IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 September 2025, accepted 14 September 2025,
date of publication 30 September 2025, date of current version 10 October 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3615908

RESEARCH ARTICLE

Toward Automated Chemical Analysis of Materials Using Secondary Electron Hyperspectral Imaging and Unsupervised Learning

JINGQIONG ZHANG[1], **NICHOLAS T. H. FARR**[2,3], **JAMES NOHL**[2,4], **YUFENG LAI**[1], **KERRY J. ABRAMS**[2], **KATE BLACK**[5], **JON WILLMOTT**[1], **CORNELIA RODENBURG**[2,3], **AND LYUDMILA MIHAYLOVA**[1,3], **(Senior Member, IEEE)**
[1]School of Electrical and Electronic Engineering, The University of Sheffield, S1 3JD Sheffield, U.K.
[2]School of Chemical, Materials and Biological Engineering, The University of Sheffield, S1 3JD Sheffield, U.K.
[3]Insigneo Institute for In Silico Medicine, The University of Sheffield, S1 3JD Sheffield, U.K.
[4]The Faraday Institution, Quad One, Harwell Campus, OX11 0RA Didcot, U.K.
[5]School of Engineering, University of Liverpool, L69 3GH Liverpool, U.K.

Corresponding author: Jingqiong Zhang (jqzhang7777@hotmail.com)

This work was supported in part by the U.K. EPSRC ‘‘SEE MORE MAKE MORE: Secondary Electron Energy Measurement Optimization for Reliable Manufacturing of Key Materials’’ under Grant EP/V012126/1, in part by EPSRC funding through the SMART Feasibility Study from ‘‘MAPP: EPSRC Future Manufacturing Hub in Manufacture using Advanced Powder Processes under Grant EP/P006566/1, and in part by the Faraday Institution Project under Grant FITG028B.

ABSTRACT Advancements in materials science have significantly transformed materials discovery and advanced manufacturing. This, along with the rapid development of sensing and instrumentation, results in a continuous increase in data volumes. To address the limitations of conventional manual analysis, this paper introduces an AI-driven framework for high-throughput chemical analysis of material surfaces at the micro- and nano-scale. The framework integrates unsupervised machine learning with secondary electron hyperspectral imaging (SEHI). It consists of four stages: hyperspectral image processing via tiling, spectral peak extraction, peak categorisation by probabilistic clustering, and chemical analysis. Tiling enables the capture of local spatial-spectral information and generation of a large number of training samples from a single SEHI image stack. After tile-wise spectral peak extraction, the distribution of the peak positions is accurately represented by probabilistic clustering with a Gaussian mixture model (GMM) or a Dirichlet process Gaussian mixture model (DPGMM). Each peak corresponds to a specific chemical bond or element in a material, reflecting the unique spectral characteristics. The performance of the GMM and GPGMM approaches is validated over a case study for identifying chemical elements or bonds of complex metal alloy and carbon films. The results demonstrate accurate chemical analysis, yielding relative errors within ±15% compared to the theoretical model of the valence band density of states. This work is a step forward towards automated material analysis across different tasks such as identifying chemical elements and bonds, visualizing surface (in)homogeneity in metal alloy films for guiding film printing, and supporting digital twins integration for advanced manufacturing.

INDEX TERMS Advanced manufacturing, artificial intelligence, Gaussian mixture models, material surface chemistry, microscopy, probabilistic clustering, secondary electron spectroscopy, unsupervised learning.

I. INTRODUCTION

Materials science and industry are embracing the broad adoption of machine learning and artificial intelligence (AI) for materials design, discovery, analysis, understanding,

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Ni.

prediction as well as characterisation [1], [2], [3], [4], [5], [6], [7], [8], [9].

A. MACHINE LEARNING FOR MATERIALS ANALYSIS

Machine learning has been proven to be a powerful tool to excavate complicated input–output mappings. As stated in [10], the role of machine learning and big-data science in manufacturing industry is evolving from merely providing basic machine automation to information automation, and ultimately to knowledge automation. It emphasises the growing importance of AI-aided information extraction from data in modern manufacturing processes.

The recent advancements in machine learning hold great potential to enhance diverse aspects of materials science research and industry, including processing high-dimensional data, automating data analytical workflows, and uncovering new knowledge [11], [12], [13], [14]. One common practice is to apply machine learning into the discovery of complex composition-structure-property relationships in both real and hypothetical materials [8]. For example, correlations between the macro-scale mechanical properties of heterogeneous materials and their microstructure were established by using a convolutional neural network (CNN) based deep learning method [15]. Similarly, an artificial neural network was deployed to reveal the complex relationship between the inclusion features and fatigue life of steels [16]. Yu et al. [17] applied a multimodal deep neural network (DNN) for learning the mechanical properties of steels from the material chemical composition. Besides, a Gaussian process regression (GPR) model was trained to establish, and to understand, the structure-property linkages of synthetic microstructure towards computationally aided materials design [18]. In addition to the application examples listed above, machine learning also plays an important role in providing chemical insights, in favour of the data-driven discovery (‘‘fourth paradigm of science’’) [8]. For instance, Paul et al. [19] proposed an AI solution based on mixed DNN architectures which can be used to predict chemical properties, such as activity, toxicity, as well as solubility, from two molecular representations as the inputs.

However, much of the success in supervised machine learning approaches is heavily dependent on the availability of large and representative labelled training datasets. Such achievement would be limited in many real-world situations, particularly in the cases involving experimental measurements, where both data acquisition and labeling become difficult or time-consuming. Especially, when characterising unknown material samples or discovering new structures, it would be more challenging to obtain appropriate prior knowledge, labelled or ground truth data, for deploying supervised learning algorithms in practice.

Given these challenges, there is a strong need to develop unsupervised machine learning solutions, which inherently requires less endeavors in data preparation. Some attempts have been made in this area, however, so far there is limited research undertaken to provide unsupervised learning strategies in materials applications. Vlcek et al. [20] applied a variational autoencoder to compress the atomic configurational data for identifying defects and monitoring abnormal behavior in the composition phase. Uesug et al. [21] demonstrated the validity of applying non-negative matrix factorization (NNMF) in images acquired by scanning transmission electron microscopy to deduce the underlying diffraction patterns of titanium oxide nanosheets. An unsupervised approach using swarm intelligence and emergence was proposed to classify archaeological materials [22]. Aversa et al. [23] combined supervised and unsupervised learning approaches for semi-supervised classification of nanostructured materials images acquired by scanning electron microscopy. Besides, a clustering algorithm based on neural network architecture was proposed to classify fracture surfaces [24].

These above are only several examples of unsupervised machine learning applications to materials science. The key to advancing materials science and engineering lies in the effective integration of established machine learning tools into materials engineering workflows [25]. This requires aligning the development of AI with the real-world demands and challenges in materials research, whilst remaining guided by domain-specific knowledge to ensure systematically interpretable
outcomes.

For instance, a particularly longstanding challenge here is how to bridge the theoretical or computational materials science with experimental areas such as materials characterisation. Often this includes scanning electron microscopy (SEM), one of the most widely used experimental characterisation tools in materials science. This is reflected in extensive recent research efforts exploring the adoption of unsupervised learning techniques with SEM images across various domains of materials science. The application fields include arts [26], geology [27], [28], [29], semiconductors [30], [31], building materials [32], concrete [33], additive manufacturing [34], and metal films [35]. A significant attention is given to image denoising and quality enhancement for materials [30], [31], [36], [37], [38], contour detection [39], image classification [23], segmentation [28], as well as pattern recognition [35].

These previous research studies have demonstrated the effectiveness of unsupervised learning in materials characterisation using SEMs. Other studies have shown how unsupervised learning can be integrated with SEM imaging and local property analysis obtained from energy dispersive X-ray spectroscopy (EDS), referred to as SEM-EDS [26], [32], [33], [40]. For instance, automated mineral phase analysis through image segmentation and unsupervised clustering was developed in [40]. This helps the analysis of sparse EDS spectral data and their combination with SEM images for graph construction to obtain image segments. The research work [26] presents an unsupervised learning-based data analysis approach for automatically extracting chemical information at the elemental level from painting materials using SEM-EDS instrumentation. Other studies such as [32] demonstrate that clustering methods, i.e., the k-means and Gaussian mixture model (GMM), combined with SEM-EDS techniques, can classify micromechanical properties of building materials.

B. SECONDARY ELECTRON HYPERSPECTRAL IMAGING

SEM, as one of the most popular and versatile micro- and nano-scale imaging methods, has been widely used across a diverse range of industrial applications and research fields, including manufacturing, nanotechnology, material, biological, as well as medical sciences [41]. The working principle of a SEM is to locally generate, and then detect electron emissions created by elastic and inelastic interactions between the scanning electrons and the sample. SEMs are typically employed for investigating surface morphology, topography, along with chemical composition of materials by offering detailed and high-resolution images of their surfaces [42], [43].

Secondary electron hyperspectral imaging (SEHI) is a form of hyperspectral imaging technique that utilises energy filtering in the SEM to generate a hyperspectral three-dimensional ($3D$) data cube [44]. The SEHI data cube is made up of high-resolution sequential images resulting from secondary electrons (SEs) emitted from the same sample region. Each SE image is produced from a specific energy range of the SEs that pass a low-pass energy filter before detection. By deriving SE spectra from the region of interest (ROI) of the sequential SEs images, SEHI is able to provide local information about the sample surface chemistry, making it a valuable tool for characterising materials at the micro- to nano-scale [45].

In contrast to SEM-EDS techniques, SEHI fundamentally differs in both information depth and analytical capability. While EDS offers elemental analysis at micron-scale depths, SEHI captures chemical bonding states at the nanoscale, providing unique insights into surface chemistry of nanoscale materials. For a detailed technical discussion, please see Figures S3-S5 and related analysis in Supporting Material [46] of [47]. SEHI is an emerging technology for experimentally investigating the morphology and chemical properties of material surfaces. SEHI has proved its potential, especially for carbon based materials, such as in tracking controlled changes to the surface chemistry of poly(glycerol sebacate)-methacrylate (PGS-M) polymer through plasma treatments, and the ageing surface of newly exfoliated highly oriented pyrolytic graphite (HOPG) [48], as well as complex carbon and metal/carbon systems [49], since 2019. Compared to the transmission electron microscopy, SEMs and SEHI are often used to inspect large areas of material samples. Chemical strongly localised inhomogeneity can be masked when spectral analysis is carried out from the whole area that is large compared to the localised inhomogeneity.

C. MACHINE LEARNING FOR SEHI DATA ANALYSIS

The success of using SEHI technique across real-world material applications can be limited by manual data analysis [50]. An existing problem is that manual analysis often comes along with human intuition. It can lead to inconsistency and inaccuracy due to cognitive biases among different investigators. A common source of human error here is the manual selection of the ROIs, which always greatly rely on the judgment of investigators based on their expertise or understanding [51]. In some materials systems, variations in surface chemistry are linked with the surface microstructural features [48]. However, this is not always the case. In other materials, the surface topographical features can negatively introduce cognitive bias in materials analysis. Thus, it is crucial to develop an automated and unbiased data analysis approach, towards accurate high-throughput materials characterisation from laboratory to industrial applications.

Integrating machine learning with SEHI analytical technique offers a promising solution that captures the rich spatial and spectral information from experimental $3D$ SEHI data. As stated in Section I-A, unsupervised machine learning approaches are powerful tools to tackle the challenges in such applications where no (or very little) ground truth data and only small datasets are available for learning. In response to such need, this paper proposes a novel AI-driven framework by combining SEHI with probabilistic clustering algorithms which can be used for characterising materials surface chemistry down to nano-scale. This framework is expected to promote automated materials chemical analysis, particularly for investigating complex carbon-based material systems where the atoms of carbon can bond together in diverse ways.

D. MAIN CONTRIBUTIONS

One of the primary contributions of this work to the materials science research community consists in the development of unsupervised learning approaches that can facilitate the integration of theoretical knowledge with simulated data into an SEM-based experimental workflow for materials characterisation. Specifically, the simulated data refers to the density of states (DOS) model obtained from first-principles calculations [52], which can be retrieved from the publicly available database in *The Materials Project* [53]. The DOS model essentially represents the number of available quantum states, e.g. electron energy levels per unit energy within a material [54]. It serves as a theoretical reference for validating our experimental results of SEHI data analysis.

Unlike previous studies that combined secondary electron imaging with EDS, SEHI utilises the same signal (SE electrons) for both imaging and spectroscopy. This offers high spatial resolution, which is crucial for characterising complex material systems, e.g., containing nano-particles. In addition, SEHI enables insights into chemical bonding-level information for materials characterisation that was not previously accessible through SEM-EDS techniques. However, the complexity and large volume of spectral image data obtained from

SEHI necessitate the development of a new automated data analysis workflow. To demonstrate this, we have presented an example of a complex material system containing metallic nano-particles dispersed in a carbon matrix, fabricated via 3D printing, and requiring reliable analysis of experimental SEHI data.

The main contributions of this paper are the following:

1) It proposes a novel AI-driven framework for automating SEHI data analysis. It creates chemical information maps and links them with the theoretical DOS model. This enables accurate chemical identification and characterisation of materials (in)homogeneity at micro- and nano-scales. A unique aspect of this work is that it bridges experimental results with theoretical analysis.
2) It develops an efficient tiling partitioning of a SEHI image stack into a large number of localised spatial-spectral regions. This allows capturing the rich spatial-spectral information and feeding diverse data into the downstream machine learning task.
3) Combined with the image tiling process, we adopt probabilistic clustering models, a GMM and a Dirichlet process Gaussian mixture model (DPGMM), for unsupervised learning. The developed framework does not require labelled training datasets. It offers two pathways for categorising spectral peaks and meets user needs in different scenarios where the cluster number of the material being analysed is either known a priori or needs to be inferred automatically.
4) The performance of this new framework is thoroughly evaluated over the real SEHI data collected from a complex metal alloy and carbon material system, showing accurate chemical identification by comparison with the theoretical DOS model.

To the best of the authors' knowledge, for the first time, this work proposes an automated analytical workflow for SEHI data, and quantifies the experimental measurement errors in the identification of chemical bonds using SEHI, with reference to the theoretical DOS model. Additionally, research outputs include publicly available SEHI datasets [55] and an AI-powered data analytical tool [56] to benefit the wider research community.

The rest of this paper is organised as follows: Section II introduces the mathematical background of the adopted probabilistic clustering approaches, GMM and DPGMM. Section III presents the overall framework proposed. Section IV demonstrates our framework through a real-world use case, detailing the problem, datasets, implementation, and evaluation metrics. Section V provides the experimental validation results and ablation studies. Section VI summarises the conclusions and discusses future work.

II. PRELIMINARIES

Clustering is a fundamental problem in machine learning, data mining, and statistical analysis [57]. Unsupervised clustering identifies inherent data groups, no need of prior knowledge of the group labels. Under the assumption that data is generated by a specific statistical model, probabilistic clustering approaches estimate the model parameters to fit the data distribution. In contrast to some conventional clustering techniques which rely on distance-based or density-based metrics, probabilistic clustering approaches are flexible, robust and versatile [58]. They provide a probabilistic interpretation of the clustering process, hence making the results statistically meaningful, and less sensitive to data outliers. In this paper, two probabilistic clustering methods, GMM and DPGMM, are adopted for clustering all spectral peaks produced from the image tiling on SEHI data.

A. GAUSSIAN MIXTURE MODELLING

GMM represents a complex distribution as a weighted combination of multiple Gaussian distributions for a dataset $\mathbf{X}$ of N data points $\mathbf{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i, \boldsymbol{x}_N\}$, which can be formulated as

$$p(\boldsymbol{x}_i) = \sum_{k=1}^{K} \omega_k \mathcal{N}(\boldsymbol{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{1}$$

where K is the cluster number and $\mathcal{N}(\boldsymbol{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ denotes the k-th Gaussian component with its mean vector $\boldsymbol{\mu}_k$, covariance matrix $\boldsymbol{\Sigma}_k$, and ω_k the mixture probability (or called component proportion). The sum of ω_k is equal to 1, namely $\sum_{k=1}^{K} \omega_k = 1$.

The model parameters of these Gaussians, $\{\boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$, are iteratively computed by maximising the log-likelihood function [59]

$$lnL(\mathbf{X} \mid \boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{N} ln\left(\sum_{k=1}^{K} \omega_k \mathcal{N}(\boldsymbol{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right). \tag{2}$$

where ln is the natural logarithm, and $lnL(\mathbf{X} \mid \boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the log-likelihood function.

To maximise this log-likelihood function, the model parameters are iteratively estimated via the expectation-maximisation (EM) algorithm [60], [61]. The posterior probability, $\boldsymbol{r}_{i,k}$, given the model parameters $\{\boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$, is expressed as follows

$$\boldsymbol{r}_{i,k} = p(k \mid \boldsymbol{x}_i, \boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\omega_k \mathcal{N}(\boldsymbol{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k=1}^{K} \omega_k \mathcal{N}(\boldsymbol{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}. \tag{3}$$

where $\boldsymbol{r}_{i,k}$ represents the posterior probability that data point $\boldsymbol{x}_i$ belongs to cluster k, reflecting the confidence of the assignment. The main challenge of the GMM clustering algorithm is the necessity of choosing the appropriate number K of clusters. Then the Bayesian information criterion (BIC) can be employed for model selection, by measuring the balance of the goodness of model fit with the simplicity of model [62].

B. DIRICHLET PROCESS GAUSSIAN MIXTURE MODELLING

By employing a Dirichlet Process (DP) prior into mixture modelling, DPGMM extends the ability of GMM to deal with the unknown number of components in a mixture. DP prior is commonly served in non-parametric Bayesian approaches, enabling the number of clusters inferred from the dataset automatically [58]. Additionally, the singularity problem can be effectively solved through the use of a Bayesian prior into probabilistic mixture modelling. However, exact inference for DPGMM is not tractable, which means it cannot be evaluated analytically. In such scenarios, approximation techniques need to be implemented so as to estimate the target posterior distribution. In this work, model learning in DPGMM is accomplished via the collapsed Gibbs sampling algorithm, which belongs to Markov Chain Monte Carlo (MCMC) sampling methods [63], [64].

A symmetric Dirichlet distribution, Dir, can be parameterised by a single hyperparameter, α, also known as the concentration parameter. Intuitively α controls how likely to create a new cluster. Large α results in more clusters expected. A DP can be acquired by taking the limit of $K \rightarrow \infty$. Accordingly, $G \sim Dir(\alpha)$, and G becomes an infinite dimensional probability vector, representing the component proportions in the mixture modelling. Chinese restaurant process (CRP) is a probabilistic interpretation of the DP, which defines a distribution over an infinite number of clusters. This allows for a flexible, non-parametric clustering approach where the number of clusters is inferred from the data rather than fixed in advance [65]. In the CRP, a new customer $\boldsymbol{x}_i$ selects a table depending on the number of customers already sitting at each table occupied [66]. As the Dirichlet distribution is a conjugate prior of categorical distribution Cat, it is commonly utilized in Bayesian inference as a prior distribution [67]. Namely $z_i \mid G \sim Cat(G)$, where z_i denotes the discrete latent variable. z_i represents the cluster assignment for data point $\boldsymbol{x}_i$. It corresponds to the table assigned to the customer $\boldsymbol{x}_i$ in the CRP.

Regarding the mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$ of each Gaussian component, a Normal-Inverse-Wishart (NIW) distribution is utilized as the prior distribution. It is governed by four hyperparameters $\boldsymbol{\theta} = \left\{\lambda_0, \boldsymbol{\mu}_0, \nu_0, \boldsymbol{S}_0\right\}$ and $\left\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right\} \sim NIW(\boldsymbol{\theta})$. The NIW distribution is a conjugate prior of a multivariate normal distribution $\mathcal{N}$ [65], [68]. The relevant expressions are,

$$\boldsymbol{\mu} \mid \boldsymbol{\mu}_0, \lambda_0, \boldsymbol{\Sigma} \sim \mathcal{N}\left(\boldsymbol{\mu} \mid \boldsymbol{\mu}_0, \frac{\boldsymbol{\Sigma}}{\lambda_0}\right), \tag{4}$$

$$\boldsymbol{\Sigma} \mid \boldsymbol{S}_0, \nu_0 \sim W^{-1}\left(\boldsymbol{\Sigma} \mid \boldsymbol{S}_0, \nu_0\right), \tag{5}$$

$$\boldsymbol{x}_i \mid z_i, \left\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right\} \sim \mathcal{N}\left(\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}\right). \tag{6}$$

During each iteration of model learning process based on the Gibbs sampler, each data point is selected one by one for performing the assignment of clusters, either to an existing cluster or a new cluster. After the cluster assignment, the model parameters are updated accordingly,

$$\boldsymbol{\mu}_n = \frac{\lambda_0 \boldsymbol{\mu}_0 + n\bar{\boldsymbol{x}}}{\lambda_n}, \tag{7}$$

$$\lambda_n = \lambda_0 + n, \tag{8}$$

$$\nu_n = \nu_0 + n, \tag{9}$$

$$\boldsymbol{S}_n = \boldsymbol{S}_0 + \boldsymbol{J} + \frac{\lambda_0 n}{\lambda_n}\left(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0\right)\left(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0\right)^T, \tag{10}$$

$$\boldsymbol{J} = \sum_{i=1}^{n}\left(\boldsymbol{x}_i - \bar{\boldsymbol{x}}\right)\left(\boldsymbol{x}_i - \bar{\boldsymbol{x}}\right)^T, \tag{11}$$

where n is the number of data points which have been observed, and $\bar{\boldsymbol{x}}$ denotes the mean of observations.

Benefiting from the conjugate prior distributions, several parameters can be marginalised analytically when model learning via the collapsed Gibbs sampler. The posterior probability for the data point $\boldsymbol{x}_i$ assigned to cluster k can be presented by,

$$\begin{aligned} &p\left(z_i = k \mid \boldsymbol{x}_i, \boldsymbol{X}_{-i,k}, z_{-i}, \alpha, \boldsymbol{\theta}\right) \propto \\ &p\left(z_i = k \mid z_{-i}, \alpha\right) \cdot p\left(\boldsymbol{x}_i \mid z_i = k, \boldsymbol{X}_{-i,k}, \boldsymbol{\theta}\right). \end{aligned} \tag{12}$$

where $z_{-i} = \left\{z_j \mid j \neq i\right\}$ is the cluster assignments excluding z_i, and $X_{-i,k} = \left\{\boldsymbol{x}_j \mid z_i = k, j \neq i\right\}$ refers to the set of data points already assigned to that cluster k, except $\boldsymbol{x}_i$.

It can be seen from (12) that the posterior probability is derived from two terms, namely the prior probability and the likelihood. The expression of the prior term $p\left(z_i = k \mid z_{-i}, \alpha\right)$ is straightforward under the CRP. The likelihood term $p\left(\mathbf{x}_i \mid z_i = k, \boldsymbol{X}_{-i,k}, \boldsymbol{\theta}\right)$ can be calculated via the posterior predictive distribution, $p(\boldsymbol{x}_i \mid \boldsymbol{X}_{-i,k}, \boldsymbol{\theta})$, which describes the likelihood of the current data point x_i under the given observed data. It is proved to be equal to the probability density of a multivariate t-distribution [65].

III. OVERALL FRAMEWORK

The proposed framework consists of four main stages: hyperspectral image processing via tiling, spectral peak extraction, peak categorisation by probabilistic clustering and surface chemical analysis. Fig. 1 presents the overall AI-powered framework of the automated SEHI data analysis approach.

A. HYPERSPETRAL IMAGE TILING

As depicted in Fig.1, a SEHI data cube can be described in a $3D$ coordinate system, where the xy-plane denotes the image plane and the z-axis represents the energy of the SEs detected in the SEM. In order to overcome the limitations of manual selection of ROI in earlier work [48], [49], [69], [70], [71], we propose to represent the hyperspectral images by subdividing into small regions, called *tiles*, in the xy-plane. Next, the clustering and spectral analysis is performed directly on these tiles, instead of selected ROI or the whole field of view (FOV).

Firstly, sequential hyperspectral images in a SEHI data cube should be aligned based on image morphology features, before performing subsequent spectral analysis [70]. Image alignment is usually accomplished through a template matching algorithm. For using template matching here, python code can be found [72] and a standalone Matlab application is publicly available [73]. Then we choose non-overlapping and square tiles so as to preserve the local details and morphology information of the hyperspectral images. The tile size can be

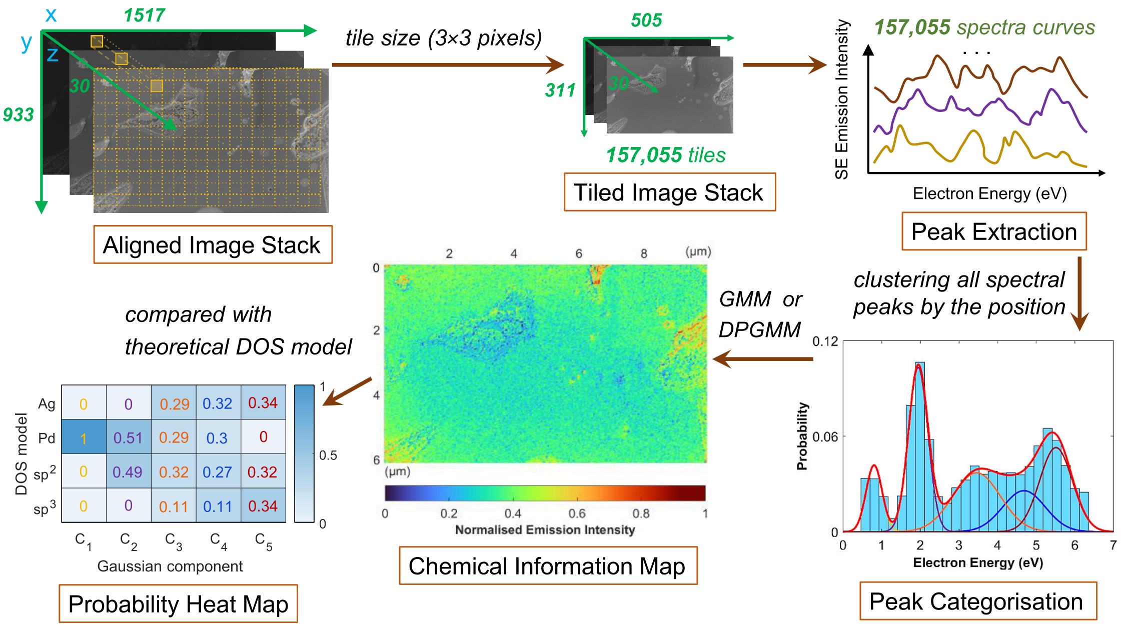


FIGURE 1. The overall framework of the proposed SEHI data analysis approach for automated materials chemical analysis. The framework consists of: image tiling, spectral peak extraction, peak categorisation by unsupervised clustering, and chemical interpretation along with probability heat maps to support decision-making. We segment a $3D$ SEHI data cube into small tiles along its image plane, and extract the peaks from the spectral curves for each tile. Then the distribution of these spectral peaks is accurately modelled by probabilistic clustering methods, GMM and DPGMM.

defined by the users based on specific application needs. And the tiles traverse vertically and horizontally over the whole FOV. This tiling process automatically creates numerous ROIs, instead of relying on just a single or few selected ROIs by the users. For every image slice, the intensity value of each tile is computed by averaging its pixels belonging to this tile. After tiling, the raw SEHI data cube is downsized over the xy-plane. Then, by differentiating the intensity values along the z-axis, a spectrum curve of SEs emission is extracted from each tile. In addition, the spectral curves are further smoothed along the z-axis by averaging the intensity values of every two adjacent data points to reduce noise, as described in [70].

B. SPECTRAL PEAK EXTRACTION

Peaks can be defined as local or sometimes global maximums, in comparison with the adjacent data points. In this work, we initially define a peak as a local maximum with two neighbours on each side, and then set reasonable criteria to select predominant and informative peaks from all potential peaks. The spectral peaks are determined using the following steps:

1) FILTER OUT WEAK PEAKS BY THE PEAK HEIGHT

Small peaks with heights below a specified threshold value are removed to account for noise and fluctuations in the spectra curves. The threshold value is defined by the ratio of peak height with respect to the global maximum. It controls the sensitivity of peak detection. Lower threshold values preserve weak spectral features but increase the chance of noise inclusion. In this work, the threshold value was empirically set as 0.4. The threshold can be adjusted by end user, to adapt to different materials and instruments in practice.

2) REMOVE THE LOWER PEAK WHEN TWO PEAKS ARE VERY CLOSE TO EACH OTHER

The idea here is to keep "true" peaks in cases where a cascade of peaks can appear in a consistent upward or downward trend. The spectral curves represent the fluctuations in the intensity of SEs emissions from the material's surface being analysed. The peaks in the spectra curves are typically regarded as spectral signatures of the material for chemical identification [74].

Other peak-related properties, such as peak height or width, may contain supplementary information. For instance, peak width can be indicative of disorder in some materials, as demonstrated for semicrystalline polymers [75] and perovskite materials [75], [76]. Yet, these properties are highly sensitive to experimental conditions, and instrument parameters [75]. They are fundamentally affected by complex interactions between SEs emitted and material surfaces. For example, peak height can be strongly influenced by experimental conditions, including the instrument utilised, and the working distance during the use of the instrument that can affect the detection of SEs [45]. In contrast, peak position is a more robust and reliable feature for chemical identification. The peak positions across different instruments can be corrected and aligned through calibration against reference materials [70].

Our choice to utilise spectral peak positions as the sole feature for clustering is driven by the study's core objective of achieving reliable and accurate chemical identification and assignment. While extending this framework to other applications (e.g., structural disorder analysis) is promising, such extensions require careful consideration of technical challenges depending on the specific needs of applications that fall beyond the scope of this study.

C. PEAK CATEGORISATION BY PROBABILISTIC CLUSTERING

After extracting the spectral peaks from all tiles, the subsequent task is to categorise these peaks. In other words, it is to label the spectral peaks based on their energy positions, and sort them according to their labels. By employing probabilistic clustering, the inherent characteristics of the spectral peaks can be inferred from its probability distribution. We adopt the GMM and DPGMM approaches for clustering the spectral peaks collected from all tiles, as the spectral data and its distribution inherently exhibit multi-peak behaviour. To meet the needs in real-world circumstances, depending on whether the cluster number is known, two clustering schemes are presented: (1) The GMM has a finite number of Gaussian components with a specified cluster number, (2) The DPGMM, with an infinite number of Gaussian components, handles an unknown number of clusters.

When adopting the GMM, one primary question is how to choose a appropriate cluster number K. In cases where the number of clusters (e.g., chemical components) in material sample is known, investigators can set the cluster number K directly. When K is unknown and computational efficiency is a concern, the GMM can be used in combination with the Bayesian information criterion (BIC) [77] or Akaike information criterion (AIC) to find the optimal K. BIC and AIC measures are widely used for model selection, balancing the goodness of model fit with model complexity [78]. Both criteria follow the principle that lower values indicate better performance. In practice, the BIC is often preferred over the AIC to mitigate model overfitting, as it imposes a stronger penalty on model complexity [79], particularly for large datasets. Thus, the BIC is used in this work. In contrast, DPGMM is able to automatically learn the cluster number from a DP prior. It is flexible in estimating the effective complexity of the mixture model, however Bayesian inference requires additional computation [80].

The parameters of the Gaussian components, derived from clustering, have significant physical meanings related to the material chemistry. As the peak positions for clustering are one-dimensional data, its variance $\boldsymbol{\sigma}^2$ is used, instead of $\boldsymbol{\Sigma}$, in later descriptions for simplicity. For the k-th Gaussian component, its mean μ_k denotes the centre of the Gaussian. It is the most likely peak position that can be used for chemical identification. The standard deviation σ_k describes how concentrated the bell-shaped curve is around its centre μ_k. This is beneficial for calculating the associated confidence interval, which helps quantify the uncertainty of the peak identification. The mixing coefficient ω_k reflects the proportion of SEs emission from the corresponding energy excitation occurring at μ_k.

D. MATERIAL SURFACE CHEMICAL ANALYSIS

1) CHEMICAL INFORMATION MAP

Following the clustering results, we can decompose a SEHI image stack into K intensity maps or called layers of SEs emission. It helps to visualise the individual Gaussian components obtained. And each intensity map is associated with a specific chemical bonding type or a combination of multiple chemical bonds (or elements), which can be inferred from the spectral peak positions. The maps integrate the chemical information with the topography or morphology of the material sample surfaces, in favour of surface chemical analysis at micro- to nano-scale.

To generate the k-th intensity map L_k, we reallocate the intensity values over their corresponding spatial locations on the image plane. For the j-th tile that have associated i-th spectra peak being assigned to the k-th cluster, the corresponding emission intensity $I_{i,j,k}$ is picked. And then we locate this intensity $I_{i,j,k}$ at the j-th tile from which the spectra peak originates. In terms of the tiles without any associated spectral peaks labelled as k, zero-padding is carried out on these tiles. By doing so, the raw intensity map L_k is produced. Moreover, in order to make L_k more informative, we incorporate probabilistic confidence into $I_{i,j,k}$. Specifically, $I_{i,j,k}$ associated with the i-th spectra peak is weighted by its posterior probability $r_{i,k}$ which measures the uncertainty of the cluster assignment.

2) PHYSICAL INTERPRETATION OF CLUSTERS

The physical interpretability of the proposed AI framework is preserved via three key aspects of our methodology:

1) **Physically meaningful features:** Using the spectral peak positions as the feature for clustering, which directly correspond to chemical informatics (bonds or elements) using SEHI. As demonstrated in [69], [81], and [82], the peak position is leveraged as an informative spectral feature for the chemical identification of material specimens.
2) **Domain-specific constraints:** To ensure good practice of SEHI measurement, we have implemented:
(a) Energy range restrictions (0 to 7 eV) during the data collection, as SEHI provides reliable measurements in low energy ranges, as explained in [69].
(b) Peak intensity threshold for filtering weak peaks, potentially noise-induced, in the step of peak extraction. The threshold can be adjusted by end user, to adapt to different materials and instruments in practice;
(c) Optional user input of cluster numbers when the prior knowledge is available, when using the GMM

clustering method. It provides controlled clustering to meet different scenarios.

3) **Probabilistic clustering:** Benefiting from the GMM and DPGMM, their probabilistic nature offers robustness to noise and outliers in the experimental data.

Potential ambiguities in chemical assignment could occur, and it is reflected in the probability heatmap (shown in Fig. 8). These ambiguities primarily stem from instrumentation and measurement limitations, rather than from the clustering methods. They are mainly due to the common gap between theoretical predictions and experimental capabilities. When the spectral resolution in the experimental SEHI data is insufficient, it is difficult to distinguish between chemical bonds or elements with closely spaced theoretical energies. In such cases, it would be helpful to improve the spectral resolution of the instrument, or to consider cross-validation with other established spectroscopies. From an analytical viewpoint, end user can set a confidence threshold to indicate low-confidence assignments, based on the posterior probability of clustering results.

E. THEORETICAL ALIGNMENT FOR CHEMICAL IDENTIFICATION

This part presents how to align the Gaussian components, derived from experimental SEHI data, with the theoretical DOS model. The DOS model is retrieved from *the Materials Project* [53], [83], [84] through its open API to obtain theoretical spectra data as reference. Through comparison with the theoretical DOS model, we are able to reveal and rank the likely chemical elements or bonds associated with the Gaussian components.

To perform the theoretical alignment, one manual way is to match the experimental spectral peaks (from SEHI data), especially the predominant and well-separated peaks, with the theoretical peaks (from the DOS model). Matching these experimental spectral peaks with the theoretical peaks, by the peak positions, suggests the presence of corresponding chemical elements or bonds present in the material sample. However, such manual peak matching can lead to less accurate chemical identification.

To overcome this limitation, we propose an automated peak matching solution for materials chemical identification. This approach evaluates the similarity of the spectral peaks derived from experimental and theoretical data. Thus it can support further decision-making based on the similarity measure. To clarify, the theoretical spectra data is retrieved from the DOS model of likely chemical species such as specific chemical elements and bonds. The experimental spectra data is obtained using the proposed AI framework with SEHI data collected from the material specimen under test. The purpose of peak matching is to figure out the most likely chemical species or bonds associated with the Gaussian components obtained from the analysis.

Algorithm 1 Automated Peak Matching Workflow

```
Require:
(1) Input: Gaussian components obtained from SEHI ($C_k$), theoretical DOS candidate ($D_m$)
(2) Preliminary selection:
for each Gaussian $C_k$ with parameters $(\mu_k, \sigma_k)$ do
    Find $D_m$ with peak locations $\mu_D \in [\mu_k - 2\sigma_k, \mu_k + 2\sigma_k]$
end for
(3) DOS candidate evaluation:
for each Gaussian $C_k$ do
    Number of candidates $D_m$ satisfying peak range $\rightarrow N_D$
    if $N_D = 1$ then Yes                                  ▷ single candidate
        Assign $D_m$ to $C_k$ with probability 1
        $\rightarrow$ Jump to Output
    else No                                                ▷ multiple candidates
        (4) Data preparation for similarity check:
            Downsample DOS spectral data;
            Find peaks in downsampled DOS and SEHI data;
            Generate synthetic spectra data;
        (5) Similarity measure by cross-correlation
        (6) Decision making based on probability scores
    end if
end for
Output: Heat map visualising probability assignment of chemical elements/bonds to Gaussian components
```

Our approach addresses the challenge of accurately characterising material composition by combining preliminary peak matching with quantitative similarity assessment for probabilistic chemical assignment. The pseudocode implementation of the automated peak matching approach is presented in Algorithm 1. As explained by Fig. 2 and Algorithm 1, the proposed peak match workflow accomplishes six steps:

(1) *Data collection of experimental and theoretical spectra*;

(2) *Preliminary selection of potential candidates of the DOS model*: For each Gaussian component, find the promising candidates of the DOS models which have the spectral peaks located in the corresponding peak range (energy band). The energy range of a given Gaussian component is set as $\boldsymbol{\mu} \pm 2\boldsymbol{\sigma}$, corresponding to 95% confidence level. Only DOS models that meet this preliminary selection criteria are retained as promising candidates. Specifically, if no peaks in the DOS model fall within the energy range associated with a given Gaussian component, it indicates that this Gaussian component is unlikely to represent the relevant chemical species;

(3) *Check how many potential DOS candidates meet the preliminary criteria*;

(4) *If there are multiple candidates, need data preparation for running subsequent cross-correlation*: Details on the data preparation of synthetic spectra are provided in the following paragraph;

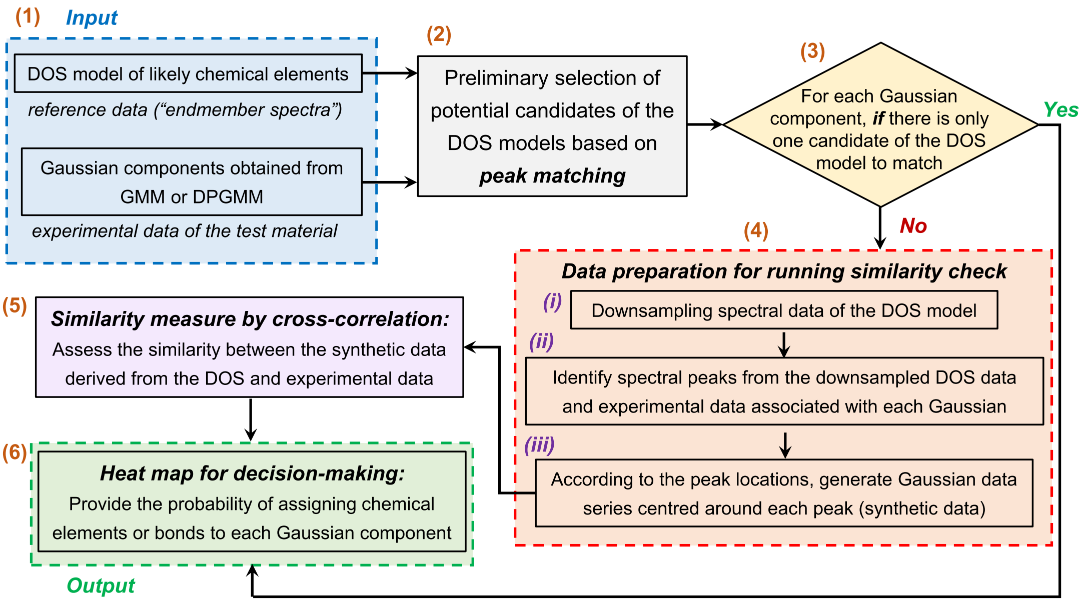

(1) Input
DOS model of likely chemical elements
reference data ("endmember spectra")
Gaussian components obtained from GMM or DPGMM
experimental data of the test material
(2)
Preliminary selection of potential candidates of the DOS models based on peak matching
(3)
For each Gaussian component, if there is only one candidate of the DOS model to match
Yes
No
(4)
Data preparation for running similarity check
(i) Downsampling spectral data of the DOS model
(ii) Identify spectral peaks from the downsampled DOS data and experimental data associated with each Gaussian
(iii) According to the peak locations, generate Gaussian data series centred around each peak (synthetic data)
(5)
Similarity measure by cross-correlation:
Assess the similarity between the synthetic data derived from the DOS and experimental data
(6)
Heat map for decision-making:
Provide the probability of assigning chemical elements or bonds to each Gaussian component
Output

FIGURE 2. Flow diagram representation of the automated peak match workflow for materials chemical identification. The reference spectra are retrieved from the theoretical DOS models of likely chemical species (specific chemical elements and bonds).

(5) *Similarity measure using cross-correlation*: Cross-correlation algorithm is used to determine the similarity between the experimental and theoretical data. The highest cross-correlation coefficient infers the most relevant DOS model among the DOS candidates, namely the most likely chemical components associated with the Gaussian components;

(6) *Provide the probability map for decision-making upon the likely chemical bonds or elements*: Heat map provides the probability of assigning a specified chemical component to a given Gaussian component. The probability tells how likely the chemical identification is, according to the cross-correlation coefficient calculated.

As stated above in the peak matching workflow step (4), synthetic spectra are generated as a superposition of several Gaussian distributions at identified peak centre locations. This approach effectively captures the peak position information while mitigating the influence of emission intensity (peak heights). It facilitates cross-correlation for measuring spectra similarity merely based on the peak locations, avoiding the effect of the peak height variations. The details of data preparation in (4) are: *(i)* Usually the theoretical spectra data derived from the DOS model have higher energy resolution than those from the SEM in practice. Thus first, we downsample the theoretical spectra data to match the energy resolution and range of the experimental spectra data; *(ii)* Identify spectral peaks in both theoretical (after downsampling) and experimental spectra data; *(iii)* Utilising the identified spectral peak locations from theoretical and experimental data, a group of Gaussian data series is generated and superposed, called synthetic data here. Specifically, the peak locations serve as the mean of the standard Gaussian distribution, with an associated standard deviation set as 1.

To clarify, the Gaussian kernel width in the synthetic spectra was chosen as 1 here, to ensure robust cross-correlation matching for accommodating potential calibration misalignment between experimental and theoretical spectra. For best practice across diverse applications, we recommend tuning the kernel width in a typical range from 0.5 to 1.5 eV. The selection should consider: (a) instrument-related parameters, particularly the energy resolution of the experimental spectra; (b) the expected degree of calibration misalignment between experimental and theoretical data; (c) the peak spacing within the material spectral signatures. While larger kernel widths improve tolerance to the spectral misalignment, they may reduce sensitivity for distinguishing closely spaced peaks. Overall, the automated multi-peak matching strategy within a constrained energy range, combined with a probabilistic confidence assessment for decision-making, enhances the reliability of chemical assignments in our proposed framework.

IV. CASE STUDY

A. BACKGROUND AND PROBLEM DESCRIPTION

We adopt the experimental SEHI data from [49] for evaluating the performance of the proposed framework. The study from [49] aimed to understand the underlying relationships between carbon and metal species on the nano-scale by SEHI, which can boost the optimization and fabrication of relevant key applications such as catalytic materials. The complex metal alloy (palladium & silver) and carbon

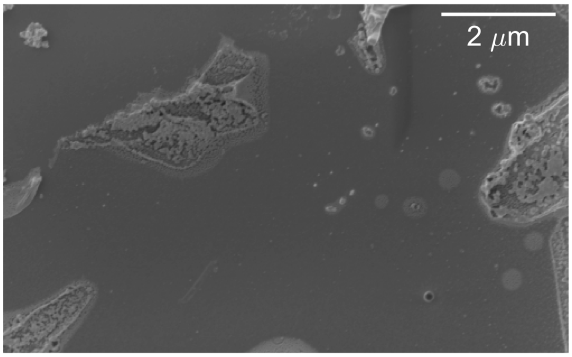


FIGURE 3. **SEHI micrograph of a thick Pd-Ag-C metal film (named as "*A*6 PdAg thick" in [55]). Scale bar denotes 2 μm.**

films, abbreviated as Pd-Ag-C, were printed by University of Liverpool and the raw SEM images were collected via a Helios Nanolab G3 UC microscope [49]. The microscope instrument was operated at a low-voltage energy ranging from 0 to 7 eV to capture chemical information better, with measurements taken in increments of 0.24 eV. Accordingly, this results in an energy resolution, R_{eV}, of 0.24 eV for the collected spectral SEM images. Details about material samples preparation, and experimental settings can be found from [47] and [49]. The study [49] provides four SEHI datasets collected from different Pd-Ag-C film specimens with varying film thickness and surface roughness.

Fig. 3 shows a high-resolution surface micrograph acquired by SEHI, depicting a thick Pd-Ag-C film sample with relatively smooth surface morphology. We mainly investigated this thick Pd-Ag-C specimen as an example to demonstrate this framework. Here we first examine whether different selections of ROIs can affect associated spectra curves. It can help understand the limitations caused by manual selection and conventional spatial averaging of ROIs. Fig. 4 (a) depicts the result when the whole FOV is chosen, representing the "global" information. In contrast, examples of the spectra curves extracted from different small tiles in 3 pixels × 3 pixels are given in Fig. 4 (b)-(d), denoting the "local" information. In the "global" case, the spectrum curve clearly exhibits only two predominant peaks, located at 1.99 eV and 5.29 eV, shown in Fig. 4 (a). These two spectral peaks well agree with the two peaks observed from the earlier work [49]. In comparison, the peaks in the "local" examples are more diverse, with varying peak locations and peak numbers in Fig. 4 (b)-(d). It suggests that by spatially averaging over larger areas, manual selection of the ROIs (or simply using the whole FOV) would limit the comprehensiveness and accuracy of SEHI data analysis results.

B. DATASETS AND DATA PREPARATION

The SEHI datasets used in this study are publicly available in our data repository [55]. It [55] contains SEHI data associated with four Pd-Ag-C specimens. For each specimen, 30 sequential SEM image slices were collected under the low-energy range of 0 to 7 eV. As mentioned above, we mainly analysed the SEHI data associated with the thick specimen which is named as "*A*6 PdAg thick" in [55]. The raw image slices are first aligned using a template matching algorithm [72]. The aligned SEHI data cube has dimensions of 1517 pixels × 933 pixels × 30 spectra. The aligned image stack is then partitioned into tiles in 3 × 3 pixels along the image plane. It yields a total of 157,055 tiles, from which 522,098 spectral peaks are identified from 157,055 spectral curves.

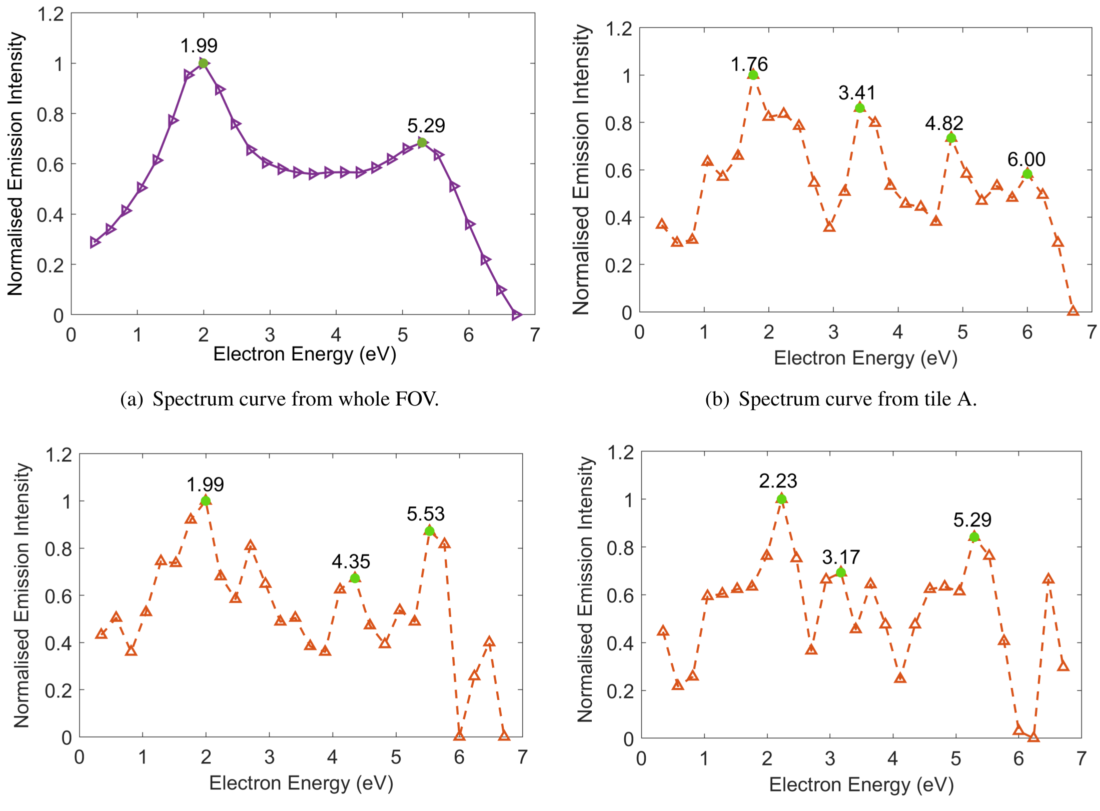
These 522,098 spectral data points are split into training, validation, and test datasets, with 70% (365,469 data points) used for training, 15% (78,315 data points) for validation, and 15% (78,314 points) for testing, respectively. It is not necessary to train a GMM model using the entire training dataset. To improve computational efficiency for the downstream clustering task, the training dataset is further subsampled, as the computational costs of GMM and DPGMM scale with the number of data points. Here, we have a distribution-preserving subsampling strategy with a reduction ratio of approximately 0.2. Our objective is to preserve the probability distribution of the whole training dataset while reducing the number of data points. Inspired by stratified random sampling [85], we first construct a probability histogram by dividing the whole training dataset into groups (or bins) based on their energy values. Then, a certain random sample can be drawn separately from the bins. As shown in Fig. 5, the subsampled dataset preserves a probability distribution closely resembling that of the entire training set while efficiently reducing the dataset to 73,045 data points.

C. IMPLEMENTATION DETAILS AND EVALUATION METRICS

After image tiling, two clustering methods, GMM and DPGMM, are adopted to model the multi-peak distribution of the spectral peaks. The implementations for both methods are detailed below.

1) GMM IMPLEMENTATION

For scenario where the cluster number is not predefined, the GMM is implemented alongside a model selection strategy to determine the optimal number of clusters (K). This process involves the elbow method to analyse the relative changes in BIC values [78], [86], complemented by silhouette score analysis [87] for cluster quality evaluation. Fig. 6 (a) depicts the relative changes in BIC values as the number of clusters increases. A clear "elbow" is observed when K equals four, indicating that beyond this point, the model fit does not improve significantly with more clusters. In addition to BIC, silhouette score (s_n) and the standard deviation of Gaussian components (σ_k), are also used to determine the optimal K. The silhouette score measures how similar a data point is to its assigned cluster (cohesion) compared to other clusters (separation) [87]. s_n ranges from -1 to 1, where



(a) Spectrum curve from whole FOV.

(b) Spectrum curve from tile A.

(c) Spectrum curve from tile B.

(d) Spectrum curve from tile C.

FIGURE 4. Examples of spectra curves extracted from the material sample shown in Fig. 3: (a) produced from the whole FOV, (b)-(d) generated from individual, distinct 3 × 3 tiles. This illustrates the influence of manually selecting different ROIs.

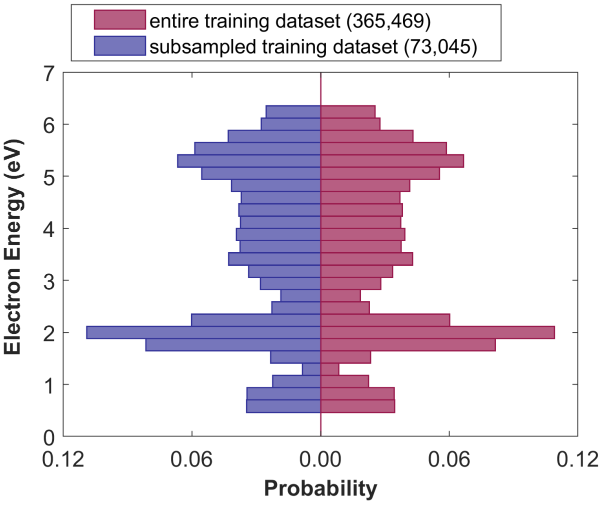


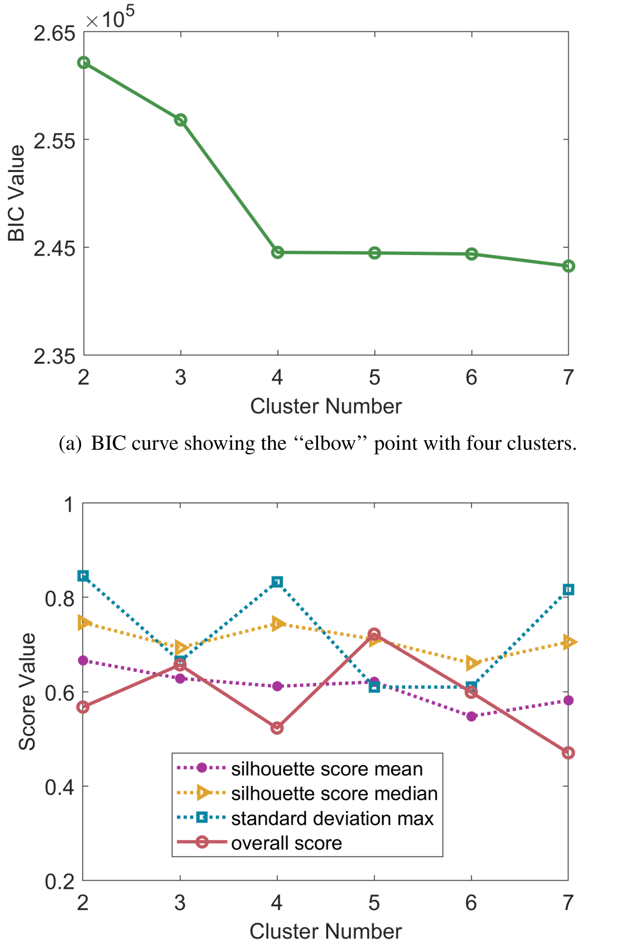
FIGURE 5. Comparison between the probability distributions of subsampled training dataset (73,045 data points) and entire training dataset (365,469 points). By visualising side-by-side probability histograms, it demonstrates our distribution-preserving subsampling strategy.

higher values indicate well-separated clusters, while negative values suggest misclassification. It is often used to determine the optimal number of clusters [88], particularly as internal evaluation metrics without requiring explicit labels. In this study, the silhouette scores, s_n, are calculated using Euclidean distance.

We introduce an overall score, $s^{overall}$, as defined in (13). It is used to determine the optimal number of clusters K, in the absence of ground truth labels. This overall score, $s^{overall}$, combines the silhouette score s_n analysis with a penalty term based on σ_k. A higher score, $s^{overall}$, indicates better clustering performance. The optimal number of clusters is selected as the K that maximizes $s^{overall}$ after the "elbow" point observed in the BIC curve. As shown in (13), $s^{overall}$ combines: (1) the average and median silhouette scores across all data points as clustering quality measures, (2) a penalty term based on ${\sigma_k}^{max}$, the maximum standard deviation of all Gaussian components. This penalty term is motivated by specific application considerations, where a larger standard deviation is less preferred because it reflects a broader chemical range for the associated Gaussian component, showing greater uncertainty in the chemical identification.

$$s^{overall} = s^{avg} + s^{med} - \beta {\sigma_k}^{max}. \tag{13}$$

where s_n denotes the silhouette score of each data point, and is calculated using Euclidean distance. s^{avg} and s^{med} represent the average and median of s_n across all N data points, respectively. ${\sigma_k}^{max}$ is the maximum standard deviation



(a) BIC curve showing the "elbow" point with four clusters.

(b) Average and median silhouette scores, penalty term based on σ_k, and overall score.

FIGURE 6. BIC values, silhouette scores and the overall score for finding the optimal number of clusters for GMM implementation. The optimal number of clusters is selected as 5 which maximises the overall score $s^{overall}$ after the "elbow" point.

among all K Gaussian components. β is a hyperparameter that controls the weight of the penalty term $\sigma_k{}^{max}$. $\beta \in [0, \infty)$ and is set to 1 in this study.

Fig. 6 (b) illustrates the variation in the relevant scores as the number of clusters changes. Consequently, the optimal number of clusters, K, was set as 5, which gives the maximum $s^{overall}$ after the elbow point in the BIC curve. Other parameters set in GMM include: the maximum number of iterations as 1500, the tolerance of objective function termination as 10^{-7}, full covariance matrices, initial value setting method as the K-means++ algorithm, a small regularization value ranging from 10^{-5} to 10^{-2}, and 10 repetitions of the EM algorithm using a new set of initial values. This small regularisation term is added to the covariance matrices to avoid numerical instability. It is set to 5.55×10^{-4}, derived from $(0.1 \times R_{eV})^2$, where R_{eV} represents the energy resolution of the discretised spectral data, which is 0.24 eV in this case.

2) DPGMM IMPLEMENTATION

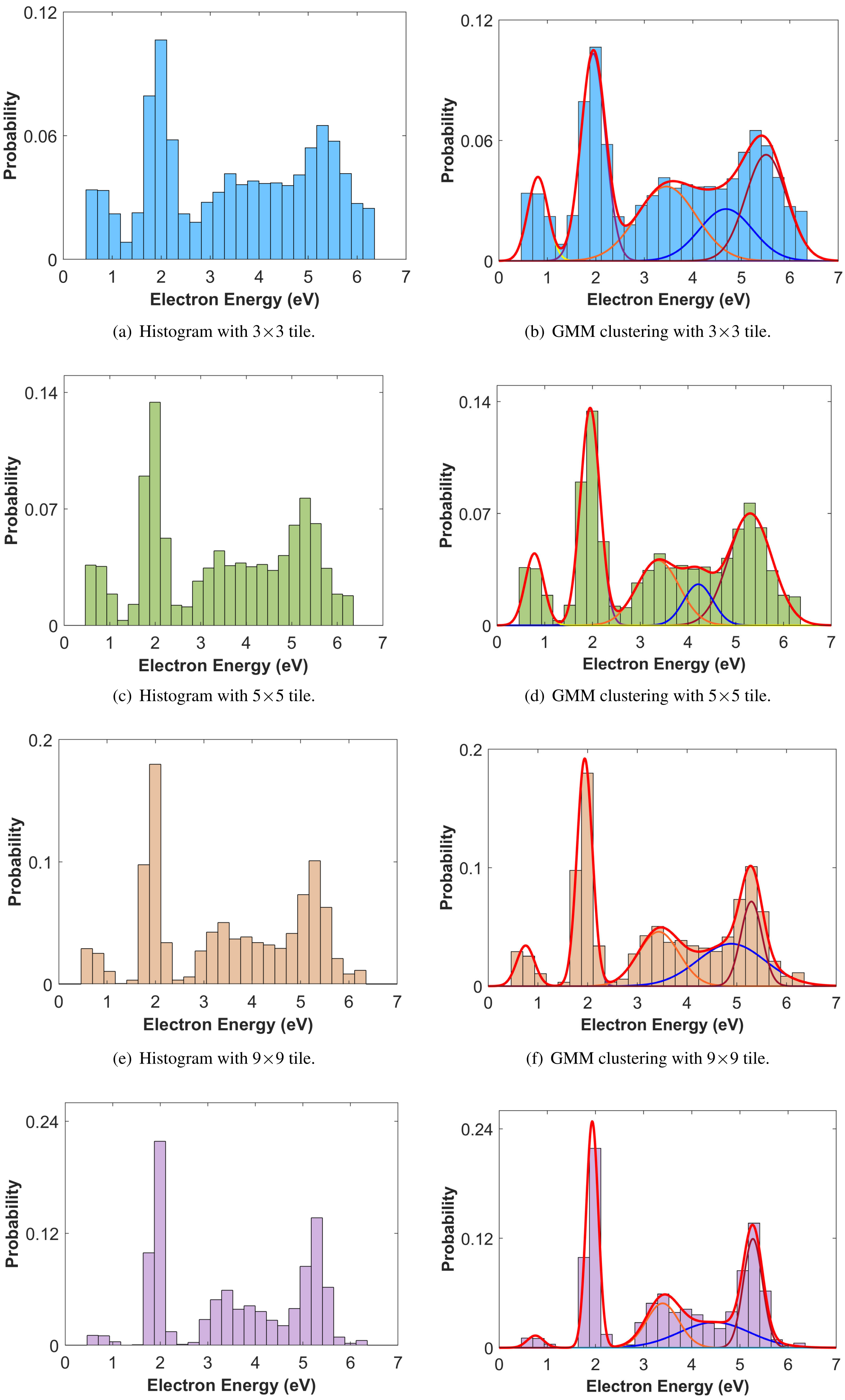
TABLE 1. Evaluation of GMM fitting on different datasets.

Data	Data Points	s^{avg}	s^{med}	D_{KL}
subsampled	73,045	0.6206	0.7112	0.0269
training	365,469	0.6207	0.7112	0.0269
validation	78,315	0.6221	0.7115	0.0275
test	78,314	0.6196	0.7091	0.0282
all	522,098	0.6207	0.7110	0.0274

The primary advantage of the DPGMM algorithm is the cluster number is inferred from a DP prior automatically. The concentration parameter in the DP prior, α, governs the probability of yielding new clusters, which means that a greater α leads to more clusters. α is typically initialised to 1 for weakly informative prior [66], [68]. Here the hyperparameters α and λ_0 are assigned small values to impose weak priors, which allow flexibility in the cluster parameters estimation. The DPGMM algorithm is fine-tuned by changing the hyperparameters α and λ_0. Based on our experimental evaluation (see Section V-C3 and Table 8 for detailed results), we choose α as 1, and λ_0 as 0.6. Regarding other hyperparamters in the NIW prior, the parameters ν_0 is initialised to 1, $\boldsymbol{S}_0$ as the identity matrix, whilst $\boldsymbol{\mu}_0 = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i$, with $N = 157,055$ represents all sample data points in this case.

Variability across runs is a well-characterised and inherent property of the DPGMM, arising directly from its Bayesian non-parametric construction via the CRP [67]. This variability is primarily influenced by the concentration parameter α. A higher α promotes the creation of new clusters, while a lower α encourages the data points to join existing groups. Consequently, different runs may yield different cluster labels and slightly different numbers of components, particularly for low-probability clusters. It is important to note that the observed variability is not solely a function of the DPGMM algorithm but also depends on the underlying distribution and structure of the data. In our work, the DPGMM is executed over 5 independent runs, with a maximum of 300 iterations per run that ensured convergence in all runs. The final model is selected as the one with the smallest negative log-likelihood *NLL*. In addition, based on the mixture weights ω_i, negligible Gaussian components with weights below 0.5% are discarded from the results, to reflect the dominant distribution and structure of the spectral data. To adapt the DPGMM method to different data structures and application needs, we advise end users to tune key hyperparameters, such as the concentration parameter α and the maximum number of iterations, and to perform multiple runs as a best practice.

In addition to the silhouette score, we evaluate the performance of the trained Gaussian mixture model with the Kullback–Leibler (KL) divergence [89]. While the silhouette score offers an internal measure of clustering quality without requiring ground truth, it does not reflect how well the GMM model captures the true distribution of spectral peaks. KL divergence complements this by comparing the distribution of real spectral peak data with the estimated



(a) Histogram with 3×3 tile.

(b) GMM clustering with 3×3 tile.

(c) Histogram with 5×5 tile.

(d) GMM clustering with 5×5 tile.

(e) Histogram with 9×9 tile.

(f) GMM clustering with 9×9 tile.

(g) Histogram with 15×15 tile.

(h) GMM clustering with 15×15 tile.

FIGURE 7. Comparison of the probability histograms of spectral peaks (left) and associated GMM clustering results (right) when changing tile sizes.

TABLE 2. Evaluation of DPGMM fitting on different datasets.

Data	Data Points	s^{avg}	s^{med}	D_{KL}
subsampled	73,045	0.6318	0.6981	0.0273
training	365,469	0.6318	0.6981	0.0265
validation	78,315	0.6333	0.6965	0.0273
test	78,314	0.6314	0.6954	0.0285
all	522,098	0.6320	0.6975	0.0274

distribution from the trained GMM model. To adopt KL divergence into our study, we first sampled synthetic data from the trained GMM, matching the sample size of the real spectral dataset, to ensure a statistically consistent comparison. Then the generated data points were discretised by rounding to the nearest value in real spectral data, aligning with the resolution of the real data. This ensured identical binning schemes for the synthetic and real datasets. Finally, we obtained probability distributions from the binned datasets and computed the KL divergence to quantify the difference between the true and modelled distributions. A lower KL divergence indicates a closer alignment between the trained model and the real data distribution, serving as a complementary external measure of model performance.

V. RESULTS AND VALIDATION

A. RESULTS WITH GMM AND DPGMM

1) MODEL GENERALISATION EVALUATION

The clustering models are trained on the subsampled training dataset. To examine the model generalisation capability, the performance is evaluated across five datasets: the subsampled training dataset, the entire training dataset, the validation data, the test data, as well as all data. Performance metrics include the average silhouette score s^{avg}, the median silhouette score s^{med}, and the KL divergence D_{KL} which quantifies the difference between the real spectral data distribution and that approximated by the trained GMM model.

Tables 1 and 2 summarise the performance evaluation across different datasets by using GMM and DPGMM, respectively. The results demonstrate strong model generalisation capabilities for both methods, with consistent performance observed across all evaluated datasets. The nearly identical evaluation performance between the subsampled training and the entire training datasets validates the effectiveness of the distribution-preserving subsampling strategy. Both models, trained with GMM and DPGMM, show no significant performance degradation when evaluated on the validation and test datasets compared to the training data. It demonstrates the good model generalisation to maintain performance on unseen data.

2) RESULTS INTERPREATION

Fig. 7 (a) displays the probability distribution of spectral peak positions extracted after tiling in 3 × 3 pixels. Fig. 7 (b) shows the GMM results, where the red contour represents the estimated distribution as a superposition of five Gaussian components. Table 3 presents the results obtained from GMM and DPGMM methods. After fine-tuning, the DPGMM yields a five-component Gaussian mixture. This enables a direct performance comparison between the GMM and DPGMM results with the same number of components in the mixture. As shown in Table 3, these five Gaussian components are denoted by $C_1,\ldots,C_5$. It can be seen that the results from GMM are quite close to that of DPGMM, especially for C_1, C_2, and C_5. The reason is that the peak regions related to C_1, C_2, and C_5 are well-separated, clearly seen in the histogram Fig. 7 (a). Whereas, distinguishing peaks between 3-5 eV (C_3, C_4) is challenging, due to the electron contamination effects in SEHI measurements and spectral overlap. Among the model parameters, mean $\boldsymbol{\mu}$ is used for chemical identification. The $\boldsymbol{\sigma}$ and $2\boldsymbol{\sigma}$ parameters are utilised to quantify the uncertainty associated with 68%, 95% confidence level of the mean $\boldsymbol{\mu}$. The related uncertainty analysis on mean $\boldsymbol{\mu}$ is given in Table 3. The component mixing proportion $\boldsymbol{\omega}$ reveals the prevalence of SEs emitted from the corresponding chemical elements or bond types.

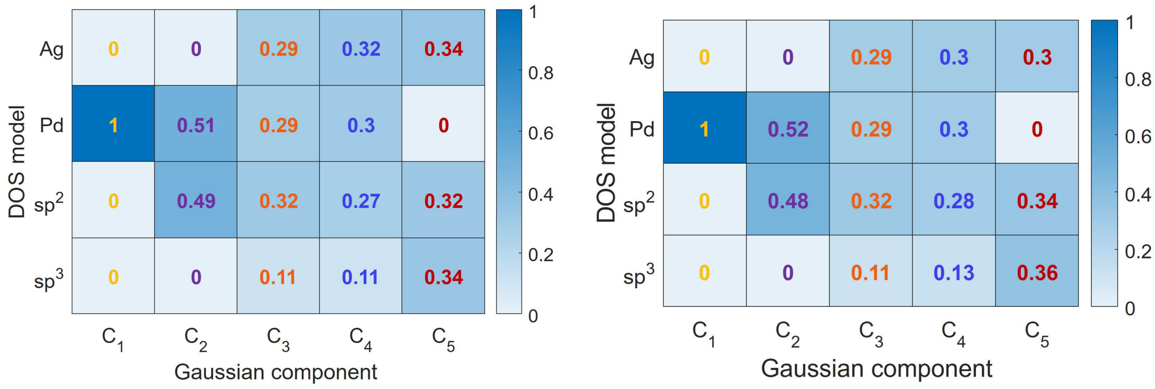
TABLE 3. Gaussian components obtained from GMM and DPGMM after tiling in 3×3. Numbers highlighted in bold represent the mean μ. The associated uncertainties are quantified using σ (68% confidence interval) and 2σ (95% confidence interval).

Gaussian Component	Clustering Method	Mean (eV) 68% confidence	Mean (eV) 95% confidence
C_1	GMM	**0.81** ± 0.21	**0.81** ± 0.42
	DPGMM	**0.81** ± 0.21	**0.81** ± 0.41
C_2	GMM	**1.95** ± 0.25	**1.95** ± 0.50
	DPGMM	**1.95** ± 0.25	**1.95** ± 0.50
C_3	GMM	**3.35** ± 0.55	**3.35** ± 1.10
	DPGMM	**3.25** ± 0.51	**3.25** ± 1.02
C_4	GMM	**4.50** ± 0.60	**4.50** ± 1.19
	DPGMM	**4.36** ± 0.66	**4.36** ± 1.31
C_5	GMM	**5.49** ± 0.43	**5.49** ± 0.86
	DPGMM	**5.49** ± 0.43	**5.49** ± 0.86

Conventional data analysis by analysing of the whole FOV reveals only two dominant peaks, located at approximately 2 eV and 5.3 eV, as shown in Fig. 4 (a). In contrast with the earlier study [49], the image tiling not only allows for recognising the two leading peaks, but also captures three weaker peaks, illustrated by the histogram in Fig. 7 (a) and (b). This highlights the advantage of the proposed framework, which offers more accurate chemical analysis compared to conventional methods that rely on manual selection of ROIs.

B. CHEMICAL ANALYSIS AND ACCURACY EVALUATION

In this case study, the relevant chemical elements and bonds of the complex Pd-Ag-C films being investigated include silver Ag, palladium Pd, sp^2 carbon and sp^3 carbon, from



(a) Probability heat map obtained from GMM.

(b) Probability heat map obtained from DPGMM.

FIGURE 8. Probability heat map which depicts how likely to assign the chemical elements or bonding types to the Gaussian components C_1 to C_5. (a) and (b) are obtained from GMM and DPGMM, respectively.

the expert's prior knowledge. Accordingly, four relevant DOS models are retrieved from the Materials Project data [53] and used as references to deduce the associated materials with the five Gaussian components obtained from the proposed AI framework.

Fig. 8 gives the probability heat map using the proposed peak matching approach. Following the peak matching step (1) as described in Section III-E, the preliminary selection is conducted by comparing the spectral peak locations in the four reference DOS models with the five Gaussian components. Specifically, a probability of 0 indicates that no spectral peaks in a reference DOS model fall within the energy range for a given Gaussian component. For instance, for the Gaussian component C_1, only one DOS candidate, Pd, meets the preliminary selection criteria, thus being assigned a probability of 1. Following the preliminary selection, for C_2 to C_5, multiple DOS candidates meet the preliminary criteria. Thus further peak matching steps (4) and (5) are performed to obtain similarity measures. Then the normalized cross-correlation coefficient is used to calculate the probability of assigning chemical species to the Gaussian components (see Section III-E for method details). Given by Fig. 8, the probability outcomes from GMM and DPGMM are relatively close to each other, particularly for the Gaussian components C_1, C_2 and C_3.

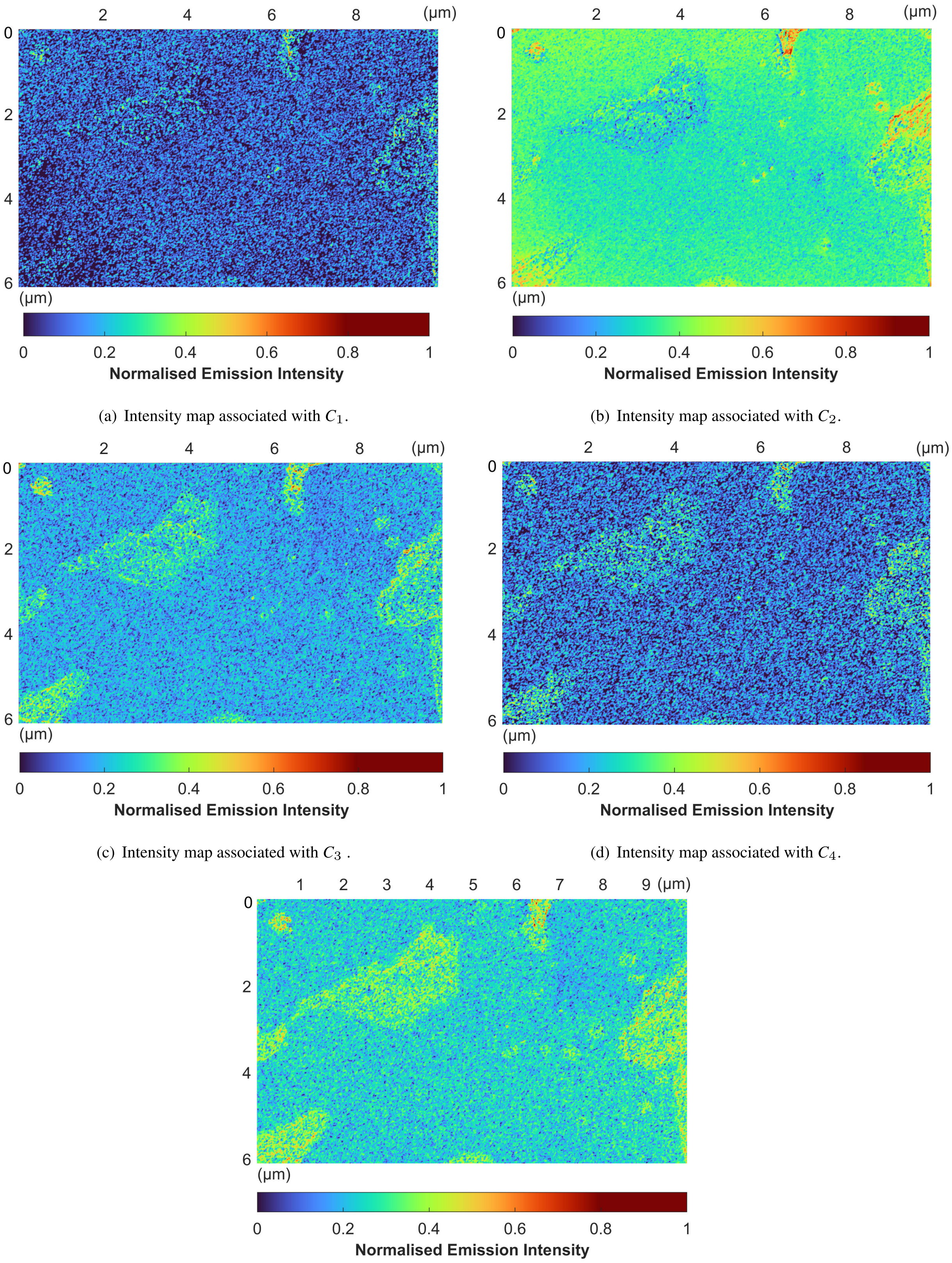
According to the probability heat maps given in Fig. 8, the most likely chemical elements associated with each Gaussian component are, C_1: Pd, C_2: Pd and sp^2 carbon, C_3: sp^2 carbon, Ag and Pd, C_4: Ag, Pd and sp^2 carbon, C_5: sp^3 carbon, Ag and sp^2 carbon. Accordingly, Fig. 9 (a)-(e) shows the five informative chemical maps associated with five Gaussain components (see Section III-D1 for method details). These maps allow the visualisation of the surface morphology, combined with the materials chemistry informatics. This holds significant scientific value for advancing our fundamental understanding of structure-property relationships in materials, and offering industrial benefits by enabling advanced precision manufacturing of complex functional materials. For instance, through the comparison between the maps associated with C_1 and C_5, given in Fig. 9 (a) and (e), we can observe the heterogeneous spatial distribution of two metals, Pd and Ag. This is valuable for analysing metal alloy composition, which can reveal structure-property-process links to guide the film printing process.

The accuracy of chemical identification is assessed with reference to the theoretical DOS models. Here the peak position obtained from experimental SEHI data is denoted as $\mu^E_{m,k}$, whilst that derived from the theoretical DOS model is represented as $\mu^T_{m,k}$. Accordingly, their relative errors, denoted as $\Delta\mu_{m,k}$, can be calculated in percentage, with respect to the theoretical values $\mu^T_{m,k}$. The relative error, $\Delta\mu_{m,k}$, quantifies the accuracy of assigning the m-th chemical bonding types (or elements) to the k-th Gaussian component. Then we quantify the overall error of the k-th Gaussian component in chemical bonding identification. By comparing the spectral peak locations of the relevant experimental and theoretical data, the overall error is computed as follows

$$\Delta\mu_k = \sum_{m=1}^{N_k} h_{m,k}\Delta\mu_{m,k} = \sum_{m=1}^{N_k} h_{m,k}\frac{\mu^E_{m,k} - \mu^T_{m,k}}{\mu^T_{m,k}}100\% \tag{14}$$

where $\Delta\mu_k$ denotes the overall error of the k-th Gaussian component (C_k) for chemical bonding identification. $h_{m,k}$ is the probability of assigning m-th chemical bond to C_k. $\Delta\mu_k$ is the weighted sum of relevant errors $\Delta\mu_{m,k}$ associated with C_k, whilst N_k is the number of likely chemical bonds with C_k.

Since each Gaussian component C_k can result from a mixture of several chemical species, the number of relevant likely chemical species is denoted as N_k. The probability $h_{m,k}$ is utilized as the weight when summing all relevant errors $\Delta\mu_{m,k} (m = 1, \cdots, N_k)$ associated with C_k. As explained by (14), the overall error $\Delta\mu_k$ is calculated, used for the



(a) Intensity map associated with C_1.

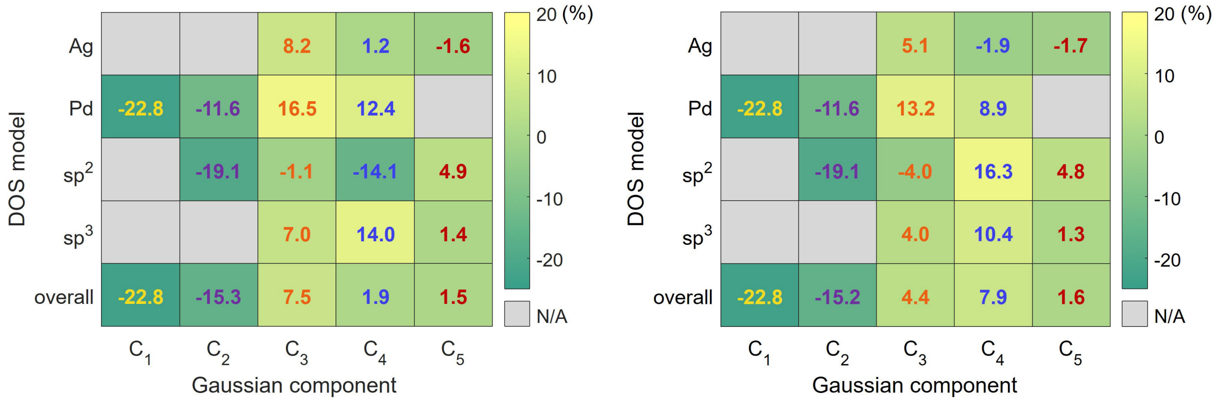
(b) Intensity map associated with C_2.

(c) Intensity map associated with C_3 .

(d) Intensity map associated with C_4.

(e) Intensity map from C_5 .

FIGURE 9. Individual intensity maps associated with Gaussian components C_1-C_5, derived from the film shown in Fig. 3. These Gaussian components are most likely, C_1: Pd, C_2: Pd and sp^2-carbon, C_3: sp^2-carbon, Ag and Pd, C_4: Ag, Pd and sp^2-carbon, C_5: sp^3-carbon, Ag and sp^2-carbon.



(a) Accuracy heat map obtained from GMM.

(b) Accuracy heat map obtained from DPGMM.

FIGURE 10. Accuracy heat map depicting the relative errors (calculated from (14)) in assigning the chemical bonding or elements to the corresponding Gaussian components. Note that the negative sign indicates that the SEHI experimental results are smaller than the theoretical DOS values. (a) and (b) are obtained from GMM and DPGMM, respectively.

overall accuracy evaluation when assigning the chemical species to the Gaussian component C_k.

Fig. 10 depicts the outcomes of accuracy evaluation for assigning the potential four chemical elements or bonds (Ag, Pd, sp^2 carbon and sp^3 carbon) to the five Gaussian components $(C_1, \ldots, C_5)$ obtained from GMM and DPGMM. The overall errors are computed by combining the probability heat map (see Fig. 8) with the relevant errors of each likely chemical bond or element, as given by (14). In other words, the overall errors are the weighted sum of the individual errors by column-wise. The error results are shown in percentage. As observed in Fig. 10, two methods, GMM and DPGMM, achieve comparable performance in terms of the overall accuracy with reference to the DOS model. Besides, it can be seen that except the Gaussian component C_1, the errors of other four Gaussian components are in the range of $\pm 15\%$, which well demonstrates the good capability and accuracy of this AI framework for chemical identification. The error of C_1 is larger than those of $C_2, \ldots, C_5$, since C_1 is associated with the lowest spectral peak in the energy range. This results from higher signal noises in the lower energy range, due to the instrumentation limitations in the SEs detector. From the Gaussian components C_1 to C_5, as the energy values of their spectral peaks become higher, the corresponding relative errors are getting lower.

C. ABLATION STUDIES

1) EFFECT OF TILE SIZE

In principle, when a finer tile is deployed, more tiles are produced from a single SEHI data cube. It increases the data points of spectral peaks that constitute the distribution. This makes the distribution statistically representative and meaningful to meet the assumption of Gaussian distributions. Nevertheless, very fine tiles would become sensitive to signal noise and other sources of uncertainty, which can reduce the reliability and stability of data analysis. Due to these concerns, we investigate how the tile size affects the distribution of spectral peaks and the subsequent probabilistic clustering results.

In comparison with the small tile of 3 × 3 pixels applied above, three larger sizes, 5 × 5, 9 × 9, and 15 × 15 in pixels, are examined. The GMM algorithm is implemented with a fixed cluster number equal to five. And details of the hyperparametes settings in GMM and DPGMM are provided in Section IV-C. Fig. 7 depicts the probability histograms and the model fitting results by GMM. From the histograms, through comparisons of the four tile sizes, the good agreement of two outstanding peaks, at around 2 eV and 5.3 eV, can be found. With larger tile sizes, the leading portions of these two predominant peaks are more noticeable. It is because less "local" information is accounted into the peak distribution. In contrast, with the fine tile of 3 × 3 pixels, more small peaks, appearing below 1 eV and ranging in 3–5 eV, are distinguished. As shown in Fig. 7, larger tile sizes lead to spectral peak suppression due to spatial averaging. This key finding clearly demonstrates a fundamental limitation of the conventional manual workflows that utilise the whole FOV or large ROIs for material analysis with microscopy. We find that such spatial averaging can obscure fine-grained chemical heterogeneity, and suppress weak spectral signatures, particularly problematic when examining subtle chemical variations at the sub-micron and nano-scale.

According to the distribution histograms given by Fig. 7, we recommend using tile sizes ranging from 3×3 to 9×9 pixels to effectively capture "local" information and subtle chemical variations, in such case of the SE image in size of 1517 × 933 pixels. It is worth noting that the optimal choice of tile size is application-specific and involves a trade-off between several key considerations. In practical applications, the selection of tile size is highly application-dependent. This choice is influenced by multiple factors, including microscopy parameters (e.g., horizontal field of view), material properties under analysis (e.g., physical scale of the target chemical features required for analysis, chemical information depth), and experimental conditions (e.g., noise characteristics). In addition to these application needs, selecting an appropriate tile size also involves trade-offs between spatial resolution, spatial size, noise sensitivity, computational efficiency, and spectral fidelity (e.g., preserving weak spectral signatures). We acknowledge that a comprehensive study on the tile size is a promising direction for future work, particularly for validating and optimising the framework across diverse material systems. To facilitate reproducibility and further experimentation, the code for the AI-assisted tool and the SEHI dataset are publicly available from [55] and [56]. Tables 4 and 5 summarise the model parameters computed by GMM and DPGMM. The results are collected under four levels of tile sizes, detailing about the associated tile number, data points for training, and computation time. By comparison between GMM and DPGMM results, good agreement in the estimated model parameters can be found. While DPGMM requires much longer computation time than GMM, DPGMM slightly outperforms in terms of the robustness to the tile size variations. This point is evident for the mean energy values μ, which are important for accurate chemical identification. In contrast, the component proportion ω and standard deviation σ show more variability across tile sizes. This is reasonable because the tile size directly determines how much the "local" information is captured, thus affecting the observed peak distributions as seen in Fig. 7. For large-scale industrial deployment, the DPGMM method could be accelerated by parallel processing of independent tiles across multiple CPU cores. Besides, GPU acceleration could further improve processing speed, making real-time analysis feasible for high-throughput applications.

TABLE 4. GMM clustering results with varying tile sizes.

Tile	Tile	Data	Comput.	Model	Gaussian Component				
Size	Number	Points	Time[1]	Param.[2]	C_1	C_2	C_3	C_4	C_5
3*3	157,055	73,045	7 mins	ω (%)	9.24	27.91	19.55	18.27	25.05
				μ (eV)	**0.81**	**1.95**	**3.35**	**4.50**	**5.49**
				σ (eV)	0.21	0.25	0.55	0.60	0.43
5*5	56,358	25,653	4 mins	ω (%)	9.28	29.43	19.13	8.37	33.80
				μ (eV)	**0.78**	**1.95**	**3.37**	**4.22**	**5.30**
				σ (eV)	0.19	0.20	0.44	0.31	0.45
9*9	17,304	15,174	3 mins	ω (%)	6.55	31.47	20.06	25.71	16.21
				μ (eV)	**0.75**	**1.94**	**3.43**	**4.89**	**5.29**
				σ (eV)	0.18	0.15	0.41	0.68	0.21
15*15	6,262	13,150	2 mins	ω (%)	2.40	33.28	17.49	21.56	25.27
				μ (eV)	**0.75**	**1.93**	**3.39**	**4.47**	**5.27**
				σ (eV)	0.17	0.13	0.34	0.73	0.20

[1] Matlab code running on a laptop with Intel Core i7-11800H processor.
[2] Model parameters include component mixing proportion ω (%), mean μ (eV) and standard deviation σ (eV).

TABLE 5. DPGMM clustering results with varying tile sizes.

Tile	Tile	Data	Comput.	Model	Gaussian Component				
Size	Number	Points	Time[1]	Param.[2]	C_1	C_2	C_3	C_4	C_5
3*3	157,055	73,045	3.6 hrs	ω (%)	9.22	27.93	15.23	22.35	25.11
				μ (eV)	**0.81**	**1.95**	**3.25**	**4.36**	**5.49**
				σ (eV)	0.21	0.25	0.51	0.66	0.43
5*5	56,358	25,653	1.4 hrs	ω (%)	9.29	29.11	21.70	11.43	28.46
				μ (eV)	**0.79**	**1.95**	**3.46**	**4.52**	**5.37**
				σ (eV)	0.19	0.20	0.53	0.54	0.43
9*9	17,304	15,174	45 mins	ω (%)	6.55	31.39	12.62	29.97	19.42
				μ (eV)	**0.75**	**1.94**	**3.32**	**4.52**	**5.29**
				σ (eV)	0.18	0.15	0.32	0.84	0.24
15*15	6,262	13,150	35 mins	ω (%)	2.40	33.27	19.32	19.96	24.96
				μ (eV)	**0.75**	**1.93**	**3.42**	**4.55**	**5.27**
				σ (eV)	0.17	0.12	0.36	0.71	0.20

2) FEATURE ANALYSIS

To investigate the significance of spectral peak position as the sole feature, we compare it with a 2D feature combining peak position and peak height. As shown in Fig. 11, the bivariate histogram visualises the joint probability distribution of peak position and peak height. It demonstrates that peak height varies significantly at the fixed peak positions. This observation reveals the variability of peak height which can be strongly influenced by material surfaces roughness, experimental instruments and conditions, as discussed in Section III-B. By comparison with Fig. 7 (a), it can be found that using peak position alone leads to more distinct and separable clusters than the combination with peak height.

We further evaluate the impact of adding the peak height as an additional feature through a feature ablation study. The clustering quality is assessed using silhouette scores as a quantitative measure, in the absence of ground truth labels. The distribution of the silhouette score s_n across all N data points is characterised by four statistics, which include the average value s^{avg}, first quartile s^{Q1}, median s^{med}, and third quartile s^{Q3}. Table 6 gives the clustering performance obtained from the GMM approach with a fixed cluster number K of 5 for comparability. It compares three feature sets, including peak position alone, peak position with raw peak height, and peak position with normalised peak height ranging from 0 to 1. The results show that incorporating peak height as a second feature leads to lower silhouette scores, compared to using peak position alone. It suggests that including peak height as an additional feature reduces the clustering quality. Thus, only the peak position is used as the spectral feature for reliable chemical identification of materials.

3) SELECTION OF HYPERPARAMETERS

To explore the effect of hyperparameters on the model performance, experiments are conducted with tile size of 3 × 3 pixels. The performance of GMM and DPGMM are

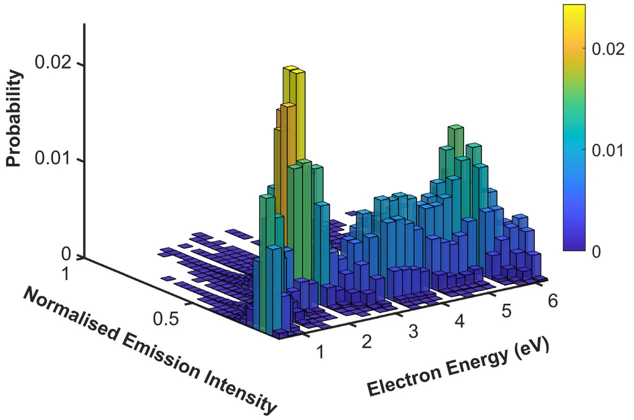


FIGURE 11. Bivariate histogram showing the joint probability distribution of spectral peak position and peak height, with a 3×3 tiling. The peak height represents SEs emission intensity, scaled to the range of 0 to 1 using min–max normalisation.

TABLE 6. GMM clustering performance using different feature sets. Feature sets include peak position alone, peak position with raw peak height, and peak position with normalised peak height.

Feature Set	s^{avg}	s^{Q1}	s^{med}	s^{Q3}
peak position	0.6221	0.5688	0.7115	0.7755
peak position and height	0.1768	-0.1515	0.2488	0.4745
peak position and normalised height	0.5034	0.3994	0.6748	0.7493

TABLE 7. GMM performance under different hyperparameters.

Regularisation Value	Comput. From[3]	Initialisation	K[4]	NLL[5]	$s^{overall}$	D_{KL}
1.39×10^{-2}	$(0.5 \times R_{eV})^2$	k-means++	3	128441	0.6424	0.1532
		random	4	122666	0.5962	0.0402
2.22×10^{-3}	$(0.2 \times R_{eV})^2$	k-means++	5	122189	0.6455	0.0282
		random	4	122215	0.5394	0.0295
5.55×10^{-4}	$(0.1 \times R_{eV})^2$	k-means++	5	122156	0.7349	0.0269
		random	5	122195	0.4504	0.0281
1.39×10^{-4}	$(0.05 \times R_{eV})^2$	k-means++	5	122155	0.7498	0.0270
		random	6	118703	0.3739	0.0235
3.47×10^{-5}	$(0.025 \times R_{eV})^2$	k-means++	5	122155	0.7400	0.0268
		random	6	117602	0.3767	0.0226

[3] Regularisation value is scaled based on the energy resolution, R_{eV}, which is roughly 0.24 eV in this study.

[4] K is the optimal cluster number with GMM (see Section IV-C for details).

[5] NLL denotes the negative log-likelihood, which equals $-lnL(\mathbf{X} \mid \boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ given in Eq. 2.

TABLE 8. DPGMM performance under different hyperparameters.

α	λ_0	K[6]	NLL[5]	$s^{overall}$	D_{KL}
0.6	0.4	4	122199	0.5310	0.0278
0.6	0.6	5	122190	0.5662	0.0275
0.6	0.8	4	128125	0.5468	0.1432
0.8	0.4	4	122198	0.5232	0.0273
0.8	0.6	5	122172	0.6103	0.0281
0.8	0.8	5	122196	0.5191	0.0275
1.0	0.4	4	122198	0.5226	0.0283
1.0	0.6	5	**122157**	**0.6739**	**0.0273**
1.0	0.8	6	122158	0.7092	0.0290
1.2	0.4	4	124140	0.5446	0.0662
1.2	0.6	5	122160	0.6800	0.0276
1.2	0.8	6	122114	0.4458	0.0270

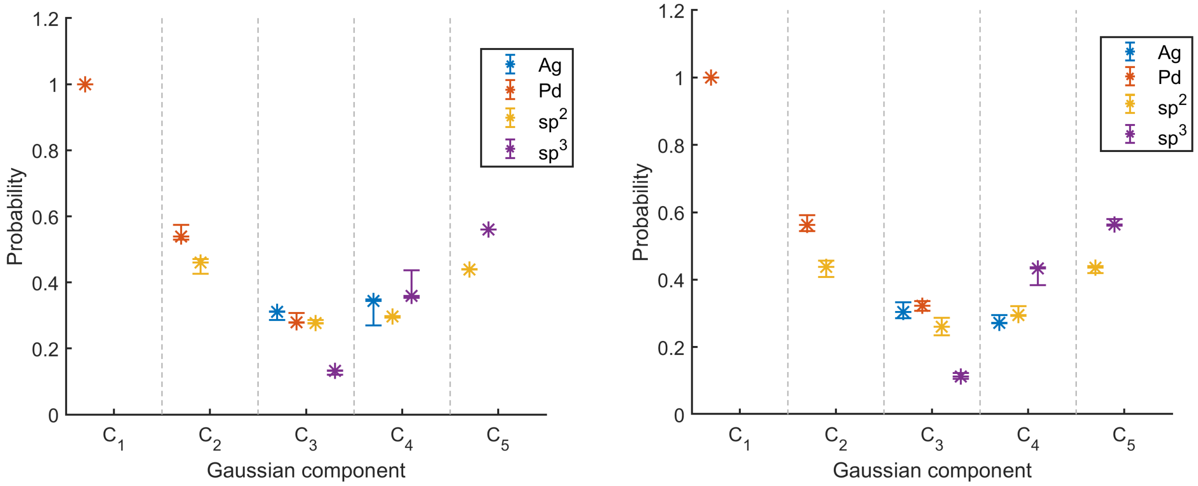
[6] K is the cluster number automatically inferred from DPGMM .

displayed in Tables 7 and 8, respectively. The performance evaluation include the negative log-likelihood NLL, the overall silhouette score $s^{overall}$ and the KL divergence D_{KL}, which describe the goodness of model fitting. Table 7 shows the results by GMM with the regularisation value changing from 10^{-5} to 10^{-2} with two initialisation methods. As observed, the results by GMM are relatively stable with the k-means++ initialisation and the regularisation value changing from 10^{-5} to 10^{-4}, which corresponds to $(0.025 \times R_{eV})^2$ to $(0.1 \times R_{eV})^2$. The k-means++ initialisation shows superior modelling performance and greater robustness compared to random initialisation. Thus, we choose the k-means++ initialisation and set the regularisation value as $(0.1 \times R_{eV})^2$ to improve numerical stability in the GMM implementation.

Table 8 shows the results by DPGMM with changing two hyperparameters α and λ_0. They control the level of the DP prior in creating new clusters and the Gaussian components in the NIW prior, respectively. To allow flexibility in the cluster structure, small values are assigned to α and λ_0, over $\alpha \in \{0.6, 0.8, 1.0, 1.2\}$ and $\lambda_0 \in \{0.4, 0.6, 0.8\}$. The performance results of DPGMM with a combination of changing α and λ_0 are listed in Table 8. It can be observed that larger α leads to more clusters. And the results are relatively stable with changing α. The results are more sensitive to λ_0 than α. Two sets of $\alpha = 1$, $\lambda_0 = 0.6$ and $\alpha = 1.2$, $\lambda_0 = 0.6$ yield optimal performance, as evidenced by low values NLL, high values $s^{overall}$ and low D_KL. Thus we set $\alpha = 1$ and $\lambda_0 = 0.6$ for the DPGMM implementation, which shows the lowest NLL.

4) UNCERTAINTY QUANTIFICATION

Considering that signal noise in the collected spectral data is a primary source of uncertainty, we evaluate its influence on the overall framework in this section. The impact of noise propagation on the final chemical identification is quantified through Monte Carlo simulations. Specifically, two levels of Gaussian noise are injected into the raw SEHI data to simulate slightly and heavily contaminated spectral signals. Following the recent work [90], the injected Gaussian noise has a mean of zero. For the slight and heavy noise conditions, the standard deviations of the Gaussian noise are set to 0.1 and 0.5 times the standard deviation of the raw SEHI data, respectively. For each noise level, 25 Monte Carlo runs were performed. To assess the uncertainty in the resulting chemical



(a) Uncertainty of material assignment under slight noise.

(b) Uncertainty of material assignment under heavy noise.

FIGURE 12. Uncertainty quantification of material assignment using Monte Carlo simulations. The median values of 25 Monte Carlo outputs are marked with *, and error bars represent the 2.5th and 97.5th percentiles as lower and upper bounds. It shows 95% confidence intervals for assigning the chemical elements or bonding types (Ag, Pd, *sp*^{2} carbon, *sp*^{3} carbon) to five Gaussian Components C_1 to C_5. (a) and (b) are computed under slight and heavy noise conditions, respectively.

probability maps, 95% confidence intervals are computed. The median value across the 25 Monte Carlo outputs is used to represent the central estimate of the material assignment. The confidence bounds are determined using the 2.5th and 97.5th percentiles of the Monte Carlo results. Fig. 12 shows the uncertainty of the material assignment associated with 95% confidence intervals obtained under two varying noise levels.

Monte Carlo simulation results indicate that the proposed framework maintains good robustness and reliability in material identification, under varying levels of noise. For the component C_1, the associated material is confidently identified as Pd. The reason is that its characteristic spectral peak occurs at a low energy below 1 eV, which is not present in any of the other candidate materials. As shown in Fig. 12 (a), when slight noise is injected, no observable uncertainty appears in the material assignments for the components C_1 and C_5. In contrast, the components C_2 and C_4 exhibit some variability in the 2–4 eV energy range, suggesting larger sensitivity to noise in material assignments. Under heavy noise injection, C_1 remains unchanged, and C_5 shows only a negligible shift, as shown in Fig. 12 (b). The increased uncertainty in C_3 and C_4 can be attributed to the close proximity of their cluster centres in terms of the peak position, making them more susceptible to noise. Moreover, the presence of multiple mixed materials within C_3 and C_4 further complicates accurate material identification.

D. VALIDATION ACROSS DIFFERENT MATERIAL SAMPLES

To further validate the performance of this framework, experiments are conducted with different material samples. As mentioned in Section IV, there are four Pd-Ag-C film specimens with varying surface roughness and film thickness. These four samples correspond to $A5$, $A6$, $A7$, and $A8$ in the data repository [55]. They are characterized as thick porous, thick smooth, thin smooth, and thin porous, respectively. Based on the performance analysis above, the GMM and DPGMM approaches generally achieve comparable performance. Thus for high computational efficiency, the GMM with tiling in 3 × 3 pixels is employed here. Table 9 summarises the outcomes of GMM clustering and corresponding performance metrics, from four Pd-Ag-C film specimens. As observed, the framework demonstrates consistent and good performance with different material samples in terms of s^{med} and D_{KL}. Compared with the previous study [49], this framework enables more comprehensive chemical analysis through the tile-wise processing. It provides a promising solution for analysing subtle variations in chemical composition between morphologically diverse specimens, even within the same material type. This can facilitate systematic experimental studies of complex material systems.

TABLE 9. GMM results on four SEHI datasets from different Pd-Ag-C film samples.

Sample	Tile	Data	Model	Gaussian Component					s^{avg}	s^{med}	D_{KL}
	Num.	Points	Para.	C_1	C_2	C_3	C_4	C_5			
thick			ω (%)	1.25	27.84	19.84	25.52	25.55			
porous	153,384	70,916	μ (eV)	**0.97**	**1.97**	**3.18**	**4.36**	**5.44**	0.5936	0.6921	0.0256
($A5$)			σ (eV)	0.28	0.25	0.68	0.62	0.36			
thick			ω (%)	9.24	27.91	19.55	18.27	25.05			
smooth	157,055	73,045	μ (eV)	**0.81**	**1.95**	**3.35**	**4.50**	**5.49**	0.6206	0.7112	0.0269
($A6$)			σ (eV)	0.21	0.25	0.55	0.6	0.43			
thin			ω (%)	1.47	31.97	35.78	15.14	15.65			
smooth	156,663	72,308	μ (eV)	**0.88**	**1.99**	**3.92**	**5.16**	**5.67**	0.5832	0.6713	0.0272
($A7$)			σ (eV)	0.23	0.23	0.71	0.4	0.31			
thin			ω (%)	1.11	29.08	33.03	14.71	22.07			
porous	157,170	73,709	μ (eV)	**0.79**	**1.88**	**3.55**	**4.66**	**5.46**	0.4782	0.6173	0.0282
($A8$)			σ (eV)	0.19	0.25	0.79	0.57	0.4			

E. DISCUSSION

As shown in the case study, only two prime spectral peaks, denoting Ag and sp^3-dominant carbon materials, are noticeable from the previous study [49]. However, more than two peaks are expected to be observed in the printed complex Pd-Ag-C film. This novel AI framework proposed enables us to detect five peaks, providing more effective and accurate surface chemical analysis. The results show good accuracy and robustness to changes in the tile size, and different material samples, which demonstrates the reliability of this AI framework for materials applications. The relevant SEHI dataset [55] and the Matlab code package [56] presented in this study are publicly shared. A unique aspect is that this framework links the theoretical DOS model [83], [84] with the experimental findings with SEHI technique. *The Materials Project* [53] provides over 89k database entries of the DOS model, publicly available as a reference to this framework proposed for automated chemical analysis.

Although the present case study focuses on Pd-Ag-C films, the proposed approach is generic, and can be applied to other material systems. SEHI has previously demonstrated its applicability across a range of material systems, including the ability to image oxidation processes in polymers [48], [91] and to differentiate between different forms of carbon [82], underscoring its chemical sensitivity across diverse classes of materials. These capabilities suggest that the approach can be readily extended to other relevant systems, where nanoscale variations in chemistry play a critical role.

As mentioned above, SEHI provides chemical bonding-level information at the nanoscale. Owing to the fundamental differences in information depth and analytical capabilities, a direct comparison for external validation, such as between SEHI and SEM-EDS techniques, is not feasible within the scope of this study. It would be beneficial to consider how this AI-powered SEHI-based approach could complement other established surface-sensitive methods. For example, X-ray photoelectron spectroscopy (XPS) provides highly quantitative surface chemical information, with sensitivity to elemental composition and bonding states. But it is limited in spatial resolution and requires ultra-high vacuum conditions. In contrast, SEHI can deliver nanoscale, spatially resolved contrast linked to surface chemistry under more flexible imaging conditions. This suggests that, when employed in combination, XPS and SEHI techniques have the potential to offer enhanced chemical specificity and spatial mapping. Similarly, atomic force microscopy (AFM) yields topographical and mechanical property information at the nanoscale, but does not directly probe chemical variation. SEHI can complement AFM by providing chemically sensitive contrast in the same regions where mechanical or morphological variations are observed [92], thereby strengthening correlations between structure, chemistry, and function.

This work mainly relys on the direct link with the theoretical DOS model for chemical assignments and accuracy evaluation. Meanwhile, please note that the DOS model derived from density functional theory (DFT) calculations may exhibit systematic shifts or inaccuracies when compared to experimental SEHI spectra, particularly due to surface sensitivity and instrument-specific effects. Such factors should be considered when interpreting chemical assignments, and proper calibration should be implemented in practice to ensure meaningful comparison between theoretical and experimental data. In this work, energy alignment was calibrated using an HOPG reference specimen, by applying an energy shift based on comparison to sp^2 and sp^3 carbon bonding [69]. Detailed calibration procedures can be found from the Supporting Information in [69] shown in Fig. S10 (d).

Our results demonstrate that the proposed approach provides reliable decisions. However, ambiguities could occur when multiple peaks have similar amplitudes and are very close to each other. These close peaks might be identified as corresponding to the same chemical components. Such situations will be reflected in the probability heatmap. To facilitate the decision in such cases, specific experimental reference samples could be designed and used for existing experimental data collection. We demonstrated the impact of experimental samples for carbon [82]. Additionally, experimental collection parameters [70] or instrumentation that allows collection with high spectral resolution, such as add-on spectrometers [74], might be necessary. Alternatively, complementary nano-spectroscopy such as nanoscale Fourier transform infrared spectroscopy (nano-FTIR) might be adopted as demonstrated in [75].

VI. CONCLUSION

This paper proposes a novel AI framework that leverages unsupervised machine learning to enable automated characterisation of materials' surface chemistry. The proposed framework caters for data-driven materials discovery as well as systematic experimental studies of functional materials. It offers valuable insights into identifying chemical elements and bonds, and characterising chemical inhomogeneity on complex material surfaces at the micro- and nano-scale.

The focus of this work is on the integration of image tiling and unsupervised clustering with SEHI data analysis. A unique contribution is that the proposed framework connects the theoretical models with the experimental SEHI results. Image tiling captures the diversity of localised spatial-spectral information within a SEHI data cube, avoiding the manual selection of ROIs. Unsupervised probabilistic clustering methods, GMM and DPGMM, are adopted to model the energy distribution of spectral peaks extracted from each tile. The case study on complex metal alloy and carbon films demonstrates that this framework effectively uncovers previously undetected chemical properties within SEHI experimental data. It achieves high accuracy in chemical identification, with relative errors under $\pm 15\%$ when compared to the theoretical DOS model, except for one Gaussian component associated with the lowest energy. Future work will focus on automating the chemical analysis for a wide range of materials and exploiting additional features such as peak widths and heights, with enhanced instrumentation. Besides, future studies could benefit from multi-modal measurements, by integrating SEHI with complementary techniques, to further enhance the analytical confidence and robustness. Validation using synthetically generated data, along with comparisons between SEHI and SEM-EDS results, is another open avenue that can provide deeper insights into both techniques.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

DATA AND CODE AVAILABILITY

The SEHI data [55] used for the performance evaluation is publicly available on Figshare (https://doi.org/10.15131/shef.data.22202923). The Matlab source code [56] of this proposed approach is free to download inline with MIT license on Figshare (https://doi.org/10.15131/shef.data.27757590)

ACKNOWLEDGMENT

The authors acknowledge the support from the Insigneo Institute for In Silico Medicine, Sheffield.

REFERENCES

[1] S. M. Moosavi, K. M. Jablonka, and B. Smit, "The role of machine learning in the understanding and design of materials," *J. Amer. Chem. Soc.*, vol. 142, no. 48, pp. 20273–20287, Dec. 2020.

[2] J. F. Rodrigues, L. Florea, M. C. F. de Oliveira, D. Diamond, and O. N. Oliveira, "Big data and machine learning for materials science," *Discover Mater.*, vol. 1, no. 1, pp. 1–27, 2021.

[3] M. Nord, P. E. Vullum, I. MacLaren, T. Tybell, and R. Holmestad, "Atomap: A new software tool for the automated analysis of atomic resolution images using two-dimensional Gaussian fitting," *Adv. Structural Chem. Imag.*, vol. 3, no. 1, pp. 1–12, Dec. 2017.

[4] S. V. Kalinin, B. G. Sumpter, and R. K. Archibald, "Big-deep-smart data in imaging for guiding materials design," *Nature Mater.*, vol. 14, no. 10, pp. 973–980, Oct. 2015.

[5] G. Hermann, N. Coudray, J.-L. Buessler, D. Caujolle-Bert, H.-W. Rémigy, and J.-P. Urban, "ANIMATED-TEM: A toolbox for electron microscope automation based on image analysis," *Mach. Vis. Appl.*, vol. 23, no. 4, pp. 691–711, Jul. 2012.

[6] J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques, "Recent advances and applications of machine learning in solid-state materials science," *Npj Comput. Mater.*, vol. 5, no. 1, p. 83, Aug. 2019.

[7] B. DeCost, J. Hattrick-Simpers, Z. Trautt, A. Kusne, E. Campo, and M. Green, "Scientific AI in materials science: A path to a sustainable and scalable paradigm," *Mach. Learn., Sci. Technol.*, vol. 1, no. 3, Sep. 2020, Art. no. 033001.

[8] K. M. Jablonka, D. Ongari, S. M. Moosavi, and B. Smit, "Big-data science in porous materials: Materials genomics and machine learning," *Chem. Rev.*, vol. 120, no. 16, pp. 8066–8129, Aug. 2020.

[9] B. Selvaratnam and R. T. Koodali, "Machine learning in experimental materials chemistry," *Catal. Today*, vol. 371, pp. 77–84, Jul. 2021.

[10] Z. Ge, Z. Song, S. X. Ding, and B. Huang, "Data mining and analytics in the process industry: The role of machine learning," *IEEE Access*, vol. 5, pp. 20590–20616, 2017.

[11] M. Botifoll, I. Pinto-Huguet, and J. Arbiol, "Machine learning in electron microscopy for advanced nanocharacterization: Current developments, available tools and future outlook," *Nanosc. Horizons*, vol. 7, no. 12, pp. 1427–1477, 2022.

[12] J. Peng, D. Schwalbe-Koda, K. Akkiraju, T. Xie, L. Giordano, Y. Yu, C. J. Eom, J. R. Lunger, D. J. Zheng, R. R. Rao, S. Muy, J. C. Grossman, K. Reuter, R. Gómez-Bombarelli, and Y. Shao-Horn, "Human- and machine-centred designs of molecules and materials for sustainability and decarbonization," *Nature Rev. Mater.*, vol. 7, no. 12, pp. 991–1009, 2022.

[13] S. Muto and M. Shiga, "Application of machine learning techniques to electron microscopic/spectroscopic image data analysis," *Microscopy*, vol. 69, no. 2, pp. 110–122, Apr. 2020.

[14] E. Hoq, O. Aljarrah, J. Li, J. Bi, A. Heryudono, and W. Huang, "Data-driven methods for stress field predictions in random heterogeneous materials," *Eng. Appl. Artif. Intell.*, vol. 123, Aug. 2023, Art. no. 106267.

[15] X. Li, Z. Liu, S. Cui, C. Luo, C. Li, and Z. Zhuang, "Predicting the effective mechanical property of heterogeneous materials by image based modeling and deep learning," *Comput. Methods Appl. Mech. Eng.*, vol. 347, pp. 735–753, Apr. 2019.

[16] C. Gu, Y. Bao, S. Prasad, Z. Lyu, and J. Lian, "Defect engineering of fatigue-resistant steels by data-driven models," *Eng. Appl. Artif. Intell.*, vol. 124, Sep. 2023, Art. no. 106517.

[17] Y. Yu, X. Wu, and Q. Qian, "Better utilization of materials' compositions for predicting their properties: Material composition visualization network," *Eng. Appl. Artif. Intell.*, vol. 117, Jan. 2023, Art. no. 105539.

[18] J. Jung, J. I. Yoon, H. K. Park, J. Y. Kim, and H. S. Kim, "An efficient machine learning approach to establish structure-property linkages," *Comput. Mater. Sci.*, vol. 156, pp. 17–25, Jan. 2019.

[19] A. Paul, D. Jha, R. Al-Bahrani, W.-K. Liao, A. Choudhary, and A. Agrawal, "CheMixNet: Mixed DNN architectures for predicting chemical properties using multiple molecular representations," 2018, *arXiv:1811.08283*.

[20] L. Vlcek, M. Ziatdinov, A. Maksov, A. Tselev, A. P. Baddorf, S. V. Kalinin, and R. K. Vasudevan, "Learning from imperfections: Predicting structure and thermodynamics from atomic imaging of fluctuations," *ACS Nano*, vol. 13, no. 1, pp. 718–727, Jan. 2019.

[21] F. Uesugi, S. Koshiya, J. Kikkawa, T. Nagai, K. Mitsuishi, and K. Kimoto, "Non-negative matrix factorization for mining big data obtained using four-dimensional scanning transmission electron microscopy," *Ultramicroscopy*, vol. 221, Feb. 2021, Art. no. 113168.

[22] P. A. López-García, D. L. Argote, and M. C. Thrun, "Projection-based classification of chemical groups for provenance analysis of archaeological materials," *IEEE Access*, vol. 8, pp. 152439–152451, 2020.

[23] R. Aversa, P. Coronica, C. De Nobili, and S. Cozzini, "Deep learning, feature learning, and clustering analysis for SEM image classification," *Data Intell.*, vol. 2, no. 4, pp. 513–528, Oct. 2020.

[24] S. Tsopanidis and S. Osovski, "Unsupervised machine learning in fractography: Evaluation and interpretation," *Mater. Characterization*, vol. 182, Dec. 2021, Art. no. 111551.

[25] S. S. Chong, Y. S. Ng, H.-Q. Wang, and J.-C. Zheng, "Advances of machine learning in materials science: Ideas and techniques," *Frontiers Phys.*, vol. 19, no. 1, p. 13501, Feb. 2024.

[26] G. Pastorelli, A. S. Ortiz Miranda, and A. H. Christensen, "Interpretation of X-ray spectral data using self-organising maps and hierarchical clustering: A study of vilhelm Hammershøi's use of painting materials," *X-Ray Spectrometry*, vol. 53, no. 5, pp. 392–404, Sep. 2024.

[27] M. J. Pasterski, M. Lorenz, A. V. Ievlev, R. C. Wickramasinghe, L. Hanley, and F. Kenig, "Machine learning correlation of electron micrographs and ToF-SIMS for the analysis of organic biomarkers in mudstone," *J. Amer. Soc. for Mass Spectrometry*, vol. 36, no. 1, pp. 58–71, Jan. 2025.

[28] B. Yin, Q. Hu, Y. Zhu, and K. Zhou, "Semi-supervised learning for shale image segmentation with fast normalized cut loss," *Geoenergy Sci. Eng.*, vol. 229, Oct. 2023, Art. no. 212039.

[29] M. Khodadadzadeh, C. Contreras, L. Tusa, and R. Gloaguen, "Subspace clustering algorithms for mineral mapping," *Proc. SPIE*, vol. 10789, pp. 557–564, Jul. 2018.

[30] J. Chen, X. Jang, K. Goto, T. Tsutsumi, and Y. Toyoda, "A flexible deep learning based approach for SEM image denoising," *Proc. SPIE*, vol. 12955, pp. 37–44, Jun. 2024.

[31] Y. Okada, H. Liu, C.-E. Lee, C.-H. Tien, and P. Yu, "Enhancing CD-SEM accuracy with attention-boosted Noise2Noise model," *Proc. SPIE*, vol. 12955, pp. 598–605, Apr. 2024.

[32] R. M. Arachchige, J. Olek, F. Rajabipour, and S. Peethamparan, "Phase identification and micromechanical properties of non-traditional and natural pozzolan based alkali-activated materials," *Construct. Building Mater.*, vol. 441, Aug. 2024, Art. no. 137478.

[33] M. S. S. Ahamad and E. N. M. Maizul, "Digital analysis of geo-referenced concrete scanning electron microscope (SEM) images," *Civil Environ. Eng. Rep.*, vol. 30, no. 2, pp. 65–79, Jun. 2020.

[34] A. Herbeaux, H. Aboleinein, A. Villani, C. Maurice, J.-M. Bergheau, and H. Klöcker, "Combining phase field modeling and deep learning for accurate modeling of grain structure in solidification," *Additive Manuf.*, vol. 81, Feb. 2024, Art. no. 103994.

[35] D. Alagic and J. Pilz, "Unsupervised algorithm to detect damage patterns in microstructure images of metal films," in *Proc. IEEE Int. Conf. Image Process., Appl. Syst. (IPAS)*, Dec. 2018, pp. 67–72.

[36] Z. Du, L. Pu, P. Wei, R. Yuan, J. Kim, and J. Tan, "Unsupervised neural network-based image restoration framework for pattern fidelity improvement and robust metrology," *J. Micro/Nanopatterning, Mater., Metrology*, vol. 22, no. 3, Aug. 2023, Art. no. 034201.

[37] A. Cheng, K. Kang, Z. Zhu, R. Zhang, and L. Wang, "Improving the neural segmentation of blurry serial SEM images by blind deblurring," *Comput. Intell. Neurosci.*, vol. 2023, no. 1, Jan. 2023, Art. no. 8936903.

[38] T. Okuda, J. Chen, T. Motoyoshi, R. Yumiba, M. Ishikawa, and Y. Toyoda, "Comparison between supervised and self-supervised deep learning for SEM image denoising," *Proc. SPIE*, vol. 12496, pp. 916–924, Oct. 2023.

[39] B. Dey, S. Wu, S. Das, K. Khalil, S. Halder, P. Leray, B. Samir, K. Ahi, M. P. Pereira, G. Fenger, and M. Bayoumi, "Unsupervised machine learning based SEM image denoising for robust contour detection," *Proc. SPIE*, vol. 11854, pp. 88–102, May 2021.

[40] R. Juránek, J. Výravský, M. Kolář D. Motl, and P. Zemčík, "Graph-based deep learning segmentation of EDS spectral images for automated mineral phase analysis," *Comput. Geosci.*, vol. 165, Aug. 2022, Art. no. 105109.

[41] C. Sun, S. Lux, E. Müller, M. Meffert, and D. Gerthsen, "Versatile application of a modern scanning electron microscope for materials characterization," *J. Mater. Sci.*, vol. 55, no. 28, pp. 13824–13835, Oct. 2020.

[42] K. D. Vernon-Parry, "Scanning electron microscopy: An introduction," *III-Vs Rev.*, vol. 13, no. 4, pp. 40–44, 2000.

[43] S. Nasr-Esfahani, V. Muthukumar, E. Regentova, K. Taghva, and M. Trabia, "Hyperspectral methods in microscopy image analysis: A survey," in *Proc. 18th Int. Conf. Signal Process. Multimedia Appl.*, Portugal, 2021, pp. 111–119.

[44] W. Han, M. Zheng, A. Banerjee, Y. Z. Luo, L. Shen, and A. Khursheed, "Quantitative material analysis using secondary electron energy spectromicroscopy," *Sci. Rep.*, vol. 10, no. 1, p. 22144, Dec. 2020.

[45] N. Stehling, R. Masters, Y. Zhou, R. O'Connell, C. Holland, H. Zhang, and C. Rodenburg, "New perspectives on nano-engineering by secondary electron spectroscopy in the helium ion and scanning electron microscope," *MRS Commun.*, vol. 8, no. 2, pp. 226–240, Jun. 2018.

[46] N. Farr, M. Davies, J. Nohl, K. J. Abrams, J. Schäfer, Y. Lai, T. Gerling, N. Stehling, D. Mehta, J. Zhang, L. Mihaylova, J. R. Willmott, K. Black, and C. Rodenburg. (2024). *Supporting Information for Revealing the Morphology of Ink and Aerosol Jet Printed Palladium-Silver Alloys Fabricated From Metal Organic Decomposition Inks*. [Online]. Available: https://advanced.onlinelibrary.wiley.com/action/downloadSupplement?doi=%0.1002%2Fadvs.202306561&file=advs7085-sup-0001-SuppMat.pdf

[47] N. T. H. Farr, M. Davies, J. Nohl, K. J. Abrams, J. Schäfer, Y. Lai, T. Gerling, N. Stehling, D. Mehta, J. Zhang, L. Mihaylova, J. R. Willmott, K. Black, and C. Rodenburg, "Revealing the morphology of ink and aerosol jet printed palladium-silver alloys fabricated from metal organic decomposition inks," *Adv. Sci.*, vol. 11, no. 10, Mar. 2024, Art. no. 2306561.

[48] N. Farr, J. Thanarak, J. Schäfer, A. Quade, F. Claeyssens, N. Green, and C. Rodenburg, "Understanding surface modifications induced via argon plasma treatment through secondary electron hyperspectral imaging," *Adv. Sci.*, vol. 8, no. 4, Feb. 2021, Art. no. 2003762.

[49] K. J. Abrams, M. Dapor, N. Stehling, M. Azzolini, S. J. Kyle, J. Schäfer, A. Quade, F. Mika, S. Kratky, Z. Pokorna, I. Konvalina, D. Mehta, K. Black, and C. Rodenburg, "Making sense of complex carbon and metal/carbon systems by secondary electron hyperspectral imaging," *Adv. Sci.*, vol. 6, no. 19, Oct. 2019, Art. no. 1900719.

[50] J. Zhang, J. Nohl, N. T. H. Farr, C. Rodenburg, and L. Mihaylova, "Unsupervised learning assisted secondary electron hyperspectral imaging for high-throughput cheminformatics analysis of materials," in *Proc. BIO Web Conf.*, vol. 129, 2024, p. 10012.

[51] F. Georget, W. Wilson, and K. L. Scrivener, "Simple automation of SEM-EDS spectral maps analysis with Python and the edxia framework," *J. Microsc.*, vol. 286, no. 2, pp. 185–190, May 2022.

[52] J. M. Munro, K. Latimer, M. K. Horton, S. Dwaraknath, and K. A. Persson, "An improved symmetry-based approach to reciprocal space path selection in band structure calculations," *Npj Comput. Mater.*, vol. 6, no. 1, p. 112, Jul. 2020.

[53] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, "Commentary: The Materials Project: A materials genome approach to accelerating materials innovation," *APL Mater*, vol. 1, 2023, Art. no. 011002, doi: 10.1063/1.4812323. [Online]. Available: https://pubs.aip.org/aip/apm/article/1/1/011002/119685/Commentary-The-Materials-Project-A-materials

[54] V. K. Khanna, *Introductory Nanoelectronics*. Boca Raton, FL, USA: CRC Press, 2020.

[55] J. Zhang, J. N. N. Farr, C. Rodenburg, K. Abrams, K. Black, and L. Mihaylova, "SEHI (secondary electron hyperspectral imaging) dataset of metal alloy and carbon film (palladium silver carbon complex film)," 2023, doi: 10.15131/SHEF.DATA.22202923.

[56] J. Zhang, J. Nohl, N. Farr, C. Rodenburg, and L. Mihaylova, "AI-assisted data analytical tool for secondary electron hyperspectral imaging (in MATLAB)," 2024, doi: 10.15131/shef.data.27757590.v2.

[57] A. Rizwan, N. Iqbal, A. N. Khan, R. Ahmad, and D. H. Kim, "Toward effective pattern recognition based on enhanced weighted K-mean clustering algorithm for groundwater resource planning in point cloud," *IEEE Access*, vol. 9, pp. 130154–130169, 2021.

[58] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[59] E. Ehsaeyan, "An efficient image segmentation method based on expectation maximization and salp swarm algorithm," *Multimedia Tools Appl.*, vol. 82, no. 26, pp. 40625–40655, Nov. 2023.

[60] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 381–396, Mar. 2002.

[61] G. J. McLachlan, S. X. Lee, and S. I. Rathnayake, "Finite mixture models," *Annu. Rev. Statist. Appl.*, vol. 6, pp. 355–378, Apr. 2019.

[62] K. P. Burnham and D. R. Anderson, "Multimodel inference: Understanding AIC and BIC in model selection," *Sociol. Methods Res.*, vol. 33, no. 2, pp. 261–304, Nov. 2004.

[63] H. Z. Yerebakan and M. Dundar, "Partially collapsed parallel Gibbs sampler for Dirichlet process mixture models," *Pattern Recognit. Lett.*, vol. 90, pp. 22–27, Apr. 2017.

[64] C. Mo. (2019). *PRML: MATLAB Code of Machine Learning Algorithms in Book Pattern Recognition and Machine Learning*. [Online]. Available: https://github.com/PRML/PRMLT.git

[65] R. Das, "Collapsed Gibbs sampler for Dirichlet process Gaussian mixture models (DPGMM)," Carnegie Mellon University, Tech. 2014. [Online]. Available: https://www.cs.cmu.edu/

[66] Z. Li, L. S. Mihaylova, O. Isupova, and L. Rossi, "Autonomous flame detection in videos with a Dirichlet process Gaussian mixture color model," *IEEE Trans. Ind. Informat.*, vol. 14, no. 3, pp. 1146–1154, Mar. 2018.

[67] Y. Li, E. Schofield, and M. Gönen, "A tutorial on Dirichlet process mixture modeling," *J. Math. Psychol.*, vol. 91, pp. 128–144, Aug. 2019.

[68] T. J. Rogers, K. Worden, R. Fuentes, N. Dervilis, U. T. Tygesen, and E. J. Cross, "A Bayesian non-parametric clustering approach for semi-supervised structural health monitoring," *Mech. Syst. Signal Process.*, vol. 119, pp. 100–119, Mar. 2019.

[69] J. F. Nohl, N. T. H. Farr, Y. Sun, G. M. Hughes, N. Stehling, J. Zhang, F. Longman, G. Ives, Z. Pokorná, F. Mika, V. Kumar, L. Mihaylova, C. Holland, S. A. Cussen, and C. Rodenburg, "Insights into surface chemistry down to nanoscale: An accessible colour hyperspectral imaging approach for scanning electron microscopy," *Mater. Today Adv.*, vol. 19, Aug. 2023, Art. no. 100413.

[70] J. F. Nohl, N. T. H. Farr, Y. Sun, G. M. Hughes, S. A. Cussen, and C. Rodenburg, "Low-voltage SEM of air-sensitive powders: From sample preparation to micro/nano analysis with secondary electron hyperspectral imaging," *Micron*, vol. 156, May 2022, Art. no. 103234.

[71] N. T. H. Farr, G. M. Hughes, and C. Rodenburg, "Monitoring carbon in electron and ion beam deposition within FIB-SEM," *Materials*, vol. 14, no. 11, p. 3034, Jun. 2021.

[72] J. Nohl, "Pysehi releases," Tech. Rep., 2023, doi: 10.15131/shef.data.22310068.v1.

[73] J. Nohl, N. Farr, N. Stehling, J. Zhang, F. Longman, G. Ives, L. Mihaylova, C. Holland, and C. Rodenburg, "CSEHI app 1.0," Tech. Rep., 2023, doi: 10.15131/shef.data.21647090.

[74] D. C. Joy, M. S. Prasad, and H. M. Meyer, "Experimental secondary electron spectra under SEM conditions," *J. Microsc.*, vol. 215, no. 1, pp. 77–85, Jul. 2004.

[75] R. C. Masters, N. Stehling, K. J. Abrams, V. Kumar, M. Azzolini, N. M. Pugno, M. Dapor, A. Huber, P. Schäfer, D. G. Lidzey, and C. Rodenburg, "Mapping polymer molecular order in the SEM with secondary electron hyperspectral imaging," *Adv. Sci.*, vol. 6, no. 5, Mar. 2019, Art. no. 1801752.

[76] V. Kumar, W. L. Schmidt, G. Schileo, R. C. Masters, M. Wong-Stringer, D. C. Sinclair, I. M. Reaney, D. Lidzey, and C. Rodenburg, "Nanoscale mapping of bromide segregation on the cross sections of complex hybrid perovskite photovoltaic films using secondary electron hyperspectral imaging in a scanning electron microscope," *ACS Omega*, vol. 2, no. 5, pp. 2126–2133, May 2017.

[77] J. Piper and J. Duselis, "Spectral library clustering using a Bayesian information criterion," in *Proc. Sensor Signal Process. Defence (SSPD)*, Sep. 2016, pp. 1–5.

[78] T. M. Kodinariya and P. R. Makwana, "Review on determining number of cluster in K-means clustering," *Int. J.*, vol. 1, no. 6, pp. 90–95, 2013.

[79] A. A. Neath and J. E. Cavanaugh, "The Bayesian information criterion: Background, derivation, and applications," *WIREs Comput. Statist.*, vol. 4, no. 2, pp. 199–203, Mar. 2012.

[80] P. Orbanz and Y. W. Teh, "Bayesian nonparametric models," in *Encyclopedia of Machine Learning*, 2012.

[81] N. Farr, S. Pashneh-Tala, N. Stehling, F. Claeyssens, N. Green, and C. Rodenburg, "Characterizing cross-linking within polymeric biomaterials in the SEM by secondary electron hyperspectral imaging," *Macromolecular Rapid Commun.*, vol. 41, no. 3, Feb. 2020, Art. no. 1900484.

[82] J. F. Nohl et al., "Secondary electron hyperspectral imaging of carbons: New insights and good practice guide," *Adv. Sci.*, vol. 12, no. 29, Aug. 2025, Art. no. 01907.

[83] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, "Commentary: The materials project: A materials genome approach to accelerating materials innovation," *APL Mater.*, vol. 1, no. 1, Jul. 2013.

[84] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, "Python materials genomics (pymatgen): A robust, open-source Python library for materials analysis," *Comput. Mater. Sci.*, vol. 68, pp. 314–319, Feb. 2013.

[85] P. Peterson, E. Baker, and B. McGaw, *International Encyclopedia of Education*. Amsterdam, The Netherlands: Elsevier, 2009.

[86] E. Patel and D. S. Kushwaha, "Clustering cloud workloads: K-means vs Gaussian mixture model," *Proc. Comput. Sci.*, vol. 171, pp. 158–167, Jan. 2020.

[87] A. Dudek, "Silhouette index as clustering evaluation tool," in *Classification and Data Analysis: Theory and Applications 28*. Cham, Switzerland: Springer, 2020, pp. 19–33.

[88] A. Punhani, N. Faujdar, K. K. Mishra, and M. Subramanian, "Binning-based silhouette approach to find the optimal cluster using K-means," *IEEE Access*, vol. 10, pp. 115025–115032, 2022.

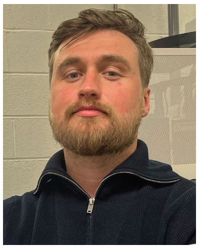
[89] S. Kullback, *Information Theory and Statistics*. New York, NY, USA: Dover, 1997.

[90] J. Cha, D. Cho, and S. Lee, "SEMNet: Deep synthetic training for unprecedented SEM image denoising and super-resolution," *BIO Web Conf.*, vol. 129, p. 10008, Apr. 2024.

[91] N. T. H. Farr, S. Roman, J. Schäfer, A. Quade, D. Lester, V. Hearnden, S. MacNeil, and C. Rodenburg, "A novel characterisation approach to reveal the mechano-chemical effects of oxidation and dynamic distension on polypropylene surgical mesh," *RSC Adv.*, vol. 11, no. 55, pp. 34710–34723, 2021.

[92] N. T. H. Farr, "Revealing localised mechanochemistry of biomaterials using in situ multiscale chemical analysis," *Materials*, vol. 15, no. 10, p. 3462, May 2022.

JINGQIONG ZHANG received the M.Sc. degree in automation and measurement from North China Electric Power University, China, in 2016, and the Ph.D. degree in electronic engineering from the University of Kent, U.K., in 2020. Since 2021, she has been a Research Associate with the School of Electrical and Electronic Engineering, The University of Sheffield, U.K. She has published more than 15 articles and holds one U.S. patent in the field of automation and machine learning. Her current research interests include data mining, machine learning, hyperspectral image processing, data-centric engineering, and multidisciplinary AI and applications.



NICHOLAS T. H. FARR received the B.Sc. degree in human biology from Loughborough University, U.K., the L.L.B. degree from the BPP Law School, U.K., and the Ph.D. degree, in 2020, under the supervision of Prof. C. Rodenburg, with a focus on tissue engineering and regenerative medicine. His Ph.D. thesis was titled "Evaluation of a Novel Scanning Electron Microscope-Based Surface Chemical Mapping Technique for Characterising Polymeric Biomaterials." His current research interests include pre-clinical testing, biomaterials characterisation, instrumentation development, and the application of mechanochemistry in evaluating medical devices, with a particular emphasis on addressing clinical challenges. He was awarded the EPSRC Prize Research Fellowship for his Ph.D. studies.

JAMES NOHL received the M.Eng. degree in materials science and engineering from The University of Sheffield, U.K., in 2020, where he is currently pursuing the Ph.D. degree with the School of Chemical, Materials and Biological Engineering. He has published three first-author articles and "pysehi," an open-source software package for secondary electron spectroscopy and hyperspectral imaging. His research interests include the application of secondary electron hyperspectral imaging to gain insights into functional materials surfaces, particularly energy storage materials.

YUFENG LAI received the Ph.D. degree from The University of Sheffield, Sheffield, U.K., in 2021. He has been a Postdoctoral Research Associate at the Sensor Systems Group, Department for Electrical and Electronic Engineering, The University of Sheffield, since then. His research interests include thermal imaging, non-contact temperature measurement, hyperspectral imaging, combustion diagnostics, fire research, and renewable energy.

KERRY J. ABRAMS received the Ph.D. degree in microscopical investigations of helium implanted silicon, in 2011. She worked in multiple research departments, including Salford, Liverpool, and Sheffield Universities. She has been at Nexperia, Manchester, since July 2021, where she is currently a Principal Failure Analysis Engineer with the Quality Department, where she utilises her materials characterisation expertise to understand and optimise semiconductor heterostacks. She has published more than 20 articles across many materials systems, including carbon materials, polymers, ceramics, and nuclear materials. Her research interest includes the electron microscopy of beam-sensitive materials.

KATE BLACK received the Ph.D. degree in materials science from the University of Liverpool, in 2008. After a research role at the University of Cambridge, she joined the University of Liverpool, in 2013, where she is currently a Professor of manufacturing with the School of Engineering. In 2019, she co-founded Meta Additive, later acquired by Desktop Metal, and founded Atomik AM, in 2022, where she serves as the CEO. Her research interests include developing functional materials for advanced printing technologies, especially in metals and ceramics. She was elected as a fellow of the Royal Academy of Engineering, in 2024, where she continues to lead in engineering innovation and advocates for equality in the field. Her contributions to engineering earned her the title of one of the "Top 50 Women in Engineering" and the Academic Entrepreneur Award, in 2022.

JON WILLMOTT received the M.Phys. and Ph.D. degrees in physics from the University of Southampton, in 1999 and 2003, respectively. He joined the Department of Engineering, University of Cambridge, as a Postdoctoral Researcher, focusing on liquid crystal technology, before transitioning to industry with Land Instruments International (now part of AMETEK Inc.). There, he specialized in the design of thermal imaging cameras, radiation thermometers, and non-contact measurement instruments, developing expertise in optical, mechanical, and electronic design, as well as temperature measurement science (metrology). He returned to academia, in 2015, joining the Department of Electrical and Electronic Engineering, The University of Sheffield, supported by an EPSRC Established Career Fellowship. He leads the Sensor Systems Research Group, Advanced Detector Centre, as a Professor of metrology. His research interests include optical and thermal imaging, emissivity measurement, and advanced metrological methods, with applications in manufacturing, medical imaging, and environmental monitoring.

CORNELIA RODENBURG received the degree in engineering from Westsächsische Hochschule Zwickau, Germany, in 1997, and the Ph.D. degree from Sheffield Hallam University, U.K., in 2001. She joined the Department of Materials Science and Metallurgy, University of Cambridge, as a Postdoctoral Researcher, before moving to the Department of Materials Science and Engineering, The University of Sheffield. She was awarded the Royal Society Dorothy Hodgkin Fellowship and after a brief spell as a University Teacher, held an EPSRC Early Career Fellowship, from 2016 to 2021. She was a Senior Lecturer of materials science and engineering, prior to her appointment as the Chair of nanostructured materials technology at The University of Sheffield, in 2022.

LYUDMILA MIHAYLOVA (Senior Member, IEEE) received the Ph.D. degree in systems and control engineering from the Technical University of Sofia, Bulgaria. She is currently a Professor with the School of Electrical and Electronic Engineering, The University of Sheffield, U.K. Her research interests include signal processing, machine learning, autonomous systems for sensing, tracking, navigation, decision making and with a variety of data from optical, thermal images, secondary electron microscopy images, radar, LiDAR, wireless sensor network data, and others. She is an Associate Editor-in-Chief of IEEE Transactions on Aerospace and Electronic Systems. She has been a Senior Editor of target tracking and multi-sensor data fusion area, since 2021, and a Subject Area Editor of *Signal Processing* (Elsevier), since 2022. She was a Guest Editor for a special issue of *Frontiers in Robotics and AI*, from 2022 to 2023.