ROYAL SOCIETY
OF CHEMISTRY

PAPER

Check for updates

XRFast a new software package for processing of MA-XRF datasets using machine learning†

Marc Vermeulen, [ID] ‡§[a] Alicia McGeachy,§¶[a] Bingjie Xu,[a] Henry Chopp,[b] Aggelos Katsaggelos,[b] Rebecca Meyers,[c] Matthias Alfeld[d] and Marc Walton [ID] §*[e]

X-ray fluorescence (XRF) spectroscopy is a common technique in the field of heritage science. However, data processing and data interpretation remain a challenge as they are time consuming and often require *a priori* knowledge of the composition of the materials present in the analyzed objects. For this reason, we developed an open-source, unsupervised dictionary learning algorithm reducing the complexity of large datasets containing 10s of thousands of spectra and identifying patterns. The algorithm runs in Julia, a programming language that allows for faster data processing compared to Python and R. This approach quickly reduces the number of variables and creates correlated elemental maps, characteristic for pigments containing various elements or for pigment mixtures. This alternative approach creates an overcomplete dictionary which is learned from the input data itself, therefore reducing the *a priori* user knowledge. The feasibility of this method was first confirmed by applying it to a mock-up board containing various known pigment mixtures. The algorithm was then applied to a macro XRF (MA-XRF) data set obtained on an 18th century Mexican painting, and positively identified smalt (pigment characterized by the co-occurrence of cobalt, arsenic, bismuth, nickel, and potassium), mixtures of vermilion and lead white, and two complex conservation materials/interventions. Moreover, the algorithm identified correlated elements that were not identified using the traditional elemental maps approach without image processing. This approach proved very useful as it yielded the same conclusions as the traditional elemental maps approach followed by elemental maps comparison but with a much faster data processing time. Furthermore, no image processing or user manipulation was required to understand elemental correlation. This open-source, open-access, and thus freely available code running in a platform allowing faster processing and larger data sets represents a useful resource to understand better the pigments and mixtures used in historical paintings and their possible various conservation campaigns.

1. Introduction

X-ray fluorescence (XRF) spectroscopy has become a go-to technique in the field of cultural heritage as an initial non-invasive and non-destructive way of investigating inorganic

[a]*Northwestern University/Art Institute of Chicago Center for Scientific Studies in the Arts (NU-ACCESS), 2145 Sheridan Road, Evanston, IL, USA*

[b]*Department of ECE, Northwestern University, Evanston, IL, USA*

[c]*National Museum of Mexican Art, 1852 W. 19th Street, Chicago, IL, USA*

[d]*Delft University of Technology, Department of Materials Science and Engineering (MSE), 2628 CN Delft, Netherlands*

[e]*M+, 38 Museum Drive, West Kowloon Cultural District, Hong Kong. E-mail: marc. walton@mplus.org.hk*

† Electronic supplementary information (ESI) available. See https://doi.org/10.1039/d2ja00114d

‡ Current address: Collection Care Department, The National Archives, Richmond, TW9 4DU, UK.

§ Formerly: NU-ACCESS.

¶ Current address: Department of Scientific Research, The Metropolitan Museum of Art, New York, NY 10028, USA.

pigments.[1-4] Development of XRF scanning for macroscopic objects (MA-XRF), wherein XRF spectra are acquired across an artwork by means of a raster scan, now allows for spatial imaging of pigment distributions based on their elemental composition. Because the technique is well suited to the analysis of flat 2D objects, it has been used extensively for the study of easel paintings,[5-10] but its application has been extended further to many other types artworks, such as illuminated manuscripts,[11-13] furniture,[14] antique statues,[15] stained glass,[16,17] wall paintings,[18-20] and works on paper.[21-23] The elemental images produced can then be used as a way of identifying areas for micro-destructive sampling but can equally be used to reveal hidden compositions[5,10,24,25] or *pentimenti*.[9,26] Importantly pigments are often inferred from these maps based on the spatial co-occurrence of certain elements: emerald green, for instance, is frequently deduced from the combination of arsenic and copper; smalt is identified based on the presence of cobalt, nickel, arsenic, and potassium; vermilion is characterized by the simultaneous presence of mercury and sulfur; cobalt

green is associated with the co-located presence of copper and zinc. However, it is essential to note that the proof positive characterization of a given pigment can only be accomplished through molecular or crystallographic methods such as Raman or infrared spectroscopy, or X-ray diffraction.

The inference of a pigment therefore requires a domain expert who knows what they are looking for and may be complicated by the ubiquitous presence of one or more of the elements, for instance, lead when used for a lead white ground. Recently, machine learning, artificial intelligence (AI), and more specifically, machine and deep learning have been applied to cultural heritage data sets to identify patterns automatically. Such approaches are not limited to MA-XRF but branch out to other imaging techniques such as reflectance imaging spectroscopy (RIS),[24,27,28] as well as to molecular techniques such as Raman and FTIR,[29–31] showing its contribution toward reducing complexity in large datasets containing 10s of thousands of spectra.

In MA-XRF, every pixel represents a spectrum that contains fluorescence lines associated with the elemental composition of surface and sub-surface pictorial layers. Beyond this, the spectra also feature continuous spectroscopic background from scattered primary radiation, especially in the case of polychrome X-ray tube sources and incomplete charge collection in the detector. This whole spectral data carries latent information about the analyzed material and can help build a robust machine learning approach toward understanding the material characteristics. Specific to cultural heritage, due to the limited number of pigments available to artists until the development of modern organic painting materials in the late 19th/early 20th century, for most works of art, one can expect a constrained palette of pigments that results in just a few unique XRF spectra. In other words, the number of pigments used to paint an artwork span a much smaller subspace than the total number of spectra collected during a single MA-XRF scan. Such conditions, where similar spectra are repeated multiple times within an image data cube, are ideally suited to machine learning capable of extracting a set of bases representing the limited inorganic painting materials (*i.e.*, pigments or pigment mixtures) used in the artwork.

One popular way of identifying such bases is by clustering similar spectra using K-means,[32] principal component analysis (PCA),[33] or embedding methods like t-distributed stochastic neighbor embedding (T-SNE)[28] or uniform manifold approximation and projection (UMAP).[24] In a recent article, for instance, self-organizing maps (SOMs) were implemented in R as an unsupervised way to cluster similar spectra within an MA-XRF data cube.[34] With clustering methods, the data is modeled as distributions around discrete points, often called centroids. These approaches produce a first approximation of the spectral bases that represent the XRF signal but not necessarily a "best" set of bases that describe a latent space with the more descriptive and discriminative features.

A more robust approach is matrix factorization which is performed following a standard linear equation:

$$y_i = D\alpha_i \tag{1}$$

in which the data collected at a certain spatial location i is represented by a vector of spectra $y_i \in \mathbb{R}^P$, the dictionary $D \in \mathbb{R}^{PxN}$ is composed of columns of pure component or end-member spectra (*e.g.*, X-ray spectra representing single pigments), and α_i is the corresponding abundance vector. By placing all the data collected into a matrix Y (the i-th column of Y is the vector y_i), eqn (1) can be re-written as

$$Y = DA, \tag{2}$$

where $Y \in \mathbb{R}^{PxM}$, with M the number of spatial locations scanned, $D \in \mathbb{R}^{PxN}$ as before, and the i-th column of matrix $A \in \mathbb{R}^{NxM}$ is the vector α_i. In our formulation P is the number of channels in the X-ray spectrum and N is the number of bases or atoms. D is set as an overcomplete dictionary meaning that N is much larger than the total number of physical pigments composing the artwork. Eqn (1) and (2) will be used interchangeably.

In work concerning the analysis of cultural heritage,[35,36] the number of columns of D is taken as the pigments used in the painting which is estimated by visual inspection of the artwork. Then based on this number, initial seed spectra for D were determined by clustering, and A was calculated by performing linear least squares (LLS) fit to Y. In a subsequent step, A was held constant and the entries of D were updated. This alternating optimization between D and A has been called *multivariate curve resolution-alternating least squares (MCR-ALS)* or alternatively non-negative matrix factorization and typically implemented using off-the-shelf commercial software. However, there are disadvantages to *MCR-ALS*. Most notably, the number of pure component spectra forming D is fixed and must be known *a priori* to avoid problems of over- or under-fitting of Y with LLS.

Here we suggest an alternative approach called sparse dictionary learning,[37] in which D and non-zero N are learned from a set of training data instead of prespecifying the number of pure spectral components. The benefits of determining D and N from the spectral data cubes itself without pre-clustering has been demonstrated for synchrotron-based X-ray microscopy data.[38–40] Sparsity penalization of the abundance vectors α_i is key to this work as it reduces fitting errors to the observations y_i, especially when the number of spectra composing D is unknown. Further details about the model are provided in Section 2.

2. Model definition: a non-negative K-SVD approach

In our previous work we focused on finding sparse representations of visible wavelength spectral reflectance data[41,42] and expanded our work looking into super-resolution of XRF cubes. In this current work we apply a more complex approach to MA-XRF data in which we form an overcomplete dictionary, learned from the input data, which takes into account the whole spectrum, a number of latent experimental factors including compositional variations, paint thicknesses, attenuation of the X-ray signal on account of material layering, counting statistics, and detector noise.

The standard dictionary learning algorithm is two steps: first a sparse coding step is performed based on an initial dictionary, and second, an update of the dictionary is made to produce a better fit to the data. Sparse coding is always a constrained optimization problem such as:

$$\min_{\alpha_i}\|\alpha_i\|_0 \text{ such that } y_i = D\alpha_i \tag{3}$$

where $\|.\|_0$ denotes the ℓ_0 pseudonorm. The optimization problem in eqn (3) can be solved by greedy algorithms like matching pursuit.[43] Alternatively, a surrogate to the ℓ_0 norm can be used, such as the ℓ_1 norm, and the relaxed problem, referred to as the LASSO problem, can be solved instead.[43] In this paper, however, we follow a different approach to the sparse coding step aimed at solving the following constrained optimization problem:

$$\min_{\alpha_i}\|y_i - D\alpha_i\|_2^2 \text{ under the constraint } \alpha_i \geq 0 \tag{4}$$

where $\alpha_i \geq 0$ denotes that the entries of α_i are non-negative so that abundances are physically interpretable as will be discussed in more detail below. To determine α_i under these conditions, and also obtain a sparse solution that establishes the number of non-zero Ns, non-negative least squares was utilized[44] followed by a hard thresholding of the coefficients wherein only the largest values were retained, and small values were set to zero. This thresholding operation is equivalent to a projection operator, which represents a non-expansive mapping. Once the sparse coding is completed, D is optimized via the approximate K-SVD algorithm, in which each endmember column is improved by an approximate singular value decomposition (SVD) of nearest neighbors thereby making it similar to a generalized form of K-means clustering.[43] We also impose the constraint that the entries of the end member spectra are non-negative, denoted as $D \geq 0$. To enforce this constraint we apply a non-negativity constraint (also a non-expansive mapping) to the result of the K-SVD, that is, all negative entries of D are set to zero.

The algorithm alternates between the sparse coding step, based on the current dictionary, and updating the columns of the dictionary to better fit the data. One of the benefits of this algorithm, is that after several iterations, all correlated elemental bands and features (e.g., those elements that belong to a specific pigment) separate into distinct endmembers in D. Also, the sparse abundances α form easily interpretable pigment distribution images that can supplement the typical elemental distribution maps which are the output of classic XRF fitting methods offered by such packages of PyMCA.[45]

The entire algorithm was coded in Julia as a freely available and open-source package named XRFast and made available on the NU-ACCESS GitHub (**https://github.com/NU-ACCESS/XRFast.jl**). Julia is a high-level, high-performance programming language (comparable to C) which is well suited for numerical analysis and computational science due to its speed and ability to handle larger datasets.[46] Julia is often considered an order of magnitude faster than Python or R,[46–48] the languages regularly used thus far to process MA-XRF or

hyperspectral imaging datasets.[24,34] Furthermore, in order to improve performance, Julia does not require the vectorization or other optimization of the data like is often needed in R or Python. Consequently, with the ever-growing size of the datasets under study, this algorithm represents a novel and suitable option for data scientists working with complex, large datasets.

In practice, the algorithm was performed "online", meaning that with each iteration of sparse coding and dictionary learning, a random 5% selection of the original data Y is used. The 5% threshold can be modified to include fewer or greater spectra, however in a typical experiment 5% was chosen as an acceptable compromise between speed and performance. Likewise, the algorithm was found to converge quickly, with no more than 10 iterations required.

3.   Materials and methods

3.1.   Mock-up reference board

The algorithm was first tested on a mock-up reference board (Fig. 1) containing a series of stripes of pigments, applied alone or in mixtures. The methodology followed for the creation of the mock-up board is described elsewhere.[49] The pigments consist of zinc white, lead white, cobalt blue, chrome oxide green, vermilion, Prussian blue, cobalt blue, lead chromate, yellow ochre, mixture of verdigris and sodium arsenate (to mimic Emerald green), and tin oxide, as an alternative for organic red precipitated on tin oxide substrate, all applied on a white calcium carbonate ground layer. The composition of the various stripes and pigments composition within the stripes are given in Table 1.

3.2.   Historical painting

A painting belonging to the permanent collection of the National Museum of Mexican Art (Chicago, USA) was
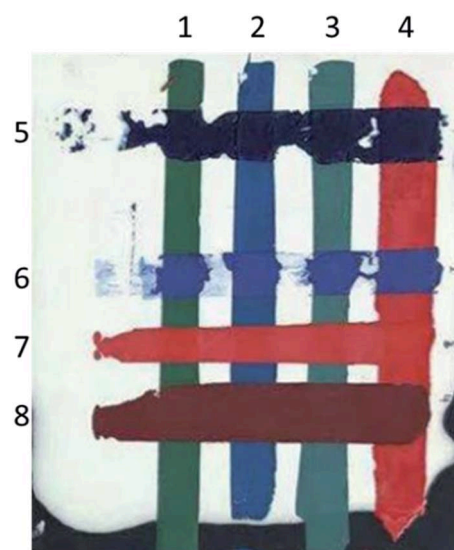

Fig. 1 Mock-up paint board with stripes (1–8) of pigments mixtures. Stripes 1 to 4 were applied first and stripes 5 to 8 were applied next and are therefore overlaid with stripes 1–4. The pigment composition of the various stripes is given in Table 1.

Table 1 Pigment composition of pigment mixtures used for the realization of the mock-up pigment board. The third column reports the elements that are expected to be identified using XRF whereas the fourth columns reports the identified endmembers associated with each pigment mixtures following the XRFast data treatment

Stripe	Composition (%)	Expected elements	Identified endmember (see Fig. 4)
1	Zinc white (10)	Zn	EM1
	Lead white (15)	Pb	
	Cobalt blue (15)	Co	
	Chrome oxide green (60)	Cr	
2	Zinc white (10)	Zn	EM2
	Lead white (15)	Pb	
	Cobalt blue (60)	Co	
	Chrome oxide green (15)	Cr	
3	Zinc white (10)	Zn	EM3
	Lead white (60)	Pb	
	Cobalt blue (15)	Co	
	Chrome oxide green (15)	Cr	
4	Vermilion (100)	Hg/S	EM5
5	Lead white (55)	Pb	EM8
	Tin oxide (10)	Sn	
	Prussian blue (25)	Fe	
	Cobalt blue (10)	Co	
6	Vermilion (30)	Hg/S	EM6
	Tin oxide (40)	Sn	
	Cobalt blue (30)	Co	
7	Vermilion (100)	Hg/S	EM5
8	Vermilion (40)	Hg/S	EM4
	Lead chromate (5)	Pb/Cr	
	Verdigris + sodium arsenate (15)	Cu/As	
	Yellow ochre (40)	Fe	
Board	Calcium carbonate (100)	Ca	EM7

scientifically investigated non-invasively to explore its color palette. The oil and canvas painting under investigation (accession number 2000.34, Fig. 2), untitled and by unknown artist, is dated from the 18th century and derived from an engraving inspired by the oil on panel sketch entitled "Franciscan Allegory in Honor of the Immaculate Conception" by Peter Paul Rubens [Flemish 1577–1640] at the Philadelphia Museum of Art.[50] The area of interest scanned using MA-XRF and reflectance imaging spectroscopy (RIS) is indicated by the red rectangle in Fig. 2.

3.3. Macro X-ray fluorescence (MA-XRF)

Macro X-ray fluorescence was carried out on two separate instruments. An XGLab Elio energy dispersive X-ray fluorescence analyzer was used to scan the pigment mock-up board and the instrument described by Pouyet et al.,[13] was used to scan the oil painting in the collection of the NMMA. For the latter, the X-ray tube and polycapillary optic were replaced by a 50 W transmission Rh anode X-ray tube (Varex Imaging, Salt Lake City, UT, USA) and a 500 μm collimator, respectively. The detector was placed at a 45-degree angle from the collimator and a 23 mm distance between the collimator and the surface of the object allows for a safe working distance and minimize as much as possible the loss of low energy lines.[51] The X-ray source was operated at 40 kV and 1 mA. Due to the large size of the painting (1.27 × 1.75 m), it was mapped with a 1.5 mm step size, and a 100 ms per pixel dwell time, for a total of 6 cubes. Only one cube is presented here due to the presence of most pigments of interest in that particular area (indicated by the red square in Fig. 2).

The XGLab Elio instrument is equipped with a high-resolution large area silicon drift detector with 130 eV at manganese (Mn) Kα with 10 kcps input photon rate (high
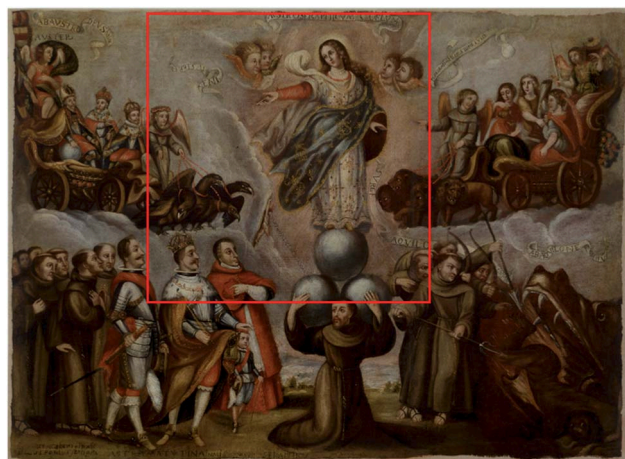

Fig. 2 18th century untitled oil on canvas after an engraving inspired by the oil on panel sketch entitled "Franciscan Allegory in Honor of the Immaculate Conception" by Peter Paul Rubens [Flemish 1577–1640] at the Philadelphia Museum of Art, 127.3 × 175.3 cm², painted by an unknown artist (National Museum of Mexican Art, Chicago, United States).

This journal is © The Royal Society of Chemistry 2022

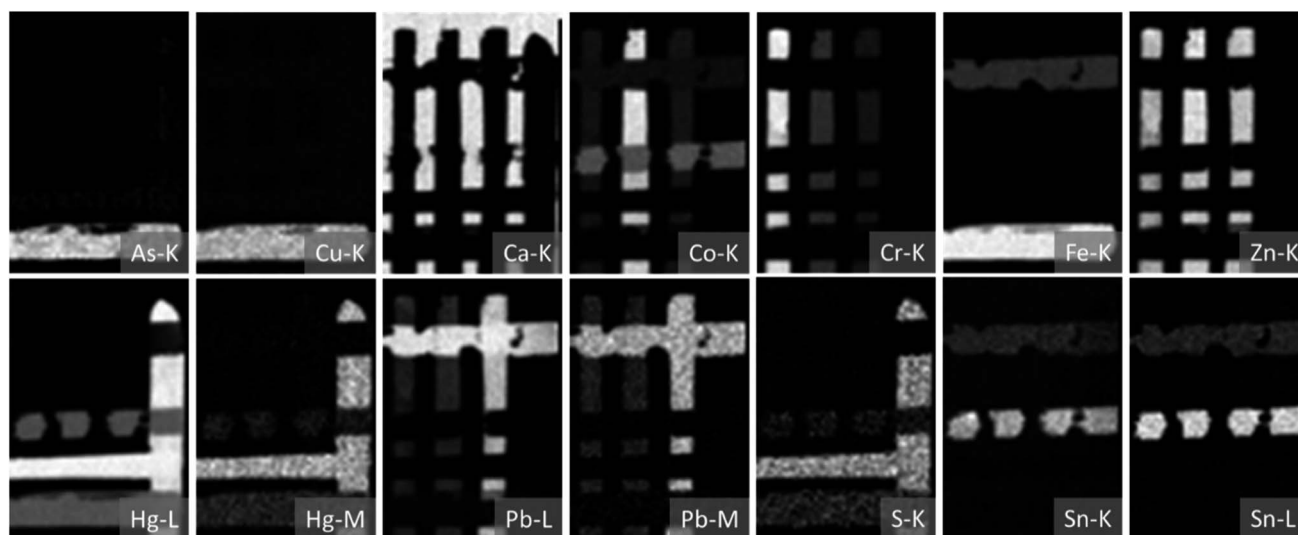J. Anal. At. Spectrom., 2022, **37**, 2130–2143 | **2133**

Fig. 3 Elemental maps obtained using PyMCA. Elements identified include arsenic (As), copper (Cu), calcium (Ca), cobalt (Co), chrome (Cr), iron (Fe), zinc (Zn), mercury (Hg), lead (Pb), sulfur (S), and tin (Sn).

resolution mode), 170 eV at Mn Kα with 200 kcps input photon rate (fast mode), changeable filters, and a rhodium (Rh) transmission target with 50 kV maximum voltage and 4 W maximum power. The size of the analyzed spot is 1 mm in diameter. Elemental 2D mapping of the surface can be achieved through automatic XY raster scanning. In our case, the following measurement conditions were used: 10 s measurement time per point, 40 kV tube voltage, 40 μA tube current, 100 ms movement delay, and a dimension of the scanned area of 5 × 7.3 cm (50 row and 73 columns with 1 mm distance between 2 rows or columns).

XRF data processing and creation of pigments/colors and comparative elemental distribution maps was undertaken through a custom-made Jupyter Notebook running in the Julia programing language for dictionary learning and with a custom made Jupyter Notebook running Python and utilizing the PyMCA library,[45] respectively. The images are given in Fig. 3. Prior to undertaking the data reduction through the dictionary learning algorithm, the MA-XRF data is compressed through a log–log-square root operator.[52] This operation is typically performed in conjunction with baseline removal (as with the SNIP algorithm).[53] Here the log–log-square root operator is used to emphasize minor spectral bands or features that would otherwise be suppressed or eliminated by the SVD decomposition.

3.4. Reflectance imaging spectroscopy (RIS)

Reflectance imaging spectroscopy data was acquired using a Resonon Pika II hyperspectral camera measuring between 400 and 900 nm with 2.1 nm spectral resolution, with a total of 240 channels. Only data between 410 and 900 nm were used for pigment identification due to noise affecting lower wavelength channels. The system was connected to a stage allowing the scanning of about 30 cm of the object's width, with a pixel size of 0.5 × 0.5 mm. During acquisition, the object was illuminated using two broad spectrum tungsten halogen lamps placed at 45°

of the objects normal. A Spectralon diffuse white reflectance standard (Labsphere, North Sutton, USA) was used as a calibration target to convert the image cubes to diffuse reflectance. RIS acquisition was performed using the SpectrononPro software (Resonon, Inc., Bozeman, MT, USA). The raw hyperspectral data cubes were converted to a tiff stack in Fiji and the partially overlapped areas of each scan were stitched together using registration and stitching plugins available in the open-source image processing package Fiji suite,[54,55] prior to further processing. Processing of the data collected was undertaken using the UMAP method described in Vermeulen *et al.*[24]

4. Results

4.1. Mock-up pigment board

The elemental maps relevant for the identification of pigments obtained through the classic MA-XRF data processing of the mock-up pigment board are given in Fig. 3. All anticipated elements were identified and consistent with the pigment mixtures used to create the mock-up pigment board, as described in Section 3.1.

The correlated element maps and their associated XRF spectra obtained through the dictionary learning approach developed in this article are given in Fig. 4. To illustrate the speed associated with using Julia programming language, the data processing of the 50 × 73 × 4096 pixels cube took 10 seconds. As illustrated, the number of endmembers (EMs) is eight (8), all presenting a unique XRF signature instead of noise components within **D**. The 8 EMs identified correspond to the 7 pigment mixtures described in Table 1 (mixtures 4 and 7 being the same pigment) and the calcium carbonate ground.

4.2. Historical painting

The maps of elements relevant for the identification of pigments obtained through the classic MA-XRF data processing of historical painting board are given in Fig. 5.
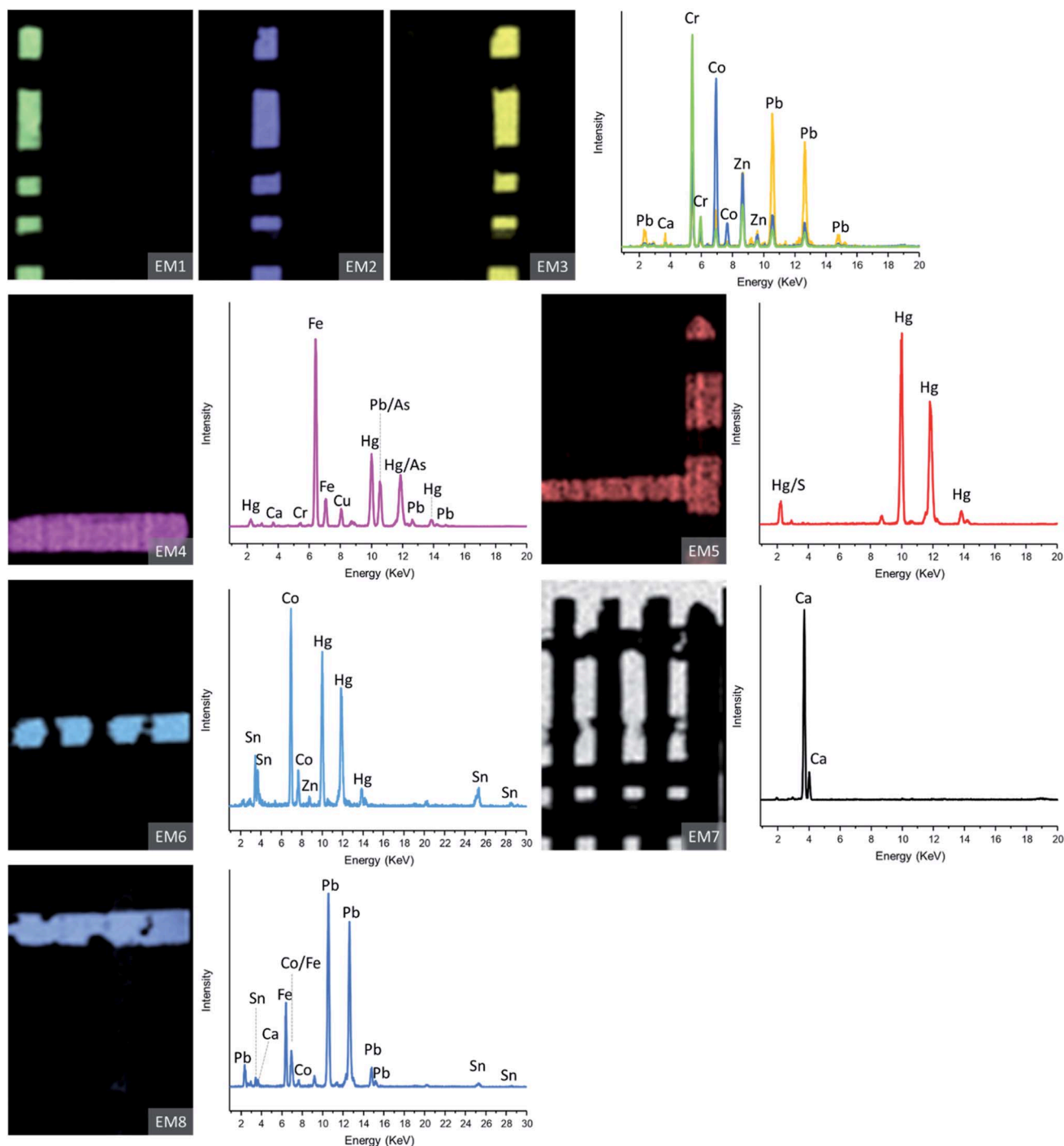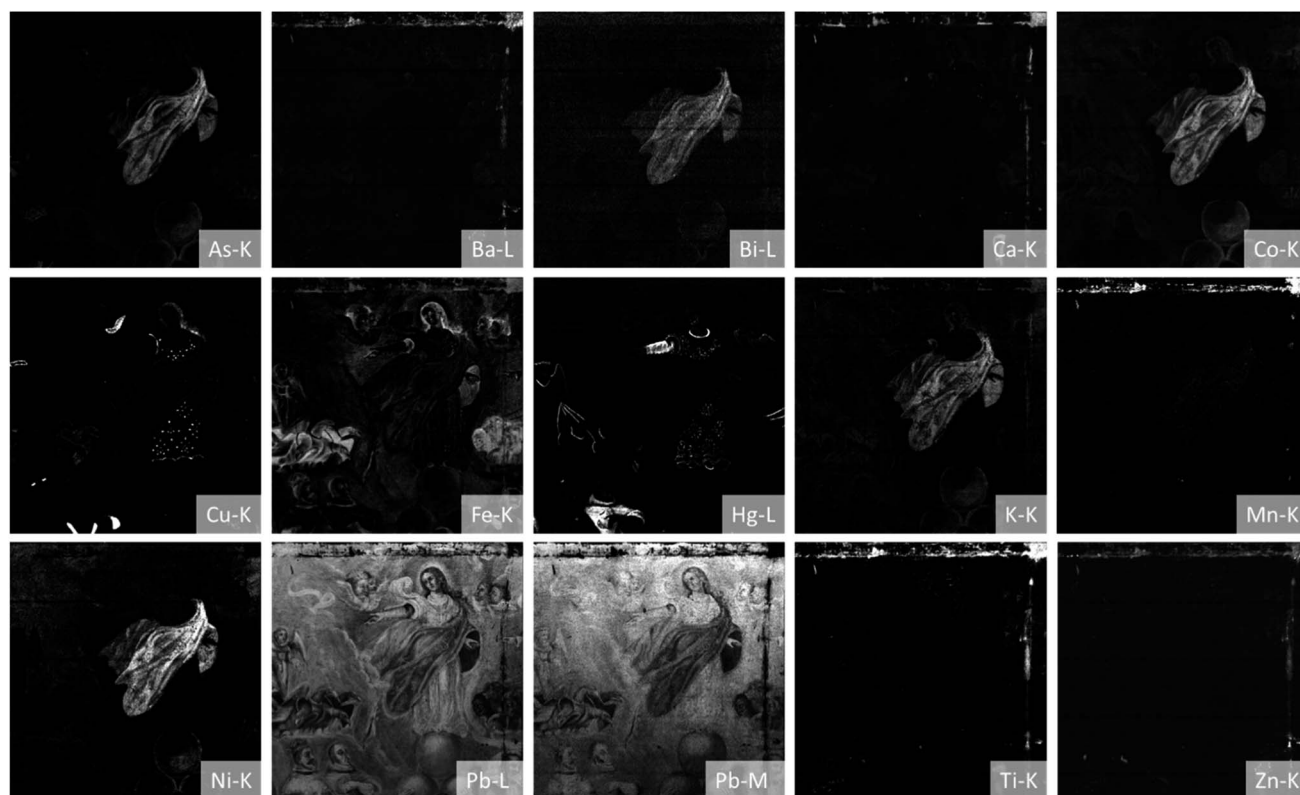
Fig. 4 Correlated element maps and associated XRF spectra (linear scale) obtained using the dictionary learning approach. See Section 3.1 for description of mock-up.

The correlated element maps and their associated XRF spectra obtained through the dictionary learning approach developed in this article are given in Fig. 6. To illustrate the speed associated with Julia, the data processing of the 536 × 548 × 4096 pixels cube took 8 minutes and did not require any fitting of the sum spectrum, nor extensive experience of the end user. As illustrated, the number of endmembers is six (6).

Some small contribution for copper (Cu) and zinc (Zn) found for endmembers 1, 2, 4, 5, and 6 can be associated with the brass used for the collimator, rather than the studied materials themselves. Similarly, iron (Fe) and lead (Pb) are found in all endmembers independently of the color of the materials. According to microsample taken on the edges of the painting (not shown) and microinvasive analyses (SEM-EDX, Fig. S1†), the signal for these elements may be associated with the red

This journal is © The Royal Society of Chemistry 2022

J. Anal. At. Spectrom., 2022, **37**, 2130–2143 | **2135**

Fig. 5 Elemental maps obtained using PyMCA. Elements identified include arsenic (As), barium (Ba), bismuth (Bi), calcium (Ca), cobalt (Co), copper (Cu), iron (Fe), mercury (Hg), potassium (K), manganese (Mn), nickel (Ni), lead (Pb), titanium (Ti), and zinc (Zn).

iron oxide/oxy-hydroxide and white Pb-rich layers found as ground/imprimaturs.

5. Discussion

5.1. Identification of pigment mixtures and pigments of known complex elemental composition

The identification of all 7 stripes of pigments as illustrated in Fig. 4, clearly highlights the ability of the algorithm to, quickly and accurately, identify pigments of complex elemental compositions. Indeed, among others, EM4 presents all elements expected for the complex mixture of vermilion (Hg), lead chromate (Pb/Cr), verdigris and sodium arsenate (Cu/As) and ochre (Fe) expected for mixture 8 (Table 1), whereas EM6 is being identified based on the elemental signature of the complex mixture 6 (Table 1), composed of vermilion (Hg), cobalt blue (Co) and tin oxide (Sn). It is important to highlight that the algorithm can differentiate between the Hg coming from mixture 6 and 8 as well as the Hg coming from mixtures 4 and 7, also suggesting the power of such an approach. The identification of the aforementioned pigment mixtures using the elemental maps presented in Fig. 3 would have required a much more difficult and time-consuming cross-identification process, encouraging possible mistakes. Similarly, yet simpler, vermilion and copper-based pigments mixed with lead white were identified in the historical painting (Fig. 6). In both cases, Pb is likely used as lead white to yield desired lighter shade. The

identification of vermilion solely based on the Hg elemental map (Fig. 5) is unambiguous as Hg is only found in this red pigment. Nonetheless, the pigment was also confirmed through RIS based on its very characteristic spectrum with inflection point at 590 nm (Fig. S2a†). The lack of additional features in the reflectance spectrum does not suggest any other pigments despite the increased absorbance in the 600–700 nm range, which has been previously observed for vermilion.[24,56] However, the presence of lead white, in mixture with vermilion, would go undetected in RIS due to the lack of features for the pigment in the visible and near infrared range used for this study. While clear from the dictionary learning approach, the additional presence of Pb in the red areas is however not evident from the Pb–L elemental map due to the overall presence of lead throughout the composition (Fig. 5). Only inspection of the signal intensity for each pixel or image manipulation such as RGB composite image composed of Hg, and Pb for the red and blue channels, respectively highlight a similar result as indicated by the pinkish magenta color created when blue (Pb) and red (Hg) overlap (Fig. 7a).

In the case of the Cu-based and Pb-based pigment mixture observed in the wing of the angel, embellishments of the blue robe gemstones of the crown and inner sleeves of the kings in the lower left quadrant (endmember map 3 (EM3); Fig. 6), and likely identified as aged verdigris or copper resinate based on its RIS features (Fig. S2b†),[57–59] while the presence of Pb is clear based on the dictionary learning approach XRF spectrum

2136 | J. Anal. At. Spectrom., 2022, 37, 2130–2143

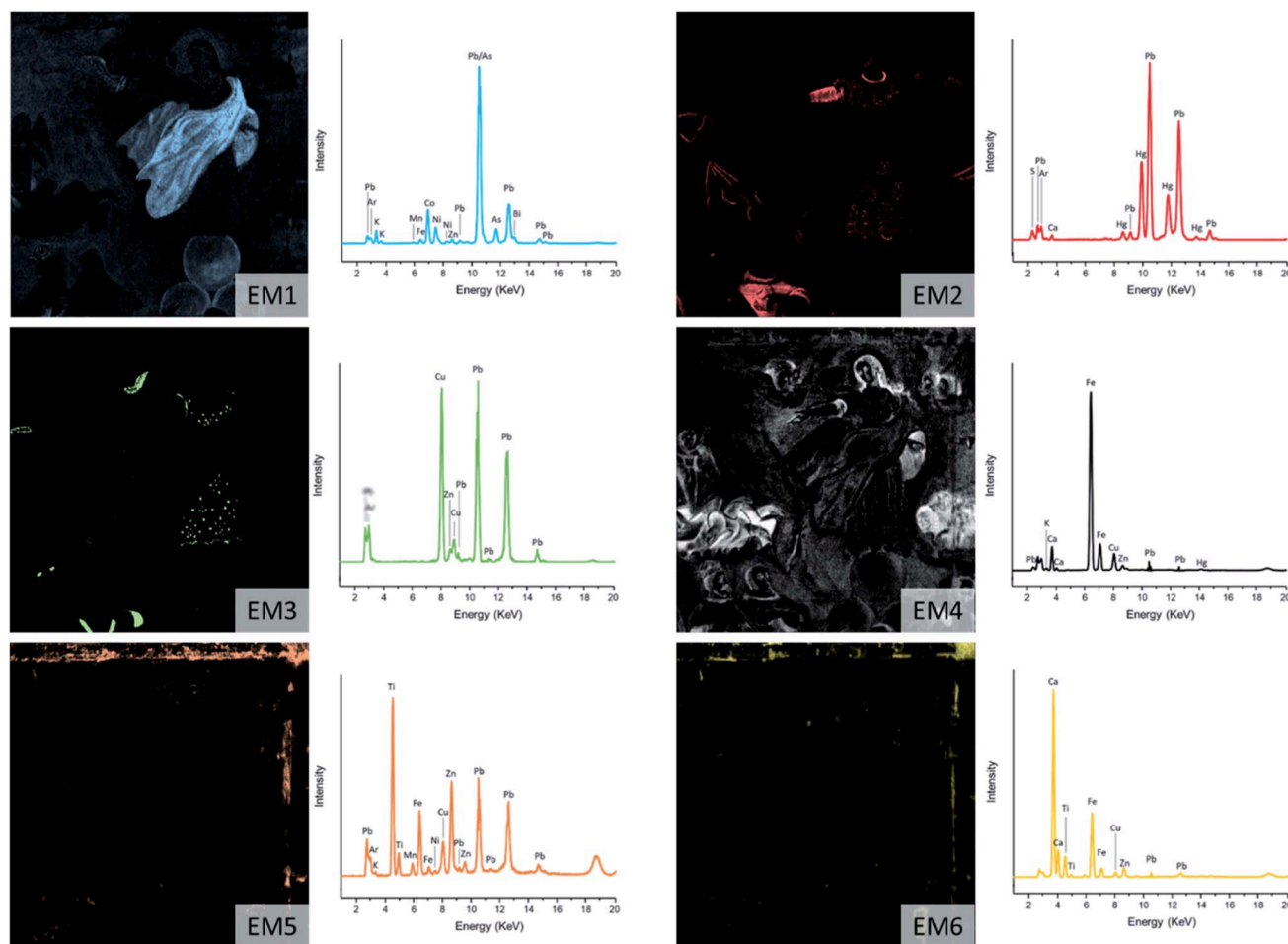This journal is © The Royal Society of Chemistry 2022

Fig. 6 Correlated element maps and associated XRF spectra (linear scale) obtained using the dictionary learning approach.

presented in Fig. 6, its presence along with Cu is not evident from the Pb map, again due to the overall presence of the element in the composition (Fig. 5). RGB Pb/Cu composite image suggests however the presence of the element mixture based on the magenta color produced by the overlapping of the two elemental maps (Fig. 7b), which shares similarities with the distribution map produced by the dictionary learning approach. This, comparatively to what was described for the Hg/Pb pigment mixture, shows the strength of the machine learning approach providing the same conclusions as what is obtained by overlapping individual elemental maps without having to
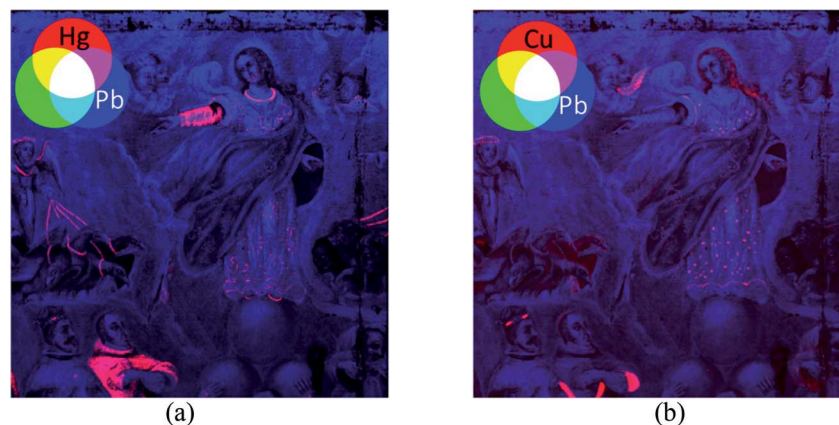
Fig. 7 (a) RGB composite image for Hg (red), and Pb (blue), showing the correlation between Hg and Pb (magenta) in the red areas of the composition, and (b) RGB composite image for Cu (red), and Pb (blue), showing the correlation between Cu and Pb (magenta) in the dark (likely to be green) areas of the composition.
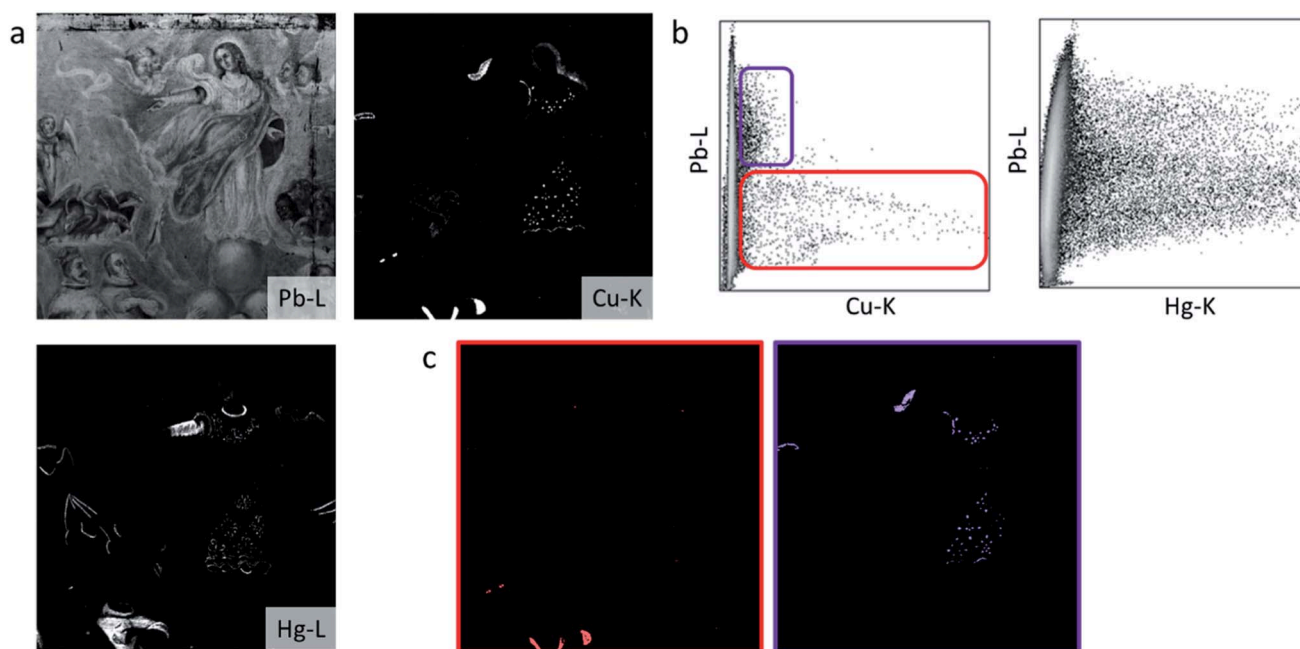
Fig. 8 (a) Elemental maps for lead (Pb), copper (Cu) and mercury (Hg), (b) Cu/Pb and Hg/Pb correlation plots highlighting a non-fully correlation between the elements, likely associated with heterogeneous mixtures of pigments and (c) back mapping of the two areas (Pb-rich, purple and Pb-poor, red) associated with the Cu/Pb maps and likely corresponding to two Cu/Pb ratios.

undertake any image manipulation, often requiring the user to know what they are looking for.

The correlation between Hg and Pb and Cu and Pb is further indicated by the correlation plots obtained using Fiji plugin ScatterJ, which enables the comparison of two images and to extract statistical and spatial information based on their pixel values. When two images are completely uncorrelated, the correlation plot will feature one horizontal and one vertical line intersecting in the origin, corresponding to the pixel values of both images (*e.g.*, Hg and Zn, Fig. S3a and b†). When the images are fully correlated, as one would expect in the case of elements found in a pigment, the correlation plot will present a diagonal line (*e.g.*, As and Ni, Fig. S3c and d†). When elements are not fully correlated but are still found together in the same area, as

a result of heterogeneous mixtures of pigments, variations in pixel intensities may be expected and may lead to further correlation variations that what is observed for fully correlated elements. This is what is observed for both the Cu/Pb and the Hg/Pb correlation plots (Fig. 8). Furthermore, the Cu/Pb correlation plot also suggests two mixtures (Fig. 8b and c): one richer in Pb (purple rectangle, dress embellishments and angel wing) and one poorer in Pb (red rectangle, crown emerald and green from outfits). Consequently, the dictionary learning approach provided the same results as what is obtained through the standard pipeline without requiring further investigation and image manipulation.

The algorithm proved suitable to identify complex mixtures but appears to be suitable to identify single pigments
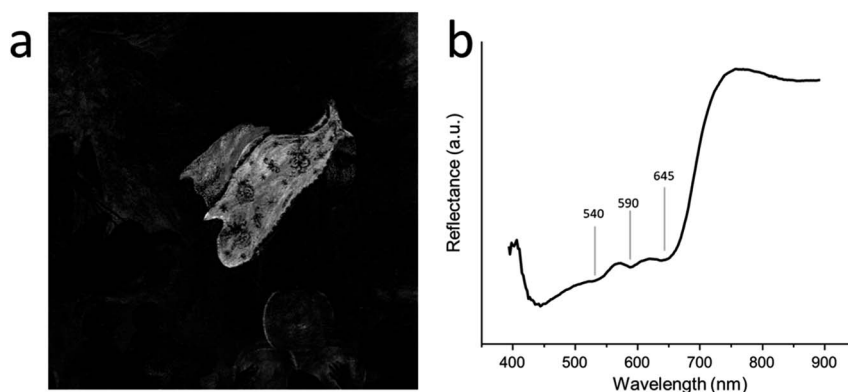
Fig. 9 Reflectance imaging spectroscopy distribution map of materials associated with smalt (a) and associated reflectance curve (b) presenting the characteristic features of smalt.

2138 | *J. Anal. At. Spectrom.*, 2022, **37**, 2130–2143

This journal is © The Royal Society of Chemistry 2022

mixtures of zinc white, lead white, cobalt blue, and chrome oxide green). Yet, the algorithm also proved valuable to identify elements likely found in minor amount in a complex mixture. Many times, such elements may go undetected in the elemental maps due to the dynamic range of the greyscale elemental maps. Indeed, when elements show large counts in certain areas, these areas appear very bright in the elemental map and will quench the range, making areas with fewer counts appear much less intense to the point it may go undetected. This is what is observed in the case of the Cr used in mixture 8 where lead chromate represents only a very small proportion of the pigment mass compared to the other pigments of the mixture accounting for 95% of the total pigment mass. Consequently, and because Cr is found in much greater quantity in mixture 1, the Cr in mixture 8 goes undetected in the Cr elemental map (Fig. 3-Cr) while being positively identified through the dictionary learning approach (Fig. 4-EM4). A similar issue is observed in the historical painting where it is difficult to say without any further user manipulation that the inner sleeve of the Virgin Mary contains anything except for Fe based on the elemental maps created using PyMCA (Fig. 10a). From the non-manipulated elemental maps, no Ca appears to be found in this area (Fig. 10b). However, the extracted XRF spectra (red

spectrum in Fig. 10e) clearly shows that, while Fe is the main element found in this area, Ca, is also present but went undetected in the elemental maps. This is most likely due to the bright Ca-rich areas that will quench the poorer (darker) elemental counts areas of the image. The presence of the Ca only becomes more evident when drastically adjusting the contrast and brightness of the image (Fig. 10c), further requiring some user input. The dictionary learning approach did however identify Ca in the Fe-rich brown areas of the composition (EM4, Fig. 10d and black spectrum in Fig. 10e). This again shows how the algorithm can work around images visual limitations and provide the users with images that reflect better the full nature of the material under study.

Ca found in the Fe-rich brown areas may have been used as additive in the Fe-based ochre pigment to yield darker shades as Ca may correspond to the addition of bone black. However, the final characterization of these materials would require further molecular analyses, not available in the frame of this study.

5.3. Identification of various conservation materials with similar composition

According to Fig. 4, the dictionary learning algorithm presented in this paper was able to easily differentiate mixtures 1, 2 and 3,
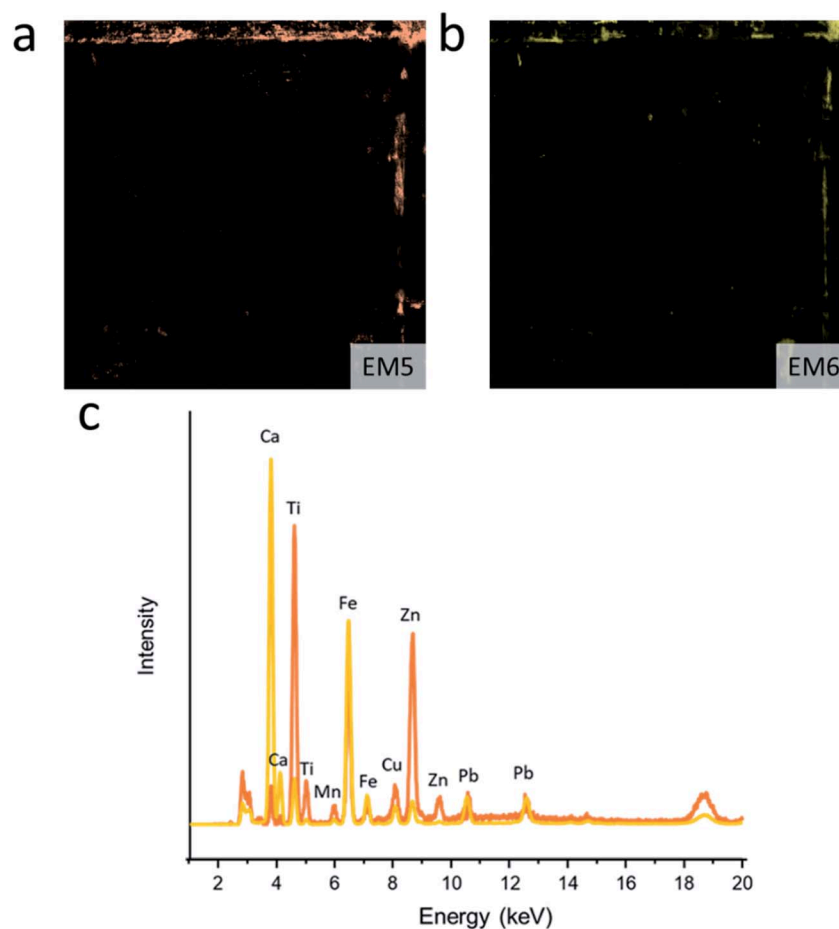

Fig. 11 Dictionary learning endmember maps (a) EM5 and (b) EM6 and associated XRF spectra highlighting the use of two different materials in the retouched areas. The spectra are also presented in linear scale (c) to further highlight the Ca, Ti and Zn elemental intensity variations that are reduced using the logarithmic scale often used.

2140 | *J. Anal. At. Spectrom.*, 2022, **37**, 2130–2143

This journal is © The Royal Society of Chemistry 2022

all presenting the same elemental composition (Zn, Pb, Co, Cr) with only variations in composition (Table 1). While it would have been possible to identify the various pigment mixtures based on elemental variations observed in the elemental maps created by PyMCA, the dictionary learning approach, by its ability to look at pixel-wise spectra and recognize patterns, appears a much easier and faster methodology. This can be a very interesting approach to differentiate between various artists having collaborated on a single painting, various stages of a painting using the same pigment mixture with variations in its composition, and various conservation campaigns, often using similar conservation materials. This is well illustrated by how the machine learning approach clearly identifies two types of complex materials containing mainly elements often associated with modern conservation materials (Ca/Ti/Fe/Mn/Zn, Fig. 11). Despite Ca and Ti found in both materials, the algorithm can differentiate the two materials as having different composition in Ti and Ca, one being Ti- and Zn-rich and the other Ca-rich. These observations provide extra understanding on the past conservation treatments undertaken on the artwork. Such level of information-digging is very difficult to achieve through classic elemental maps approach due to various factors, including the variation in dynamic range.

6. Conclusion

This study looked at the feasibility and advantages of machine-learning-based correlated elements mappings as an alternative processing of MA-XRF datasets. The algorithm was tested on a pigment mock-up board with known pigment mixtures and then applied to a historic painting from 18th century Mexico and presenting pigments with characteristic composition, pigments mixtures, and retouching areas. The open-access, user-friendly, fully available online machine learning approach, using the recently developed Julia programming language, for increased processing speed and larger data handling, showed great advantages over the traditional elemental map processing and pipelines using Python or R when it comes to identify pigments with characteristic composition such as smalt, composed of elements such as Co, Ni, As, Bi, and K. Identification of smalt with the traditional elemental map approach required the user to create the various elemental maps and compare them to one another to highlight areas of elemental co-presence when the machine learning approach create a single distribution map of the co-localized elements, identified through an associated XRF spectrum. As a result, no image manipulations to create the popular RGB composite images highlighting elemental co-presence is required, saving the user time and user-based oversights. While smalt was identified in this study, similar outcomes are expected for pigments of known specific elemental composition such as Emerald/Scheele's green (As/Cu), cadmium yellow/red (Cd/S/Se), cerulean blue (Co/Sn), cobalt green (Co/Zn), chrome yellow/orange (Pb/Cr), Egyptian blue (Cu/Ca/Si), lead tin yellow (Pb/Sn), and Naples yellow (Pb/Sb). The dictionary learning approach also allowed us to identify pigment mixtures, in our case mixtures of colored pigments (vermilion and most likely copper-based verdigris or copper resinate) with a white base (likely to be lead white), providing further information on the artists' technique. The novel machine learning approach also demonstrated its power in terms of highlighting elements which elemental counts in some areas do not allow the user to distinguish them easily as their brightness and contrast is quenched by the dynamic range of the distribution map. Finally, we were also able, using the machine learning approach, to extract more information about previous conservation treatment than the classic elemental map approach would allow us to without further data manipulation. This study demonstrates that machine learning, made available to all through the open-access data treatment pipeline, is a valuable tool for the data treatment of MA-XRF obtained on artworks and allows the user to extract more and more specific information than what is usually possible using the classic elemental mapping. However, it is important to note that this approach does not replace the user-based knowledge of the pigments and conservation materials required to contextualize the results.

Author contributions

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

References

1 K. Janssens, G. Van der Snickt, F. Vanmeert, S. Legrand, G. Nuyts, M. Alfeld, L. Monico, W. Anaf, W. De Nolf,

This journal is © The Royal Society of Chemistry 2022

J. Anal. At. Spectrom., 2022, **37**, 2130–2143 | **2141**

M. Vermeulen, J. Verbeeck and K. De Wael, *Top. Curr. Chem.*, 2016, **374**, 81.

2 A. Deneckere, L. de Vries, B. Vekemans, L. V. d. Voorde, F. Ariese, L. Vincze, L. Moens and P. Vandenabeele, *Appl. Spectrosc.*, 2011, **65**, 1281–1290.

3 A. Križnar, V. Muñoz, F. de la Paz, M. A. Respaldiza and M. Vega, *X-Ray Spectrom.*, 2011, **40**, 96–100.

4 A. G. Karydas, R. Padilla-Alvarez, M. Drozdenko, M. Korn and M. O. Moreno Guzmán, *X-Ray Spectrom.*, 2014, **43**, 138–145.

5 A. T. da Silva, S. Legrand, G. Van der Snickt, R. Featherstone, K. Janssens and G. Bottinelli, *Heritage Sci.*, 2017, **5**, 37.

6 J. K. Delaney, K. A. Dooley, A. van Loon and A. Vandivere, *Heritage Sci.*, 2020, **8**, 4.

7 A. Mazzinghi, C. Ruberto, L. Castelli, C. Czelusniak, L. Giuntini, P. A. Mandò and F. Taccetti, *Appl. Sci.*, 2021, **11**, 6151.

8 F. P. Romano, C. Caliri, P. Nicotra, S. Di Martino, L. Pappalardo, F. Rizzo and H. C. Santos, *J. Anal. At. Spectrom.*, 2017, **32**, 773–781.

9 G. Van der Snickt, S. Legrand, I. Slama, E. Van Zuien, G. Gruber, K. Van der Stighelen, L. Klaassen, E. Oberthaler and K. Janssens, *Microchem. J.*, 2018, **138**, 238–245.

10 G. Van der Snickt, A. Martins, J. Delaney, K. Janssens, J. Zeibel, M. Duffy, C. McGlinchey, B. Van Driel and J. Dik, *Appl. Spectrosc.*, 2016, **70**, 57–67.

11 P. Ricciardi, S. Legrand, G. Bertolotti and K. Janssens, *Microchem. J.*, 2016, **124**, 785–791.

12 A. Mazzinghi, C. Ruberto, L. Castelli, P. Ricciardi, C. Czelusniak, L. Giuntini, P. A. Mandò, M. Manetti, L. Palla and F. Taccetti, *X-Ray Spectrom.*, 2021, **50**, 272–278.

13 E. Pouyet, N. Barbi, H. Chopp, O. Healy, A. Katsaggelos, S. Moak, R. Mott, M. Vermeulen and M. Walton, *X-Ray Spectrom.*, 2021, **50**, 263–271.

14 L. Sottili, L. Guidorzi, A. Mazzinghi, C. Ruberto, L. Castelli, C. Czelusniak, L. Giuntini, M. Massi, F. Taccetti, M. Nervo, S. De Blasi, R. Torres, F. Arneodo, A. Re and A. Lo Giudice, *Appl. Sci.*, 2021, **11**, 1197.

15 M. Alfeld, M. Mulliez, P. Martinez, K. Cain, P. Jockey and P. Walter, *Anal. Chem.*, 2017, **89**, 1493–1500.

16 G. Van der Snickt, S. Legrand, J. Caen, F. Vanmeert, M. Alfeld and K. Janssens, *Microchem. J.*, 2016, **124**, 615–622.

17 S. Legrand, G. Van der Snickt, S. Cagno, J. Caen and K. Janssens, *Journal of Cultural Heritage*, 2019, **40**, 163–168.

18 A. Dal Fovo, A. Mazzinghi, S. Omarini, E. Pampaloni, C. Ruberto, J. Striova and R. Fontana, *Journal of Cultural Heritage*, 2020, **43**, 311–318.

19 M. Vadrucci, A. Mazzinghi, B. Sorrentino, S. Falzone, C. Gioia, P. Gioia, E. M. Loreti and M. Chiari, *X-Ray Spectrom.*, 2020, **49**, 668–678.

20 M. Alfeld, S. Pedetti, P. Martinez and P. Walter, *C. R. Phys.*, 2018, **19**, 625–635.

21 M. Vermeulen, D. Tamburini, E. M. K. Müller, S. A. Centeno, E. Basso and M. Leona, *Sci. Rep.*, 2020, **10**, 20921.

22 O. Dill, M. Vermeulen, A. McGeachy and M. Walton, *Heritage*, 2021, **4**, 1590–1604.

23 A. Couvrat Desvergnes, L. Sauvage, J. de Hond, P. D'Imporzano and M. Alfeld, *Heritage Sci.*, 2021, **9**, 119.

24 M. Vermeulen, K. Smith, K. Eremin, G. Rayner and M. Walton, *Spectrochim. Acta, Part A*, 2021, **252**, 119547.

25 M. Alfeld, W. De Nolf, S. Cagno, K. Appel, D. P. Siddons, A. Kuczewski, K. Janssens, J. Dik, K. Trentelman, M. Walton and A. Sartorius, *J. Anal. At. Spectrom.*, 2013, **28**, 40–51.

26 E. Pouyet, K. Brummel, S. Webster-Cook, J. Delaney, C. Dejoie, G. Pastorelli and M. Walton, *SN Appl. Sci.*, 2020, **2**, 1408.

27 S. Kogou, G. Shahtahmassebi, A. Lucian, H. Liang, B. Shui, W. Zhang, B. Su and S. van Schaik, *Sci. Rep.*, 2020, **10**, 19312.

28 E. Pouyet, N. Rohani, K. Katsaggelos Aggelos, O. Cossairt and M. Walton, *Pure Appl. Chem.*, 2018, **90**, 493.

29 M. Madden and A. Ryder, *Machine Learning Methods for Quantitative Analysis of Raman Spectroscopy Data*, SPIE, 2003.

30 V. Sevetlidis and G. Pavlidis, *Journal of Cultural Heritage*, 2019, **37**, 121–128.

31 Y. Chen, D. Wen, J. Pei, Y. Fei, D. Ouyang, H. Zhang and Y. Luo, *Current Opinion in Environmental Science & Health*, 2020, **18**, 14–19.

32 S. Misra, H. Li and J. He, in *Machine Learning for Subsurface Characterization*, ed. S. Misra, H. Li and J. He, Gulf Professional Publishing, 2020, pp. 129–155, DOI: **10.1016/B978-0-12-817736-5.00005-3**.

33 M. Vermeulen, L. Burgio, N. Vandeperre, E. Driscoll, M. Viljoen, J. Woo and M. Leona, *Heritage Sci.*, 2020, **8**, 62.

34 S. Kogou, L. Lee, G. Shahtahmassebi and H. Liang, *X-Ray Spectrom.*, 2021, **50**, 310–319.

35 A. Martins, J. Coddington, G. Van der Snickt, B. van Driel, C. McGlinchey, D. Dahlberg, K. Janssens and J. Dik, *Heritage Sci.*, 2016, **4**, 33.

36 A. Martins, C. Albertson, C. McGlinchey and J. Dik, *Heritage Sci.*, 2016, **4**, 22.

37 M. Elad and M. Aharon, *IEEE Trans. Image Process.*, 2006, **15**, 3736–3745.

38 R. Mak, M. Lerotic, H. Fleckenstein, S. Vogt, S. M. Wild, S. Leyffer, Y. Sheynkin and C. Jacobsen, *Faraday Discuss.*, 2014, **171**, 357–371.

39 M. Lerotic, R. Mak, S. Wirick, F. Meirer and C. Jacobsen, *J. Synchrotron Radiat.*, 2014, **21**, 1206–1212.

40 M. Alfeld, M. Wahabzada, C. Bauckhage, K. Kersting, G. Wellenreuther, P. Barriobero-Vila, G. Requena, U. Boesenberg and G. Falkenberg, *J. Synchrotron Radiat.*, 2016, **23**, 579–589.

41 N. Rohani, J. Salvant, S. Bahaadini, O. Cossairt, M. Walton and A. Katsaggelos, *Proc. 24th European Signal Processing Conference (EUSIPCO)*, 2016, pp. 2111–2115.

42 Q. Dai, E. Pouyet, O. Cossairt, M. Walton and A. K. Katsaggelos, *IEEE Transactions on Computational Imaging*, 2017, **3**, 432–444.

43 M. Elad, *Sparse and Redundant Representations: from Theory to Applications in Signal and Image Processing*, Springer New York, New York, NY, 1st edn, 2010.

44 C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*, SIAM Society for Industrial and Applied Mathematics SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104, 1995.

45 V. A. Solé, E. Papillon, M. Cotte, P. Walter and J. Susini, *Spectrochim. Acta, Part B*, 2007, **62**, 63–68.

46 D. Moura, *R vs. Python vs. Julia*, Toward Data Science, 2021, vol. 2022.

47 J. Bezanson, S. Karpinski, V. B. Shah and A. Edelman, *Julia: A Fast Dynamic Language for Technical Computing*, 2012.

48 J. M. Perkel, *Nature*, 2019, **572**, 141–142.

49 B. Xu, Y. Wu, P. Hao, M. Vermeulen, A. McGeachy, K. Smith, K. Eremin, G. Rayner, G. Verri, F. Willomitzer, M. Alfeld, J. Tumblin, A. Katsaggelos and M. Walton, arXiv:2207.12651, 2022.

50 P. P. Rubens, Franciscan Allegory in Honor of the Immaculate Conception, *Philadelphia Museum of Art #Cat. 677*, John G. Johnson Collection, 1917.

51 M. Alfeld, V. Gonzalez and A. van Loon, *X-Ray Spectrom.*, 2021, **50**, 351–357.

52 F. Li, C. Tang, H. Li and L. Ge, *X-Ray Spectrom.*, 2020, 1–9.

53 C. G. Ryan, E. Clayton, W. L. Griffin, S. H. Sie and D. R. Cousens, *Nucl. Instrum. Methods Phys. Res., Sect. B*, 1988, **34**, 396–402.

54 J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J.-Y. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak and A. Cardona, *Nat. Methods*, 2012, **9**, 676–682.

55 S. Preibisch, S. Saalfeld and P. Tomancak, *Bioinformatics*, 2009, **25**, 1463–1465.

56 M. L. Clarke, F. Gabrieli, K. L. Rowberg, A. Hare, J. Ueda, B. McCarthy and J. K. Delaney, *Heritage Sci.*, 2021, **9**, 20.

57 M. Aceto, A. Agostino, G. Fenoglio, A. Idone, M. Gulmini, M. Picollo, P. Ricciardi and J. K. Delaney, *Anal. Methods*, 2014, **6**, 1488–1500.

58 J. K. Delaney, P. Ricciardi, L. D. Glinsman, M. Facini, M. Thoury, M. Palmer and E. R. d. l. Rie, *Stud. Conserv.*, 2014, **59**, 91–101.

59 A. Coccato, L. Moens and P. Vandenabeele, *Heritage Sci.*, 2017, **5**, 12.

60 A. van Loon, P. Noble, D. de Man, M. Alfeld, T. Callewaert, G. Van der Snickt, K. Janssens and J. Dik, *Heritage Sci.*, 2020, **8**, 90.

61 L. Appolonia, D. Vaudan, V. Chatel, M. Aceto and P. Mirti, *Anal. Bioanal. Chem.*, 2009, **395**, 2005–2013.

62 A. Phenix and A. Roy, *Artists' Pigments: A Handbook of Their History and Characteristics, Volume 2*, 1995, vol. 40, pp. 69–70.

63 K. Janssens, G. Van Der Snickt, M. Alfeld, P. Noble, A. van Loon, J. Delaney, D. Conover, J. Zeibel and J. Dik, *Microchem. J.*, 2016, **126**, 515–523.

64 L. Robinet, M. Spring and S. Pagès-Camagna, Investigation of the loss of colour in smalt on degradation in paintings using multiple spectroscopic analytical techniques, in, *Proc ICOM Committee for Conservation 16th triennial meeting*, Lisbon, Portugal, 2011.

65 P. Santopadre and M. Verità, *Stud. Conserv.*, 2006, **51**, 29–40.

This journal is © The Royal Society of Chemistry 2022

J. Anal. At. Spectrom., 2022, **37**, 2130–2143 | 2143