, along with a significant increase in source brightness that ensures

sufficient signal intensity even with such additional focusing optics. As a result, covering the same sample area now requires a much larger number of measurement points. While the improved spatial resolution facilitates the detection of fine local environments and contributes to more informative ARPES measurements, it inevitably results in a substantial increase in data volume. Consequently, the analysis of ARPES spectra becomes increasingly complex and labor-intensive. Moreover, conventional analysis procedures typically involve constructing spatial maps by integrating signal intensities within manually defined energy and momentum windows. These maps are used to visualize chemical state variations in core-level spectra or spatial modulations in band dispersion in valence-band spectra. Subsequent steps often rely on manual selection of regions of interest and detailed inspection of representative spectra[10–14]. However, this workflow is heavily dependent on the expertise and prior knowledge, and requires multiple subjective decisions, such as the choice of integration ranges or selection of characteristic features. As a result, the reproducibility and objectivity of the analysis can be compromised, especially in datasets where spectral features are weak, overlapping, or spatially fluctuating in complex ways. In view of these challenges, a data-driven and unbiased analytical framework that enables efficient extraction of meaningful components without prior assumptions about the data structure is desirable.

Recently, remarkable advances in machine learning and artificial intelligence have led to the widespread application across various scientific fields. In particular, the increasing availability of open-source libraries and analytical tools many of which provide user-friendly interfaces and comprehensive documentation has significantly lowered the barrier to entry. As a result, even researchers without a background in data science can now implement and apply these techniques with relative ease. In the field of materials science, for instance, machine learning has been employed in the development of autonomous high-throughput experimental platforms utilizing robotics, as well as in the prediction and discovery of novel material phases[15–19]. In photoelectron spectroscopy, which is the focus of this study, a range of machine learning-based approaches have been reported, including the clustering of measured spatially-resolved ARPES spectra [20], Gaussian process for efficient exploration of experimental conditions [21], Bayesian spectroscopy of XPS spectra [22], and neural network-based methods for denoising and detecting spectral features[23–26]. Among these, dimensionality reduction provides a robust framework for condensing and interpreting complex experimental datasets in a comprehensive manner. Applying dimensionality reduction to experimental datasets not only reduces data size but also provides interpretable features or latent variables that can explain the observed data without the need to manually inspect each individual spectrum. This enables researchers to identify underlying correlations or latent patterns within the dataset, offering an intuitive and comprehensive understanding of the data. Such capabilities are especially valuable when dealing with large, heterogeneous datasets, as is often the case in spatially-resolved ARPES measurements involving numerous spectra.

Non-negative Matrix Factorization (NMF) is a type of dimensionality reduction technique that decomposes a non-negative data matrix  $X$  into two smaller non-negative matrices, the basis matrix  $W$  and the activation matrix  $H$  (Fig. 1) [27,28]. The basis matrix  $W$  represents common patterns in the dataset, while the activation matrix  $H$  corresponds to the weight coefficient for each  $W$  to reconstruct  $X$ . This factorization allows the original data matrix to be approximated as the product  $X \approx WH$ . A significant advantage of NMF lies in the intuitive interpretability of the results. Because all elements in both  $W$  and  $H$  are constrained to be non-negative, the resulting components

correspond to interpretable quantities. These characteristics make NMF not only valuable for general applications such as image and signal analysis, but also highly effective in material science, where the interpretability and physical relevance of extracted components are of critical importance[29–32].

We previously performed spatially-resolved ARPES measurements on a graphene containing mixed monolayer- (ML) and bilayer- (BL) derived domains, applied NMF to the resulting dataset, and subsequently conducted clustering using k-means method[33]. Representative spectra were selected from each cluster, and the spectra were interpreted based on NMF results. However, since k-means is a hard clustering method, which assigns every data point to one of the clusters without exception, it does not account for the spectral variability within the same label. In practice, spatially-resolved ARPES spectra often exhibit variations even within a single cluster, and at cluster boundaries, spectra may appear that could reasonably belong to either label. Such cases traditionally require manual inspection of each measurement point, which becomes impractical and prone to oversight when dealing with datasets comprising tens of thousands of points.

In this work, we propose a practical analytical framework based on unsupervised learning using NMF. A spatially-resolved ARPES dataset of graphene containing mixed BL- and trilayer- (TL) derived domains, which exhibits spatial variations in its electronic structure, was decomposed into physically interpretable spectral components without relying on any prior assumptions. To overcome the limitations of k-means clustering, we mapped the intensity of the activation vectors, the weights of each basis vector  $W$  in the spectra at each spatial point, thereby generating spatial distributions that directly reflect the spectral characteristics represented in the dataset. This approach creates spatial distributions that directly reflect spectral characteristics, enabling a more quantitative and efficient interpretation of the ARPES data while preserving physical insight. Analysis of spatial maps of the activation matrix revealed the distinct domain-dependent spectral features associated with local thickness, step structures, and growth conditions. These results demonstrate that NMF serves as a highly effective and versatile tool for the analysis of spatially-resolved ARPES data, enabling domain identification based on the intensity of specific band structures and elucidation of their physical significance. Compared with conventional approaches, this method provides a more direct representation of the underlying electronic structure by explicitly extracting the relevant bands.

## 2. Experiments

As a model system for this study, we selected epitaxial graphene grown on SiC substrates. Graphene exhibits pronounced changes in its Dirac cone structure at six corners of two-dimensional Brillouin zone depending on the number of layers, making it highly suitable for testing the ability of NMF to distinguish subtly varying electronic structures[34–36]. Moreover, graphene grown on SiC typically forms terraces with widths on the order of several hundred micrometers, which are comparable to the spatial resolution of the ARPES measurements used in this work. Graphene was prepared on n-doped 6H-SiC(0001) with a miscut angle of less than  $0.05^\circ$  using a face-to-face method, a well-established and widely used approach for obtaining high-quality epitaxial layers[37,38]. In this method, placing two SiC (0001) substrates face-to-face suppresses Si sublimation and promotes uniform graphene growth. The thickness of graphene layers was controlled with the annealing temperature.

Spatially-resolved ARPES measurement for graphene grown on a SiC substrate was

conducted at Saga university beamline (BL13) in SAGA-LS[39]. An incident photon energy was 40 eV. ARPES spectra were acquired using a grid scan at approximately 400 points on an (x, z) grid covering the SiC substrate approximately 4 mm  $\times$  13 mm. All ARPES spectra were indexed in conjunction with the corresponding measured locations (x and z). During the spatially-resolved measurements, the locations were adjusted with steps of 0.5 mm and 0.2 mm in the x and z direction, respectively. These step sizes were significantly larger than the spot size of the incident light, which was approximately 100  $\mu\text{m}$   $\times$  150  $\mu\text{m}$ . Dirac-cone dispersions originating from BL and TL graphene around the K point were experimentally measured and subsequently used as the dataset. All the coding for the machine learning analysis was carried out with Python, and the scikit-learn library was utilized for NMF[40]. As a preprocessing step, each observed ARPES spectrum was binned to reduce noise and decrease the data size. The photoelectron intensities were normalized to have the same maximum intensity. Then, the preprocessed ARPES spectra were converted into one-dimensional vectors and stacked to convert them into a two-dimensional matrix. Each row of the matrix corresponds to a single ARPES image at a certain measurement point. When applying NMF, the values of W and H in NMF were initialized as non-negative random matrices (init='random'). The decomposition was carried out using the coordinate descent solver (solver='cd') with a maximum of 2000 iterations (max\_iter=2000). All other hyperparameters were set to their scikit-learn defaults. For k = 7, the algorithm converged in approximately 500 iterations; the number of iterations required for convergence increased moderately with increasing k. All computations were performed on a MacBook Pro equipped with an Apple M1 Pro CPU and 32 GB of memory. For the primary binned data matrix (390  $\times$  6396), the computation time was less than 1.5 seconds per run (mean over 10 trials); for the unbinned dataset (390  $\times$  58247), the computation time remained below 10 seconds per run. The stability of the decomposition with respect to other initialization methods and the selection of the number of basis vector k were examined. In the following, columns or rows extracted from the resulting W and H matrices with NMF will be referred to as vectors.

### 3. Results and discussions

Fig. 2 (a) shows the basis vectors W obtained from the NMF. Each image was generated by extracting one column from the resulting basis matrix W and reshaping it into the two-dimensional image with the same dimensions as the original ARPES image. For ease of reference, these individual components are denoted as W0 to W6. The vertical axis represents binding energy, and the horizontal axis corresponds to the photoelectron emission angle. In ARPES analysis, the horizontal axis is typically calibrated to actual emission angles or momentum values. However, in this study, we treated it as a dimensionless quantity. The obtained basis vectors W reflect the spectral features associated with BL and TL graphene. [34–36] Variation in the central position of the Dirac cone is observed among the basis vectors: W0, W1, W3, and W4 are aligned at the same angle, while W2 and W6 are shifted to higher, and W5 to lower angles. A dotted line is drawn at the same angular coordinate in each panel and is included as a visual guide. In spatially resolved ARPES, the sample is scanned while the focal distances, kinetic energy, and angular windows are kept fixed, ensuring that all spectra share a common angular axis defined by the measurement geometry. A systematic shift of the Dirac cone was nevertheless observed for spectra acquired near the edge region, which we attribute to local electric fields at the edge. [33] The basis vectors associated with this effect (W2, W5, W6) are consistent with such field-induced spectral shifts, as

discussed in detail in Supplementary Information Fig. S4.

In NMF, the number of basis vectors  $k$  and the initialization method are key hyperparameters that influence the decomposition results. Since convergence to a global optimum is not guaranteed and the solution may be trapped in local minima depending on the initialization, it is necessary to evaluate the robustness of the results in terms of reproducibility, the physical validity of the extracted components, and the stability of the solution. The number of basis vectors was selected based on the elbow in the SSE curve and a review of the extracted basis matrix, as shown in Fig. 2(b). Here, SSE denotes the sum of squared errors between the original spectra and those reproduced from the product of the basis and activation matrices. Since SSE decreases monotonically with increasing  $k$ , no absolute threshold can define the optimal number. Choosing  $k$  based on the SSE elbow retains a degree of trial-and-error, but it offers the advantage of being broadly applicable without additional assumptions about the data distribution. The decision of the number of basis vectors requires careful comparison between the raw data and the extracted components while ensuring that the resulting decomposition remains physically interpretable.

To assess the validity of the selected  $k = 7$ , we examined how the extracted basis and activation vectors change across different values of  $k$  (Supplementary Information S1). When an excessively large number of basis vectors is chosen, NMF attempts to represent minor spectral fluctuations, often yielding multiple nearly identical basis vectors that differ only in noise content, which complicates the overall interpretation of the results. Conversely, at smaller values of  $k$ , distinct spectral components associated with BL and TL graphene tend to be merged into a single basis vector, obscuring physically meaningful distinctions. Initialization dependence was assessed as follows. First, NMF with random matrices was performed multiple times with different random seeds, and the resulting basis vectors were compared to confirm reproducibility. Second, NMF was performed using alternative initialization methods available in scikit-learn, namely 'nndsvd', 'nndsvda', and 'nndsvdar', and the output basis vectors and their SSE values were compared (Supplementary Information S2). Since the response to data sparsity differs across initialization methods, certain methods yielded higher SSE values or resulted in poorer reconstruction of the original spectra and less physically interpretable decompositions than those obtained with random initialization. Therefore, it is important to visually inspect the extracted basis vectors and compare them against the original data regardless of the initialization method employed, and to examine whether the solution has converged to a local minimum. Since similar principal conclusions are obtained across a range of  $k$  values and under various initial conditions,  $k = 7$  is adopted as the practically optimal choice for the present dataset and its physical interpretation. This selection is not intended to represent a uniquely determined optimum in the sense of a formal model-selection criterion, but reflects a comprehensive judgment based on decomposition stability, physical interpretability of the extracted basis vectors.

Yoshinari et al. demonstrated that hierarchical clustering and NMF can play complementary roles in a two-stage analytical workflow in the context of RHEED pattern analysis. [41] In that study, hierarchical clustering was first applied to determine the number of distinct surface phases present in the dataset; Ward's method naturally revealed structural phases through the dendrogram without requiring the number to be fixed in advance. Subsequently, NMF was applied to extract quantitative information specific to each identified phase. This separation of concerns, where clustering guides component number selection and NMF handles quantitative spectral decomposition, represents a valid alternative strategy to the elbow-based approach used in the

present study, and applying a similar two-stage workflow to spatially resolved ARPES data could serve as a more objective cross-check on component number selection. In the present study, the spatial map of labels obtained from k-means clustering clearly reflects the domain structure, and comparison with the activation-vector heatmaps provides valuable insight into the domain structure of the dataset (Fig. S3).

To efficiently extract and interpret structural features from the extensive ARPES dataset, we generated heatmaps of the activation vector  $H$  at each measurement location. These maps visualize the relative contributions of the NMF basis vectors  $W$ , providing a spatially-resolved representation of characteristic spectral components. Since the activation vector  $H$  represents the weighting coefficients for the basis vectors used to reconstruct each experimental ARPES spectrum, its spatial mapping corresponds to the distribution of specific spectral features. Figs. 3(a) and 3(b) show the heatmaps of the activation vectors  $H$  corresponding to weights for  $W_0$  and  $W_3$ , respectively. The vertical and horizontal axes correspond to the  $x$ - and  $z$ -axis values of the manipulator at the measurement locations, respectively. These setting values of manipulator reflect the relative positions of the spectra measured across the graphene grown on the SiC substrate. Strong contributions from both components appear on the right side of the map, where  $W_0$  dominates in the lower region and  $W_3$  in the upper region.

Figure 3(c) shows representative experimental spectra measured at points A–E within the boxed region in Figs. 3(a) and 3(b). All spectra shown here exhibit BL-derived Dirac-cone features of graphene, and similar BL-derived spectra are consistently observed within the regions where  $W_0$  and  $W_3$  are dominant, where variations in spectral intensity distribution are present. At point A, the spectral intensities above and below the Dirac point are nearly balanced. From B to E, the spectral weight on the higher-binding-energy side gradually decreases.  $W_0$  exhibits a two-band dispersion characteristic of bilayer (BL) graphene. In contrast, although no spectra characteristic of monolayer (ML) graphene are experimentally observed in the present dataset,  $W_3$  shows a single linear dispersion in appearance that resembles that of ML graphene. The Dirac cone of BL graphene consists of two bands: an inner band located at the same position as the ML Dirac cone and an additional outer band. Depending on experimental conditions such as angular misalignment and matrix element effects, the relative intensity of these two bands can vary significantly. In particular, when the inner band becomes dominant, the resulting spectrum may exhibit a single linear dispersion similar to that of ML graphene. NMF represents such variations in spectral intensity distribution as a linear combination of multiple basis vectors. Within this framework,  $W_3$  is interpreted as a component that emphasizes the inner-band contribution of BL graphene, and therefore exhibits an ML-like single linear dispersion in appearance. Both  $W_0$  and  $W_3$  are thus interpreted as basis vectors originating from BL graphene. Accordingly,  $W_3$  does not represent an independent ML-derived component, but rather corresponds to an inner band within the BL electronic structure.

Figure 3(d) shows the contribution ratios of  $W_0$  and  $W_3$  required to reconstruct each experimental spectrum in Fig. 3(c). The plotted weights indicate the extent to which  $W_0$  and  $W_3$  contribute to the spectral reconstruction at each point (A–E), thereby quantifying the continuous variation of BL spectra across the spatial region. The contribution of  $W_3$  decreases monotonically from A to E, whereas that of  $W_0$  increases. This crossover in contribution ratios accounts for the spectral evolution observed in Fig. 3(c).  $W_3$  is characterized by a single linear dispersion with a relatively symmetric intensity distribution across the Dirac point, whereas  $W_0$  exhibits two distinct

bands with a pronounced asymmetry, in which the spectral weight on the higher-binding-energy side is weaker. Consequently, as the dominant component shifts from W3 to W0 (from A to E), the reconstructed spectra naturally reproduce the experimentally observed trend, namely, a gradual suppression of intensity on the higher-binding-energy side. The basis vectors obtained via NMF are not meant to directly represent specific physical entities, such as the number of graphene layers. Instead, they are derived so that each measured spectrum can be expressed as a combination of several components shared across the dataset. In the BL region, changes in the contribution ratio of W0 and W3 account for the subtle position-dependent modulations in the spectral intensity along the binding energy axis.

Figures 4(a) and 4(b) show heatmaps of the activation vectors  $H$  corresponding to weights for W1 and W4. Both components are concentrated on the left side of the map. W1, which contains the inner and outer cones of the triple Dirac cone of TL graphene, exhibits moderate-to-high intensity across a broad region, indicating that TL graphene occupies the left half of the sample. In contrast, W4 exhibits a strong contribution in regions where the contribution from W1 is weak. W4 is composed of the outer and center cones of triple Dirac cones of TL graphene below Dirac point together with strong spectral intensity near the Fermi level. This alternating spatial pattern indicates that, although both basis vectors originate from the TL region, the area contains multiple TL domains that exhibit distinct spectral characteristics. Figure 4(c) displays representative spectra at positions F–J. As expected from the activation maps, all spectra exhibit TL-derived structures. At positions F and I, where W1 dominates, the three linear TL branches are clearly observed. At position H, where W4 contributes strongly, as well as at positions G and J, where W1 and W4 exhibit comparable contributions, the spectra exhibit features analogous to Dirac cones of TL graphene at positions slightly displaced from the K-point, similar to those measured when the sample is slightly deviated from K-point. Figure 4(d) plots the weight ratios of W1 and W4 for each spectrum. By adjusting the relative contributions of these basis vectors, NMF successfully reproduces the observed differences in spectral shape across the TL region, demonstrating that NMF effectively decomposes the domain into meaningful subcomponents.

A detailed evaluation of the spatial distribution of activation vectors  $H$  obtained from NMF enables the quantitative characterization of local properties of graphene. As shown in Figs. 3 and 4, BL graphene is predominantly distributed in the right half of the spatial map, whereas TL graphene is mainly observed in the left half in the maps. The number of graphene layers grown by the face-to-face method strongly depends on the temperature and duration of annealing. In the present study, the measured sample was grown on a single substrate under spatially uniform growth time conditions, indicating that the observed variation in layer distribution can be attributed to spatial inhomogeneities in annealing temperature. The spatial mapping of activation vectors by NMF allowed us to isolate and quantitatively reveal the spatial distribution of graphene derived from local temperature differences during sample preparation.

In Figs. 3(c) and 4(c), intensity modulations along the energy axis are observed across different measurement positions. Notably, in Fig. 4(c), spectra recorded at positions G, H, and J exhibit Dirac-cone features that appear slightly displaced relative to those at positions F and I—an observation that could be interpreted as a local shift of the Dirac point energy. Since the buffer layer in graphene on SiC can act as a charge trap, one might consider whether local variations in carrier transfer associated with defects or step structures modulate the Dirac point position. However, as discussed below, the experimental evidence points to an effective geometry origin rather than an



electronic one. Step bunching, arising from differences in step velocities between adjacent steps, leads to the formation of aggregated step structures that can locally distort the surface orientation [42–44]. Consequently, graphene may grow within domains that are inclined relative to the original step direction. The observed modulations in spectral intensity along the energy axis thus reflect the presence of such step-bunched domains, where the local tilt results in an effective angular deviation in photoemission even under fixed measurement geometry. Supporting evidence for this interpretation is provided in Supplementary Information Fig. S5, which demonstrates that similar spectral modulations—including apparent displacement of cone features—can be induced by varying the emission angle relative to the K point. Importantly, while the spectral intensity distribution across the Dirac point changes systematically with angle, the Dirac point energy itself remains unchanged. This indicates that the observed spectral variations, including those seen at points G, H, and J in Fig. 4(c), do not originate from a shift in the Dirac point energy, but rather from angle-dependent intensity modulation. In the present spatially resolved ARPES measurements, the measurement window is fixed near the K point during sample scanning. However, local tilting of the graphene surface due to step bunching effectively shifts the emission angle at each measurement position. The similarity between the angle-dependent behavior shown in Fig. S5 and the spectral modulations observed in Figs. 3(c) and 4(c) therefore supports the interpretation that the apparent spectral variations arise from local tilt effects induced by step bunching, rather than from carrier-induced Dirac point energy shifts.

It should be noted that NMF inherently extracts the dominant spectral features within the chosen energy–momentum window. When a broad window is used, the decomposition is naturally governed by the largest spectral variations in the dataset, such as those arising from differences in layer number, and minor spectral features may be obscured. Extracting such finer variations independently would therefore require a deliberately narrowed energy–momentum window, chosen in accordance with the specific features of interest, together with dedicated fitting procedures or higher-resolution measurements.

Furthermore, the interpretation of the activation maps requires particular care. The spatial maps derived from the NMF activation matrix represent the distribution of specific spectral features at each measurement point, and should not be interpreted as a direct determination of the physical graphene domain structure. The correspondence between the extracted spectral components and physical properties such as layer number and step bunching is established on the basis of well-established ARPES spectroscopy—in particular, the number of bands in the Dirac-cone structure provides a reliable spectroscopic indicator of graphene layer number—and prior knowledge of the sample system. For direct spatial validation of the physical domain structure, independent characterization using complementary techniques such as Raman mapping or atomic force microscopy would be required, and is considered an important direction for future work.

As revealed by our investigation, NMF enables the decomposition of complex spatially-resolved ARPES datasets into physically interpretable basis matrix and its activation matrix, allowing for a quantitative assessment of their spatial contributions. Compared to principal component analysis (PCA), a widely used dimensionality reduction method, NMF offers enhanced interpretability and clear physical meaning of the extracted components. NMF iteratively optimizes basis and coefficient matrices under a non-negativity constraint, making it generally more computationally demanding than PCA for datasets of similar size. Its key advantage is that the resulting decompositions are inherently interpretable and physically meaningful. For the dataset in

this study, computation times were comparable to PCA. However, for datasets comprising tens of thousands of spectra, the computational cost may become significant. While PCA has an advantage in terms of simplicity of hyperparameter tuning, the principal components it yields contain negative values, making physical interpretation of the spectral features challenging. In contrast, the result of NMF is represented as a weighted sum of physically meaningful basis spectra. This characteristic allows for intuitive distinction of spectral variations across different spatial locations. As demonstrated above, the application of NMF to spatially-resolved ARPES data enabled the quantitative and straightforward extraction of multiple growth-related factors of graphene, including inhomogeneous thermal distribution during annealing, electric field effects at the sample edges, and local variations in orientation due to step bunching on the substrate. These results highlight the effectiveness of the NMF-based analysis as a practical and integrative approach for the spatially-resolved ARPES data for complex physical phenomena.

In this study, we focused on graphene systems in which the ARPES spectral features undergo significant changes depending on the number of layers and local step structures. The same approach can be effectively applied to other graphene systems in which spectral structures are highly sensitive to heterogeneous local environments, such as the presence of adsorbates, intercalations, stacking of layers[14,44–48]. In such cases, NMF provides a useful means to decompose complex spectra and extract contributions from localized electronic states in intercalation-induced band modifications. A similar approach could be extended to spatially resolved measurements performed with finer spatial sampling, where multiple electronic phases may coexist within a single illumination area. Under such conditions, the measured spectra often represent superpositions of signals originating from different local domains; NMF-based analysis may offer a way to separate and extract the individual spectral contributions associated with each phase. At the same time, it should be noted that non-ideal effects — including local band-energy shifts, matrix-element effects on each phase, and subtle variations in orientation — can further complicate the spectral signatures and cause the extracted activation weights to deviate from true phase fractions. Careful consideration of these effects is therefore required. Once such factors are systematically evaluated, concepts that explicitly allow for mixed contributions, such as soft clustering or mixture-aware decomposition schemes, are expected to become increasingly important. Incorporating these ideas may enable more realistic and advanced analyses of spatially heterogeneous electronic structures in complex ARPES datasets.

At the stage of data acquisition, it is often unclear which physical characteristics are embedded within the acquired spectra. In such cases, unsupervised learning techniques that do not rely on pre-labeled data or prior training offer significant advantages. Due to the limited beamtime typically available at synchrotron facilities, spatially-resolved ARPES measurements often require on-site data analysis within a constrained time frame. Based on the preliminary results, one may identify points of interest for subsequent, more detailed measurements. In this context, the application of the unsupervised learning framework demonstrated in this work enables the efficient reduction of human labor and time. Furthermore, in order to make efficient use of time, it is beneficial to shorten the measurement duration by reducing the accumulation time at each measurement point. However, this often results in ARPES spectra with poor signal-to-noise (S/N) ratios. The potential applicability of the proposed approach to such low-S/N datasets presents a compelling direction for future work.

#### **4. Conclusions**

In this study, NMF was applied to spatially-resolved ARPES data to decompose the spectra into quantitatively interpretable components. The spatial distributions of the obtained basis vectors revealed distinct features such as layer thickness distribution, electric fields, and inclined layers on bunched steps arising from growth conditions or local environment. NMF provides more interpretable results by decomposing spectra into non-negative, physically meaningful components, and facilitates intuitive understanding of spectral variations across the sample. The unsupervised nature and adaptability of the NMF-based approach make it especially useful for rapid, on-site analysis during synchrotron experiments under limited beamtime. Overall, this establishes NMF as a scalable and practical tool for uncovering spatially-varying local electronic states in low-dimensional and related materials.

#### **Acknowledgements**

This work was supported by JSPS KAKENHI (Grant Numbers: 20K03821, 24K06927) and the Partnership Project for Fundamental Technology Research of the Ministry of Education, Culture, Sports, Science and Technology of Japan.

#### **Conflict of interest**

The authors declare no conflict of interest.

#### **Data availability**

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## References

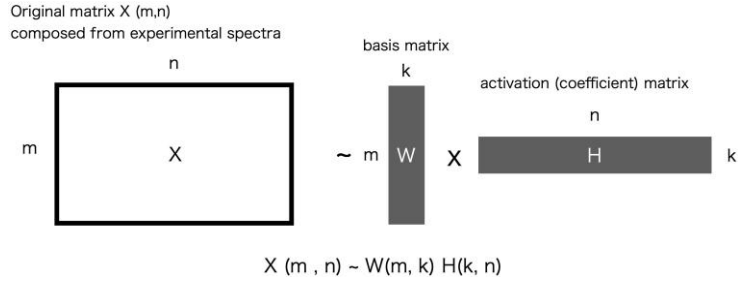
- [1] King PDC, Picozzi S, Egdell RG, et al. Angle, Spin, and Depth Resolved Photoelectron Spectroscopy on Quantum Materials. *Chem Rev.* 2021;121:2816–2856. doi:10.1021/acs.chemrev.0c00616
- [2] Boschini F, Zonno M, Damascelli A. Time-resolved ARPES studies of quantum materials. *Rev Mod Phys.* 2024;96:015003. doi:10.1103/RevModPhys.96.015003
- [3] Dil JH. Spin and angle resolved photoemission on non-magnetic low-dimensional systems. *J Phys Condens Matter.* 2009;21:403001. doi:10.1088/0953-8984/21/40/403001
- [4] Zhang H, Pincelli T, Jozwiak C, et al. Angle-resolved photoemission spectroscopy. *Nat Rev Methods Primers.* 2022;2:1–22. doi:10.1038/s43586-022-00133-7
- [5] Bostwick A, Rotenberg E, Avila J, et al. Zooming in on Electronic Structure: NanoARPES at SOLEIL and ALS. *Synchrotron Radiat News.* 2012;25:19–25. doi:10.1080/08940886.2012.720162
- [6] Avila J, Razado I, Lorcy S, et al. Exploring electronic structure of one-atom thick polycrystalline graphene films: A nano angle resolved photoemission study. *Sci Rep.* 2013;3:2439. doi:10.1038/srep02439
- [7] Hoesch M, Kim TK, Dudin P, et al. A facility for the analysis of the electronic structures of solids and their surfaces by synchrotron radiation photoelectron spectroscopy. *Rev Sci Instrum.* 2017;88:013106. doi:10.1063/1.4973562
- [8] Rösner B, Dudin P, Bosgra J, et al. Zone plates for angle-resolved photoelectron spectroscopy providing sub-micrometre resolution in the extreme ultraviolet regime. *J Synchrotron Rad.* 2019;26:467–472. doi:10.1107/S1600577519000869
- [9] Avila J, Lorcy S, Dudin P. ANTARES: Space-resolved electronic structure. *J Electron Spectrosc Relat Phenom.* 2023;266:147362. doi:10.1016/j.elspec.2023.147362
- [10] Xu L, Mao Y, Wang H, et al. Persistent surface states with diminishing gap in MnBi<sub>2</sub>Te<sub>4</sub>/Bi<sub>2</sub>Te<sub>3</sub> superlattice antiferromagnetic topological insulator. *Sci Bull.* 2020;65:2086–2093. doi:10.1016/j.scib.2020.07.032
- [11] Moriya A, Nakayama K, Kawakami T, et al. Visualizing crystal twin boundaries of bismuth by high-spatial-resolution ARPES. *Phys Rev Res.* 2023;5:023152. doi:10.1103/PhysRevResearch.5.023152
- [12] Lisi S, Lu X, Benschop T, et al. Observation of flat bands in twisted bilayer graphene. *Nat Phys.* 2021;17:189–193. doi:10.1038/s41567-020-01041-x
- [13] Zhang H, Bao C, Jiang Z, et al. Resolving Deep Quantum-Well States in Atomically Thin 2H-MoTe<sub>2</sub> Flakes by Nanoscale Angle-Resolved Photoemission Spectroscopy. *Nano Lett.* 2018;18:4664–4668. doi:10.1021/acs.nanolett.8b00589
- [14] Bao C, Yao W, Wang E, et al. Stacking-dependent electronic structure of trilayer graphene resolved by nanoscale angle-resolved photoemission spectroscopy. *Nano Lett.* 2017;17:1564–1568. doi:10.1021/acs.nanolett.6b04698
- [15] Xian RP, Stimper V, Zacharias M, et al. A machine learning route between band mapping and band structure. *Nat Comput Sci.* 2022;3:101–114. doi:10.1038/s43588-022-00382-2

- [16] Szymanski NJ, Rendy B, Fei Y, et al. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*. 2023;624:86–91. doi:10.1038/s41586-023-06734-w
- [17] Kusne AG, Gao T, Mehta A, et al. On-the-fly machine-learning for high-throughput experiments: search for rare-earth-free permanent magnets. *Sci Rep*. 2014;4:6367. doi:10.1038/srep06367
- [18] Allen AEA, Tkatchenko A. Machine learning of material properties: Predictive and interpretable multilinear models. *Sci Adv*. 2022;8:eabm7185. doi:10.1126/sciadv.abm7185
- [19] Griesemer SD, Xia Y, Wolverton C. Accelerating the prediction of stable materials with machine learning. *Nat Comput Sci*. 2023;3:934–945. doi:10.1038/s43588-023-00536-w
- [20] Iwasawa H, Ueno T, Masui T, et al. Unsupervised clustering for identifying spatial inhomogeneity on local electronic structures. *NPJ Quantum Mater*. 2022;7:24. doi:10.1038/s41535-021-00407-5
- [21] Melton CN, Noack MM, Ohta T, et al. K-means-driven Gaussian Process data collection for angle-resolved photoemission spectroscopy. *Mach Learn Sci Technol*. 2020;1:045015. doi:10.1088/2632-2153/abab61
- [22] Kumazoe H, Iwamitsu K, Imamura M, et al. Quantifying physical insights cooperatively with exhaustive search for Bayesian spectroscopy of X-ray photoelectron spectra. *Sci Rep*. 2023;13:13221. doi:10.1038/s41598-023-40208-3
- [23] Kim Y, Oh D, Huh S, et al. Deep learning-based statistical noise reduction for multidimensional spectral data. *Rev Sci Instrum*. 2021;92:073901. doi:10.1063/5.0054920
- [24] Liu J, Huang D, Yang Y, et al. Removing grid structure in angle-resolved photoemission spectra via deep learning method. *Phys Rev B*. 2023;107:165106. doi:10.1103/PhysRevB.107.165106
- [25] Sims C. Edge Detection and Image Filter algorithms for Spectroscopic Analysis with Deep Learning Applications. 2022. doi:10.48550/arXiv.2203.06820
- [26] Peng H, Gao X, He Y, et al. Super resolution convolutional neural network for feature extraction in spectroscopic data. *Rev Sci Instrum*. 2020;91:033905. doi:10.1063/1.5132586
- [27] Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *Nature*. 1999;401:788–791. doi:10.1038/44565
- [28] Gillis N. Nonnegative Matrix Factorization. Philadelphia: SIAM; 2020.
- [29] Tanimoto H, Hongkun X, Mizumaki M, et al. Non-negative matrix factorization for 2D-XAS images of lithium ion batteries. *J Phys Commun*. 2021;5:115005. doi:10.1088/2399-6528/ac3268
- [30] Mak R, Lerotic M, Fleckenstein H, et al. Non-negative matrix analysis for effective feature extraction in X-ray spectromicroscopy. *Faraday Discuss*. 2014;171:357–371. doi:10.1039/C4FD00023D
- [31] Stanev V, Vesselinov VV, Kusne AG, et al. Unsupervised phase mapping of X-ray diffraction data by nonnegative matrix factorization integrated with custom clustering. *NPJ Comput Mater*. 2018;4:43. doi:10.1038/s41524-018-0099-2

- [32] Shiga M, Muto S. Non-negative Matrix Factorization and Its Extensions for Spectral Image Data Analysis. *e-J Surf Sci Nanotechnol*. 2019;17:148–154. doi:10.1380/ejssnt.2019.148
- [33] Imamura M, Takahashi K. Unsupervised learning of spatially-resolved ARPES spectra for epitaxially grown graphene via non-negative matrix factorization. *Sci Rep*. 2024;14:24200. doi:10.1038/s41598-024-73795-w
- [34] Ohta T, Bostwick A, Seyller T, et al. Controlling the Electronic Structure of Bilayer Graphene. *Science*. 2006;313:951–954. doi:10.1126/science.1130681
- [35] Jin S, Zong J, Chen W, et al. Epitaxial Growth of Uniform Single-Layer and Bilayer Graphene with Assistance of Nitrogen Plasma. *Nanomaterials*. 2021;11:3217. doi:10.3390/nano11123217
- [36] Riedl C, Coletti C, Starke U. Structural and electronic properties of epitaxial graphene on SiC(0001): a review of growth, characterization, transfer doping and hydrogen intercalation. *J Phys D Appl Phys*. 2010;43:374009. doi:10.1088/0022-3727/43/37/374009
- [37] Yu XZ, Hwang CG, Jozwiak CM, et al. New synthesis method for the growth of epitaxial graphene. *J Electron Spectrosc Relat Phenom*. 2011;184:100–106. doi:10.1016/j.elspec.2010.12.034
- [38] Zebardastan N, Bradford J, Lipton-Duffin J, et al. High quality epitaxial graphene on 4H-SiC by face-to-face growth in ultra-high vacuum. *Nanotechnology*. 2023;34:105601. doi:10.1088/1361-6528/aca8b2
- [39] Takahashi K, Imamura M, Yamamoto I, et al. Upgrade of Saga-university beamline in SAGA-LS. *J Phys Conf Ser*. 2013;425:072007. doi:10.1088/1742-6596/425/7/072007
- [40] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12:2825–2830.
- [41] Yoshinari A, Iwasaki Y, Kotsugi M, et al. Skill-agnostic analysis of reflection high-energy electron diffraction patterns for Si(111) surface superstructures using machine learning. *Sci Technol Adv Mater Methods*. 2022;2:162–174. doi:10.1080/27660400.2022.2079942
- [42] Emtsev KV, Bostwick A, Horn K, et al. Towards wafer-size graphene layers by atmospheric pressure graphitization of silicon carbide. *Nat Mater*. 2009;8:203–207. doi:10.1038/nmat2382
- [43] Norimatsu W, Kusunoki M. Formation process of graphene on SiC (0001). *Physica E*. 2010;42:691–694. doi:10.1016/j.physe.2009.11.151
- [44] Sakakibara R, Bao J, Hayashi N, et al. Control of rotation angles of multilayer graphene on SiC (000 $\bar{1}$ ) by substrate off-direction and angle. *J Phys Condens Matter*. 2023;35:385001. doi:10.1088/1361-648X/acdebf
- [45] Yang D, Ma F, Bian X, et al. The growth of epitaxial graphene on SiC and its metal intercalation: a review. *J Phys Condens Matter*. 2024. doi:10.1088/1361-648X/ad201a
- [46] Castro Neto AH, Guinea F, Peres NMR, et al. The electronic properties of graphene. *Rev Mod Phys*. 2009;81:109–162. doi:10.1103/RevModPhys.81.109
- [47] Tsujikawa Y, Sakamoto M, Yokoi Y, et al. Controlling of the Dirac band states of Pb-deposited graphene by using work function difference. *AIP Adv*. 2020;10. doi:10.1063/5.0013797
- [48] Wu X, Zheng F, Kang F, et al. Effects of lithium intercalation in bilayer graphene. *Phys Rev B*.

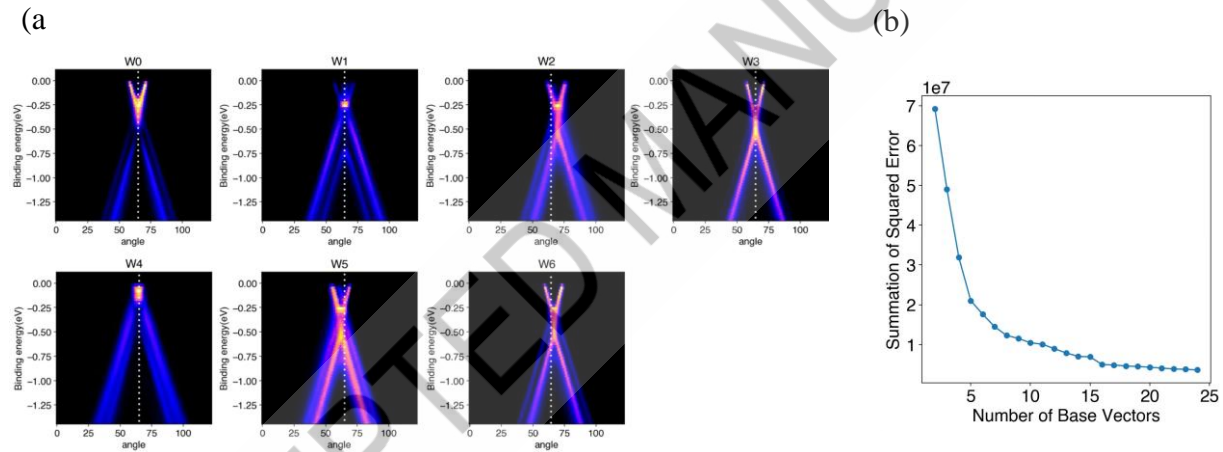
ACCEPTED MANUSCRIPT

**Fig. 1.**



**Fig. 1.** Schematic illustration of Non-negative Matrix Factorization (NMF). A non-negative data matrix  $X$  is approximately factorized into a basis matrix  $W$  and an activation matrix  $H$ .

**Fig. 2.**



**Fig. 2.** (a) Basis vectors obtained from NMF with  $k = 7$ . Dotted lines correspond to guides to the eye aligned with the Dirac point of  $W_0$ . (b) Summation of squared errors that were calculated between experimental ARPES spectra and the reconstructed spectra through NMF was plotted as a function of the number of basis vector  $W$ .



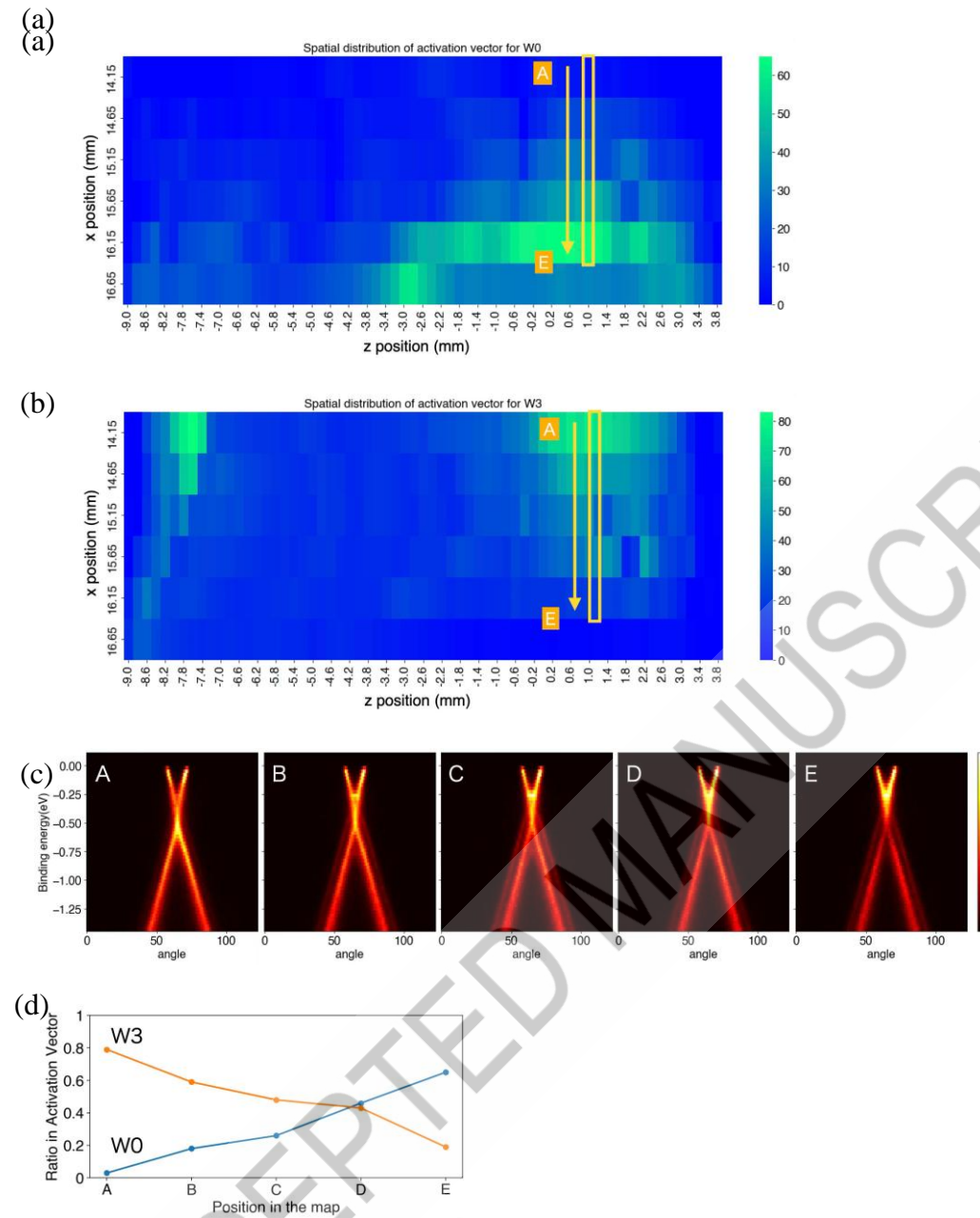
**Fig. 3.**

Fig. 3. Heatmaps visualizing the spatial distribution of relative intensities in the activation matrix  $H$ , for the basis vectors (a)  $W0$  and (b)  $W3$ , obtained by applying NMF to the spatially-resolved ARPES dataset of graphene on SiC. (c) Representative ARPES spectra measured at positions A–E indicated in (a) and (b). All spectra exhibit BL-derived Dirac-cone features, with a systematic reduction in spectral weight on the higher-binding-energy side from A to E. (d) The contribution ratio of the corresponding activation vector  $H$  at each point shown in Figs. 3 (a) and (b) for basis vectors  $W0$  and  $W3$ .

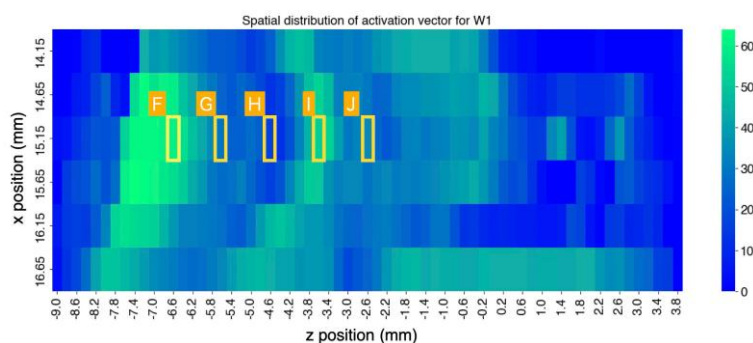
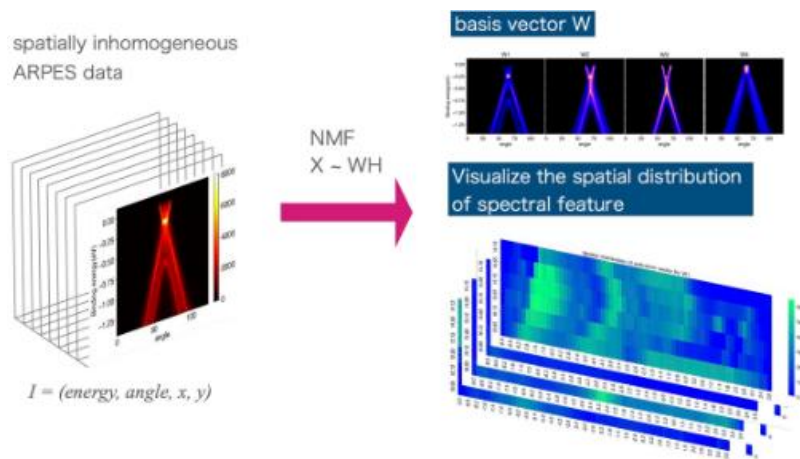
**Fig. 4.**

Fig. 4. Heatmaps visualizing the spatial distribution of relative intensities in the activation matrix  $H$ , for the basis vectors (a)  $W1$  and (b)  $W4$ , obtained by applying NMF to the spatially-resolved ARPES dataset of graphene on SiC. (c) Representative ARPES spectra measured at positions F–J indicated in (a) and (b). All spectra exhibit TL-derived Dirac-cone features. (d) The contribution ratio of the corresponding activation vector  $H$  at each point shown in Figs. 4 (a) and (b) for basis vectors  $W1$  and  $W4$ .

(b)

(c)

(d)



graficalabstract

## **IMPACT STATEMENT**

This study presents a practical analysis framework utilizing non-negative matrix factorization. It enables the automated decomposition and spatial visualization of latent electronic features in inhomogeneous samples without prior assumptions.

ACCEPTED MANUSCRIPT