Application of Uniform Manifold Approximation and Projection (UMAP) in Spectral Imaging of Artworks

Article *in* Spectrochimica Acta Part A Molecular and Biomolecular Spectroscopy · February 2021

DOI: 10.1016/j.saa.2021.119547

5 authors, including:

Marc Vermeulen
Northwestern University
21 PUBLICATIONS **209** CITATIONS
SEE PROFILE

Kate Smith
Harvard University
5 PUBLICATIONS **21** CITATIONS
SEE PROFILE

Katherine Eremin
Harvard University
64 PUBLICATIONS **1,077** CITATIONS
SEE PROFILE

Georgina Rayner
Harvard University
9 PUBLICATIONS **35** CITATIONS
SEE PROFILE

Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 252 (2021) 119547

Contents lists available at ScienceDirect

Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy

journal homepage: www.elsevier.com/locate/saa

Application of Uniform Manifold Approximation and Projection (UMAP) in spectral imaging of artworks

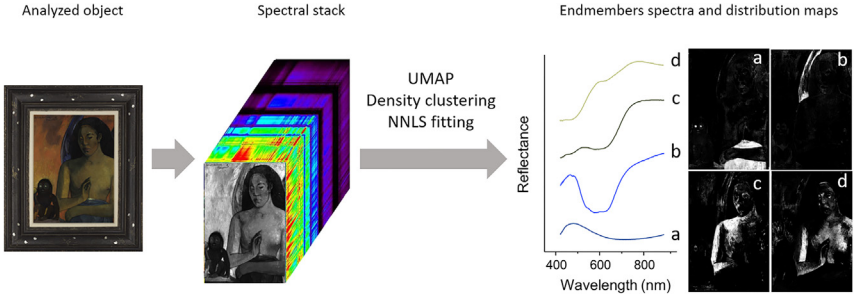
Marc Vermeulen [a], Kate Smith [b], Katherine Eremin [b], Georgina Rayner [b], Marc Walton [a,*]

[a] Northwestern University / Art Institute of Chicago Center for Scientific Studies in the Arts (NU-ACCESS), 2145 Sheridan Road, Evanston, IL, United States
[b] Harvard Art Museums, Straus Center for Conservation and Technical Studies, 32 Quincy St, Cambridge, MA, United States

HIGHLIGHTS

- Software pipeline using UMAP is described to reduce and visualize a complex spectral dataset from a Gauguin Paintings.
- Compared to t-SNE UMAP is fast and preserves the global vs. local structure balance of the data.
- Python scripts are used to extracts endmembers and produce pigment distribution maps via non-negative least square fitting.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 18 November 2020
Received in revised form 22 January 2021
Accepted 24 January 2021
Available online 4 February 2021

Keywords:
Hyperspectral imaging
Data reduction and visualization
Multivariate analysis
UMAP
Cultural heritage

ABSTRACT

This study assesses the potential of Uniform Manifold Approximation and Projection (UMAP) as an alternative tool to t-distributed Stochastic Neighbor Embedding (t-SNE) for the reduction and visualization of visible spectral images of works of art. We investigate the influence of UMAP parameters—such as, correlation distance, minimum embedding distance, as well as number of embedding neighbors— on the reduction and visualization of spectral images collected from *Poèmes Barbares* (1896), a major work by the French artist Paul Gauguin in the collection of the Harvard Art Museums. The use of a cosine distance metric and number of neighbors equal to 10 preserves both the local and global structure of the Gauguin dataset in a reduced two-dimensional embedding space thus yielding simple and clear groupings of the pigments used by the artist. The centroids of these groups were identified by locating the densest regions within the UMAP embedding through a 2D histogram peak finding algorithm. These centroids were subsequently fit to the dataset by non-negative least square thus forming maps of pigments distributed across the work of art studied. All findings were correlated to macro XRF imaging analyses carried out on the same painting. The described procedure for reduction and visualization of spectral images of a work of art is quick, easy to implement, and the software is opensource thus promising an improved strategy for interrogating reflectance images from complex works of art.

1. Introduction

Hyperspectral imaging (HSI) enables non-invasive and non-destructive spatial investigation of pigments and colorants found in artworks, either pure or in mixture to reach the hue intended by the artist. At typical spatial resolutions, a single HSI experiment produces gigabytes of data composed of millions of reflectance spectra. Such large datasets pose challenges for extracting a maximum of information. It is clear that traditional methods of spectral analysis (*e.g.*, as is routinely performed in fiber optic reflectance spectroscopy [1–15]), that require the careful examination of each

* Corresponding author.
E-mail address: marc.walton@northwestern.edu (M. Walton).

https://doi.org/10.1016/j.saa.2021.119547
1386-1425/

individual spectrum, is not possible. Because of the large volume and complexity of these data, there is an increasing need for advanced dimensionality reduction techniques, capable of automatically identifying trends and clustering data.

Visualizing the structure of data is a key exploratory step. Depending on the dataset under study, the visualization of the data may take several forms such as scatterplots, histograms/distributions, or tree maps [16–20]. Many of these visualization tools are difficult to implement on multi-dimensional data. Consequently, it becomes important to reduce dimensionality by grouping together comparable features— reflectance spectra in the case of HSI— prior to visualization, allowing insights into identification and localization of pigments used by the artists. This is made possible by embedding the large number of spectra collected into a lower dimensional space that retains the information and properties of the original high dimensional dataset.

There are several data reduction methods already widely available. Amongst these, Principal Component Analysis (PCA) has historically been the most commonly used [19–27]. PCA projects spectral data onto a variance subspace [28] which retains the global structure of the dataset by linearly recombining all spectra into orthogonal eigenvectors. Local structure (*e.g.*, spectrum to spectrum differences) within a given dataset, however, will not be preserved by PCA [16]. Also, while linear dimensionality reduction techniques, such as PCA, have shown satisfactory results in the past for extracting information from HSI data [24,29–31], these models presuppose that a linear relationship exists amongst the variables despite the fact that paint mixtures models are inherently nonlinear [32]. The non-linear chromatic behavior of paint blending and layering is well illustrated by the Kubelka-Munk (K-M) model, for which the K-M equations accurately approximate the diffuse reflectance of pigmented materials like paint, given descriptions of their constituent pigments and pigment concentrations [33–35].

To overcome the limitations of PCA, nonlinear data reduction techniques have been developed that better preserve local structure by embedding spectra with small distances as nearby points in a low-dimensional graphical representation [36–38]. t-distributed Stochastic Neighbor Embedding (t-SNE) is one such method that has proven to be a very valuable tool for HSI data analysis [39–41]. The t-SNE algorithm evaluates the similarity between one point and a given number of neighbors by calculating their distance and modelling a pair-wise probability distribution. t-SNE is a valuable dimensionality reduction tool because the number of features that can be captured in the reduced space is not restricted by the number of output dimensions selected (often set to 2 or 3). Furthermore, t-SNE maintains the local distances of the original dataset making it a good contender for data visualization. Due to these attributes, t-SNE has become a very popular data reduction technique well suited for the visualization of high-dimensional datasets [42–46]. A major drawback of t-SNE and many other nonlinear dimensionality reduction techniques, however, is their memory intensive computation. In t-SNE, a large distance matrix needs to be calculated between pixels, which consumes available RAM and makes the processing of high-resolution HSI datasets slow, at least with standard desktop and laptop computers [36,47]. Furthermore, t-SNE does not retain the global structure of the original dataset. As a result, two groupings far away from each other in the embedded space are not necessarily far away in the original data. As a general rule, distances between embedded groupings in a t-SNE plot lack meaning.

In the case of HSI data, however, finding the distances between groupings is desirable as this information could indicate similarity or dissimilarity in pigments or reveal information on pigment mixtures. As a result, there is a need for novel data reduction techniques that retain both the local and global geometric structures of the initial dataset and provides an easy way to interpret visualization. To address this gap, Uniform Manifold Approximation and Projection (UMAP) was recently developed [36,47,48]. Similarly to t-SNE, UMAP uses graph layout algorithms to arrange data in a low-dimensional space [49]. The algorithm finds an embedding by searching for a low-dimensional projection of the data that has the closest possible equivalent global shape and structure as the original dataset. This method can be used for visualization purposes in two or three dimensions. Another great advantage of UMAP over t-SNE is its faster processing time, and ability to handle larger data set [50,51]. As an example, it was reported in literature that UMAP processes 429,165 pixels with 399 bands in 857.47 s, while t-SNE employs 2905.28 s for the same dataset [52].

Since its introduction in 2018, UMAP has found many applications in the fields of bioinformatics, material and environmental sciences, and machine learning [36,38,47,51,53–60]. Yet the application of UMAP to heritage science has hitherto been limited and, when used, the optimization of its various parameters has not been explained in detail [52]. In this paper, we assess the potential of UMAP for the reduction and visualization of hyperspectral data obtained on works of art and compare its performance to t-SNE. We also systematically evaluate the influence of various UMAP parameters (*e.g.*, distance metric and number of neighbors) on the reduction and visualization of hyperspectral data obtained on works of art. Furthermore, we develop an endmember extraction pipeline which utilizes 2D histogram density maps to identify groupings and their centroids. Endmember distribution maps are made by fitting the centroid spectra to the original data using non-negative least squares. A Jupyter Notebook containing these scripts is freely available from the Center for Scientific Studies in the Arts (NU-ACCESS) Github page (https://github.com/NU-ACCESS/UMAP).

2. Materials and methods

2.1. Dataset

The historical data set used in this study was obtained on one painting by Paul Gauguin (1848–1903) from the collection of the Harvard Art Museums. The painting analyzed, *Poèmes Barbares* (Fig. 1), is dated 1896 and was painted during Gauguin's second trip to French Polynesia. The final dataset, 750 Mb in size, is composed of 1,590,292 pixels (1034 × 1538 pixels), therefore yielding a ca. 500-μm spatial resolution, common characteristics for reflectance imaging data sets.

2.2. Hyperspectral data acquisition

HSI data in the visible range was acquired using a Resonon Pika II Pushbroom system (Resonon, Inc., Bozeman, MT, USA) in the 400–900 nm range with spectral resolution of 2.1 nm, with a total of 240 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 460 × 420 μm^2. 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. Hyperspectral 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 six partially overlapped areas, each with a size of 230 × 240 mm^2, were stitched together using registration and stitching plugins available in the open source image processing package Fiji suite [61,62], prior to further processing.

Fig. 1. "Poèmes Barbares" (1896), oil on canvas, 64.8 × 48.3 cm (unframed), painted by the French artist Paul Gauguin (1848–1903), Harvard Art Museums/Fogg Museum, Bequest from the Collection of Maurice Wertheim, Class of 1906. Object Number: 1951.49 © President and Fellows of Harvard College.

2.3. MA-XRF data acquisition

The macro XGLab's ELIO XRF imaging spectrometer system (MA-XRF) was used in combination with HSI for characterizing the pigments palette used in the Gauguin painting. The instrument is equipped with a transmission Rh anode X-Ray tube, the polychromatic beam presenting an incoming angle of 63.5° prior to the sample plane, and a compact head free of X-ray optics. A collimator allowing a 1 mm diameter focused spot size at the surface of the object was used to acquire XRF maps at the surface of the painting. Two laser pointers, mounted in such a way that their intersection point coincides with the cross-point of the incident X-ray beam and detector axis, allow for optimizing both excitation and detection conditions. The X-ray detector element is a large area (active collimated area is 25 mm^2) silicon drift detector (SDD) equipped with a CUBE preamplifier, with an energy resolution of 135 eV at the Mn Kα line (5.9 keV). The instrument was operated at 50 kV and 60 μA. The elemental 2D mapping of the object surface was achieved using a 100 × 100 mm^2 automatic XY raster scanning stage mounted to a homebuilt 560 × 400 mm^2 two-dimensional motorized scanner. This macro-XRF system allowed the scanning of approximately 70% of the total surface of the painting. Rastering was executed with acquisition times of 0.5 s per point and with a step size of 2 × 2 mm^2. The various maps were stitched together using registration and stitching plugins available in the open source image processing package Fiji suite [61,62].

2.4. Data processing

Manifold learning approaches were used to embed high-dimensional data into a 2-dimensional space for visualization and investigation of nonlinear relations in the data. While the main aim of this article is to assess the application of UMAP for the data reduction and visualization of hyperspectral imaging of works of art, t-SNE (perplexity of 50, [40]) was applied to the same datasets in order to compare the visualization outcomes and the running times of both approaches. In the frame of this study, the data was reduced to 2 components for both the t-SNE and UMAP experiments.

UMAP was performed in a Jupyter Notebook running Python 3 [48]. Default parameters were used except for number of neighbors, distance, and the type of distance measured (*e.g.*, Euclidean, cosine, Manhattan, etc.). Since the reduction of hyperspectral data aims at visualizing similar reflectance curves, it is necessary to set the minimum distance parameter to 0. This parameter, as the name suggests, corresponds to the minimum distance between each point within the graph. When set to 0, points corresponding to similar spectra will be placed as close together as possible (on top of each other if identical) creating high density regions of similar spectra within the two-dimensional space of the graph. This densification of points is of foremost importance for the histogram clustering methods we employ in the endmember selection steps indicated below.

One of the first steps in data reduction and visualization is calculating distances between the spectra in the original high-dimensional space. Distance describes how similar a spectrum is to all other spectra and choice of the distance metric can have a considerable impact on the performance of UMAP to group similar spectra [56,63]. While the Euclidean distance is often used by default, including in t-SNE [42,53,56,64], the distance metrics evaluated in this study included Canberra, Chebyshev, cosine, Euclidean, Manhattan, and Minkowski, as they have been used previously in reduction of different data types [47,51,54,56,65]. The detailed formulas for each distance metric are given in Table S1. Also assessed is the influence of the number-of-neighbors metric which controls how UMAP balances local versus global structure in the data by constraining the size of the local neighborhood. As a result, low number-of-neighbors values will force UMAP to concentrate on very local structure, whereas large values will push UMAP to look at larger neighborhoods of each point when estimating the manifold structure of the data. The distance and number-of-neighbors metrics used for the various experiments are indicated in the figures' captions.

The RGB colors used in the UMAP and t-SNE embeddings plots are calculated from the reflectance spectrum themselves. First, a Python script implemented in the Fiji image processing suite, collapses the wavelength stack into a 3-band XYZ tristimulus image by multiplying each wavelength by a CIE color matching function and summing across all wavelengths. The XYZ image is then transformed into the Adobe sRGB color space as detailed previously [66,67]. By performing these actions, a per pixel RGB image is thus registered to each reflectance spectrum in the data cube. The RGB value of each pixel from the RGB image is then used to color and plot the UMAP embeddings.

For the determination of endmembers from UMAP, embeddings were converted into a 2D histogram by binning pixels to 256. Areas of high density were identified from the histogram based on user defined threshold (sensitivity to counts of normalized pixels) and nearest-neighbor values (a constraint that indicates how far away a group needs to be - in terms of pixels - to be considered as a new group). Regions of highest density were considered as cluster centroids and the associated spectra were thus taken as endmembers. This method will not differentiate single pigment and mixtures of pigments and therefore, the endmembers identified will be either pure pigments or pigment mixtures representative of what the artists used in their compositions. Finally, these endmember spectra were fitted to the original spectral data cube using non-negative least squares to produce pigment distribution maps.

UMAP and t-SNE reduction and projection were carried out on a Dell computer running with Windows 10 with a 1.80 GHz Intel Core i7 processor, with 16 GB of RAM. This setup allowed for the UMAP treatment of datasets up to 250,000 pixels. When reduction and projection tests required more memory, a Dell computer running with Windows 10 with 3.6 GHz Intel Xenon processor and 72 GB of RAM was used.

3. Results and discussion

3.1. Description of the analytical procedure

The flowchart presented in Fig. 2 illustrates the processing pipeline used for data reduction, identification of pigment clusters, and visualization of pigment distribution in the studied artworks. Data pre-processing steps, such as the transforming of the raw "band interleaved by line" (BIL) data formats into TIFF files, are not included in the Jupyter Notebook associated with this study but may be found on the previously mentioned NU-ACCESS Github page.

3.2. Dataset preparation prior to data reduction

As the first step, the data was reorganized into an n x m matrix in which n represents the number of wavelength channels and m the number of pixels. This step also provides the opportunity for the investigator to remove any zero lines artificially produced when stitching together two dissimilarly sized data cubes acquired on large objects under study.

The next step is to reduce the size of the spectral stack which can become quite large (up to several Gb/several millions of pixels) depending on the area scanned. Reducing spatial dimensions may be required, for instance, when using a computer with limited RAM which can be consumed by the large matrices utilized by UMAP. One option is to downsize the spatial dimensions of the data cube by binning pixels (averaging) as can be readily accomplished in imaging processing suits like Fiji or Photoshop. However, since averaging can produce undesirable artifacts [40], it is not optimal when the ultimate goal of the data reduction is to identify the signatures of the "pure pigments" used by the artist. We find that a better approach is to randomly select a percentage of pixels which are expected to represent the variance found in the full data cube, using a Python implementation of the random sample module. The exact percentage can be customized to fit constraints of calculation time and computer hardware. This stochastic method produces UMAP results very similar to what is obtained for the full non-reduced cube thus demonstrating its robustness in maintaining data fidelity. Finally, not all wavelength channels may be required for data processing, especially when portions of the spectrum are affected by poor signal to noise ratios (SNR) or stray light within the spectrometer. Therefore, our code offers the option to choose wavelength ranges. Typically, UMAP is run with the noisy UV portion of the spectrum eliminated using only 230 spectral bands between 414 and 892 nm.

3.3. Performance comparison between UMAP and t-SNE

The performance of the UMAP algorithm is compared, using its default values (2 components, number of neighbors of 15 and Euclidean distance metric, [68]), to the t-SNE algorithm using the same optimized parameters selected by Pouyet et. al for spectral data (2 components, perplexity of 50 and Euclidean distance metric, [40]). Both UMAP and t-SNE were applied to the analysis of the Gauguin's *Poèmes Barbares*. The resulting 2D color scatterplots are given in Fig. 3a,b.

Points in the t-SNE scatterplot (Fig. 3a) appear more dispersed than in the UMAP scatterplot (Fig. 3b). This is well illustrated by the orange points associated with the orange-red background of the painting. In the t-SNE scatterplot, these points are found both as a large diffuse group in the upper half of the plot and as smaller and denser groupings around it. In comparison, all of the orange points appear as a single dense grouping in the UMAP scatterplot.

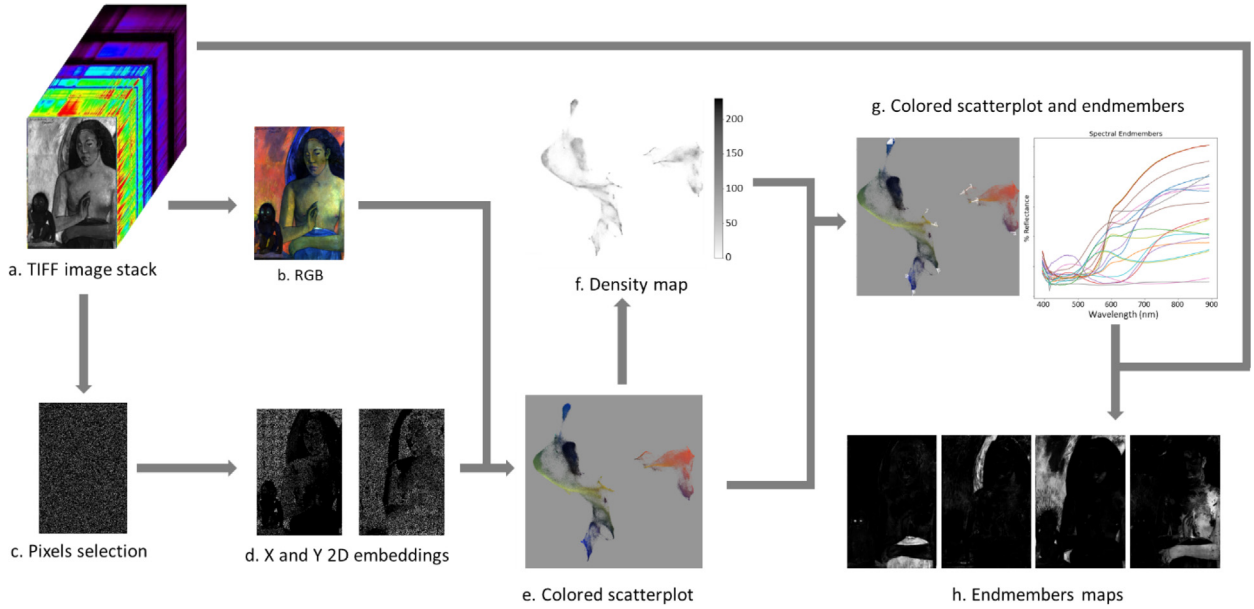


Fig. 2. Processing pipeline for reduction and visualization of hyperspectral data followed by the extraction and mapping of the various endmembers (pigments/colorants): (a) TIFF image stack resulting from the data collection and conversion from the studied work of art, (b) RGB image obtained from the TIFF image stack using the lambda stack to XYZ and XYZ to RGB macros running in Fiji, (c) pixel selection following the transformation of the image stack into a 2D matrix, (d) X and Y 2D embeddings obtained following the UMAP data reduction, (e) colored scatterplot obtained through the pixel correlation between the 2D embedding and the RGB image of the studied artwork, (f) 2D histogram density map allowing for the (e) characterization dense clusters/spectral endmembers, and (h) endmembers distribution maps created through a non-negative least square fitting of the selected endmembers to the original image stack (a).

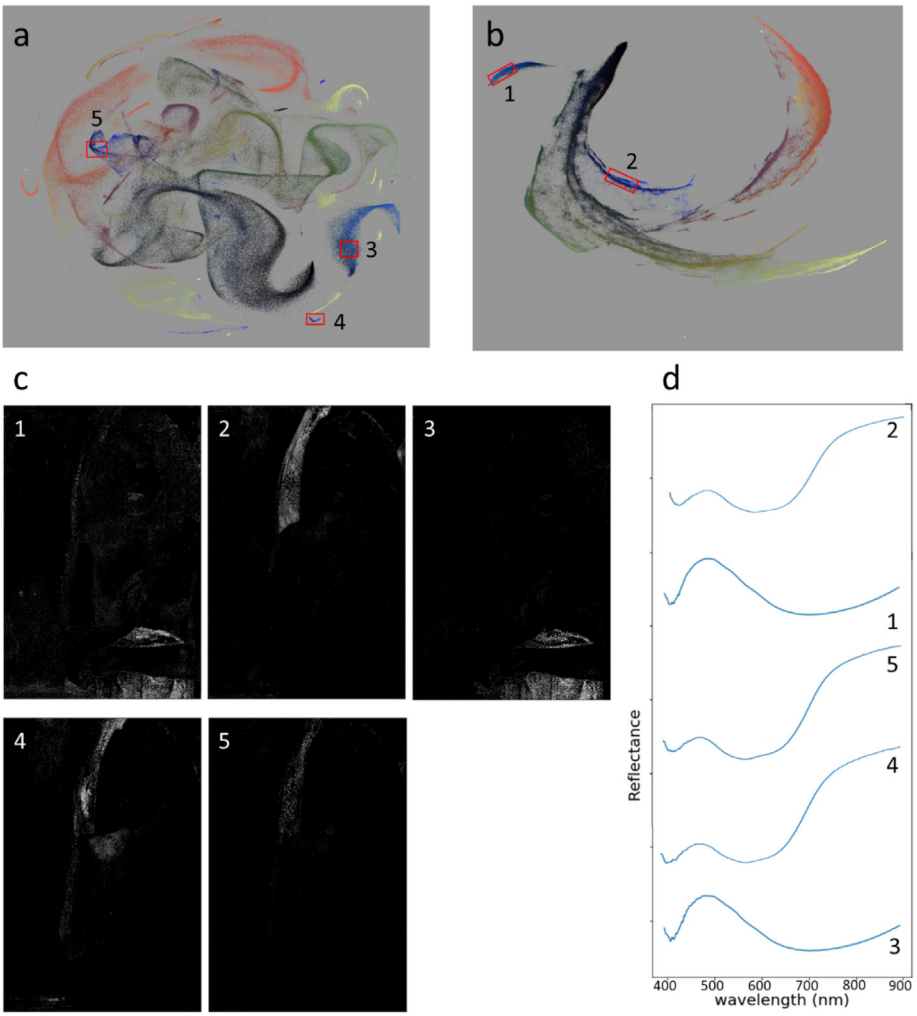


Fig. 3. Colored 2D scatterplot representation of the embeddings obtained on the Gauguin's Poèmes Barbares dataset (15% of the total pixels) using (a) t-SNE (Euclidean distance metric, perplexity of 50) and, (b) UMAP (Euclidean distance metric and number of neighbors of 15). (c) selected extracted endmembers maps for blue groupings marked with red rectangles on the 2D UMAP colored scatterplots and (d) corresponding reflectance spectra. The various blues were identified as Prussian blue (1 and 3) and ultramarine blue (2, 4, and 5) based on their reflectance curves presented in (d).

The main pigment associated with these groupings is mercuric sulfide (vermilion, HgS), as suggested by the mercury (Hg) map obtained with MA-XRF (Fig. 4-Hg). Therefore, based on the characteristic sigmoidal shape the HgS reflectance spectra with inflection point around 600 nm [1,3], one would expect to find all spectra with this feature to be grouped closely together as was correctly observed in the UMAP scatterplot. This clearly illustrate some of the previously discussed shortcomings of t-SNE compared to UMAP [36,47,51,56], which also apply to hyperspectral data obtained on works of art.

Likewise, for the blues, UMAP exhibits only two main groupings (Figs. 3b-1 and 3b-2) corresponding to the two shades applied to the wing behind the female figure's head and her garment (Figs. 3c-1 and 3c-2). Through comparison of the reflectance spectra (Figs. 3d-1 and 3d-2) with published databases [1,3,16,69], these groups were identified as Prussian blue, in the garment (Fig. 3c-1) and ultramarine used for the wings (Fig. 3c-2). The t-SNE plot, however, produced multiple scattered groupings (around 7) for the blues (Fig. 3a). When examining the spectra from hand selected groupings just associated with the female figure's head (Fig. 3c-4-5), for instance, it may be observed that each group is formed from spectra with the characteristic shape of ultramarine (Fig. 3d-4-5), with only subtle differences due to noise and intensity variations. However, by separating these groupings in the 2D scatterplot, t-SNE may tend to indicate that all these groupings may have been realized using different pigments, which is not the case based on the extracted reflectance curves. We believe that the more complex scatterplot produced by t-SNE is due to the lack of global structure, when using this technique, which places a suboptimal emphasis on noise. As a result, it becomes difficult to narrow down the number of blue pigments used in the painting by t-SNE and highlights the poor clustering potential of this method compared to UMAP.

UMAP may consolidate the data into fewer groupings, but each grouping has additional localized structure such as color gradients: *e.g.*, from light to dark orange, from yellow/light green to deep green and black, from light blue to dark blue. These gradients can be explained by the use of mixtures of pigments – dilution with white or black to achieve the various hues - rather than pure pigments to yield the desired color. This color arrangement, that is clearly not observed in t-SNE, can be of foremost importance in understanding the artistic process of pigment mixing. This also shows one of the primary differences between the t-SNE and UMAP: t-SNE retains only local structure (all neighboring points within a grouping are related) while global structure is lost (distance between clusters is meaningless). UMAP on the other hand

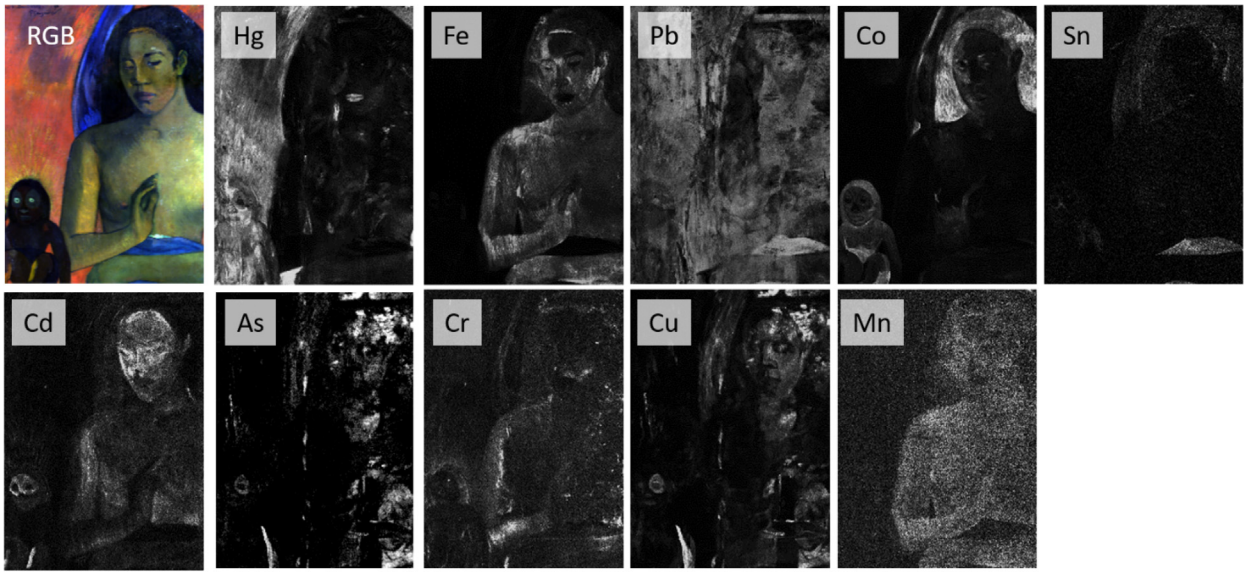


Fig. 4. RGB image and selected MA-XRF elemental maps representative of the composition.

preserves both local and global structure [36,48,51,70]. Therefore, all neighboring points within a grouping will be related while the distribution of the groupings within the 2D representation and their distance to one another will also bear meaning. For instance, two groups of points that are found close in the original complex data should also be found close in the reduced space following the UMAP reduction. It follows that the blue grouping observed curving between the black and the purple in Fig. 3b-2 most likely corresponds to the blue found on the upper middle part of the painting, located between the dark hair of the female figure and the purple from the background. This is confirmed by the distribution maps for ultramarine blue presented in Fig. 3c-2. Therefore, UMAP arranges the reduced data in a way that makes it easier to understand what pigments the artist has been using in the composition of the painting. In addition, toward the better data reduction and visualization capabilities of UMAP, a comparison of the runtimes obtained by the different algorithms (t-SNE and UMAP) was realized for the Gauguin dataset. Runtimes, 7000 sec vs. 300 sec for t-SNE and UMAP respectively, proves that UMAP runs about 20 times faster than t-SNE. Therefore, along with delivering embeddings with superior readability than t-SNE, UMAP is also computationally more efficient, as often described in published literature [48,59,60,71].

Overall, UMAP shows good runtime performance and, in comparison to t-SNE, has the major advantage of compressing the spatial and molecular information into tighter, more defined, and more meaningful groupings, resulting in easier detailed visualization and interpretation.

3.4. Influence of the distance metric and normalization

To assess the influence of the distance metric on the data reduction of hyperspectral imaging of Gauguin's *Poèmes Barbares*, the number of neighbors was set to its default value of 15.

Fig. 5 presents the UMAP color scatters obtained using Canberra (Fig. 5a), Chebyshev (Fig. 5b), cosine (Fig. 5c), Euclidean (Fig. 5d), Manhattan (Fig. 5e), and Minkowski (Fig. 5f) distance metrics.

All distance metrics lead to a clear grouping of the various colors. However, the projections obtained using Canberra, Euclidean, Manhattan, and Minkowski, all based on L_1 or un-normalized L_2 norms, appear to be very similar (Fig. 5a,d,e,f). On the contrary, Chebyshev and cosine distance metrics, respectively L_∞ and normalized L_2 norms, yield unique projections (Fig. 5b,c). With each metric there is a clear separation of the red, purple, two distinct blues, black, green, and yellow colors of the painting. Nonetheless, purples and greens, and to a lesser extent the blues, appear more scattered when the Canberra, Euclidean, Manhattan, and Minkowski distance metrics are used, whereas all colors appear more tightly grouped when Chebyshev or cosine distance metrics are used. In all cases, a color arrangement can be observed and each color groupings present a gradient, which appear to be independent from the distance metric used.

When looking at the associated density maps, it becomes clear than certain distance metrics allow denser groupings indicating a better data reduction step and potentially better endmember identification through density-based cluster identification algorithm. Looking solely at the highest density groupings, Canberra, Euclidean, Manhattan and Minkowski, all L_1 and un-normalized L_2 distance metrics, would appear as the most promising metrics to use, yielding small groupings as dense as 400 (Fig. 5a), 400 (Fig. 5d), 480 (Fig. 5e) and 350 pixels (Fig. 5f) respectively compared to 225 (Fig. 5b) and 260 pixels (Fig. 5c) for Chebyshev and cosine respectively. However, when considering the rest of the points, the groupings are found in the 50–100 pixels range for all L_1 and un-normalized L_2 distance metrics whereas Chebyshev and cosine present more groupings in the 100–200 pixels. Consequently, these distance metrics (cosine and Chebyshev) appear to form, overall, denser groupings than all the other distance metrics. As a result, cosine and Chebyshev may appear to be among the best distance metrics to use for global endmembers extraction based on cluster density as they provide a more comprehensive assessment of the artists' palette. The other L_1 and un-normalized L_2 norms metrics would identify a few high-density groupings but miss much of the palette. Due to the grouping nature of the data reduction process, the tighter and more well-defined groupings observed for cosine and Chebyshev distance metrics tend to indicate that these metrics are the most suitable for the purpose of data reduction of hyperspectral data obtained on art materials. As a result, these distance metrics were chosen as possible distance metric to be used with UMAP for data reduction of hyperspectral cubes.

While spectral data acquired is normalized against a diffuse white standard, fluctuations in intensity may be observed when pigments are found mixed in various proportions with a white pigment to yield the intended hues. Because the aim here is not to

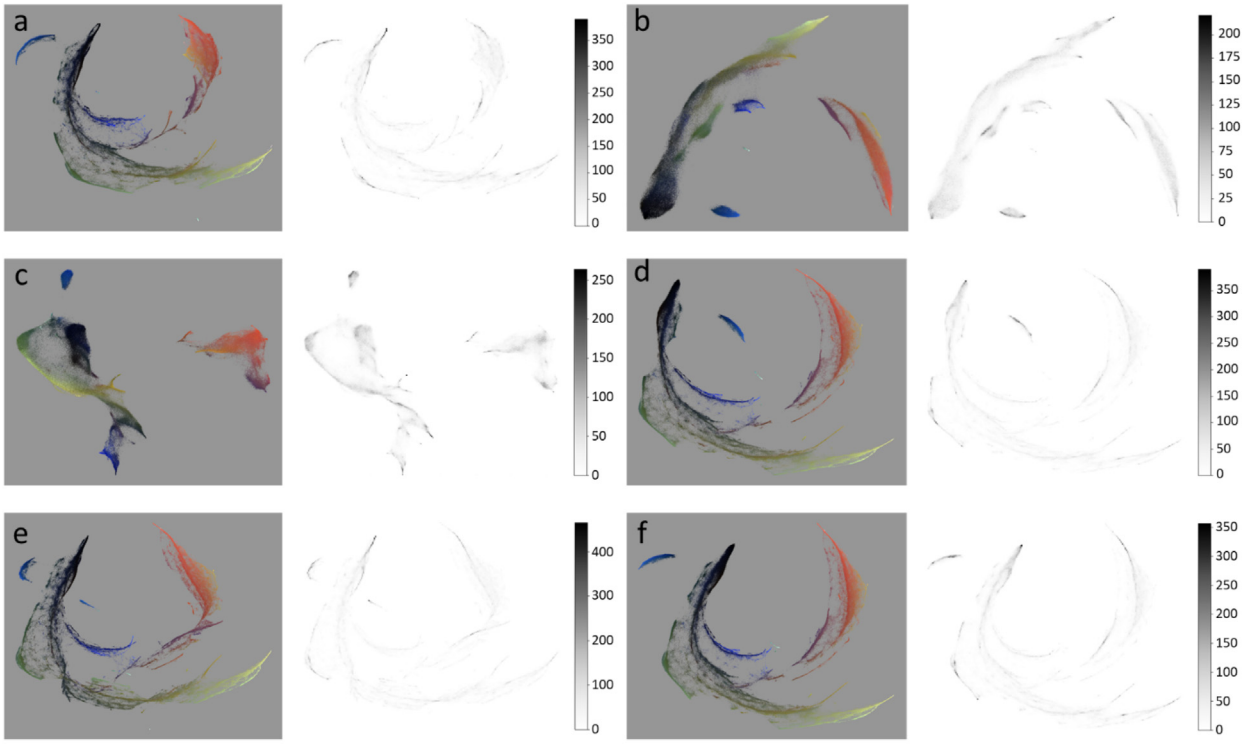


Fig. 5. UMAP 2D color scatterplots and associated density maps obtained for the Gauguin's Poèmes Barbares dataset using six different metrics (a) Canberra, (b) Chebyshev, (c) cosine, (d) Euclidean, (e) Manhattan and (f) Minkowski. Other UMAP parameters were set as follow: n_neighbors = 15, min_dist = 0, and n_components = 2.

perform a non-linear unmixing of the pigments to understand the exact composition of the mixtures [32,72], such intensity variations are less important than identifying the pigments in the mixtures themselves. Consequently, pixel-by-pixel normalization of the spectral data between 0 and 1 was undertaken to reduce the number of groupings. UMAP reduction of Gauguin's *Poèmes Barbares* and mapping of the cluster density were realized using previously shortlisted distance metrics: cosine, Chebyshev and Euclidean (the latter being used for comparison purposes with L_1 distance metrics). Their scatterplots and associated density maps are given in Fig. 6. As illustrated, normalization of the spectral stack prior to reduction and visualization does not have any impact on the UMAP embedding when using the cosine distance metric (Fig. 6a) when compared to the non-normalized cube (Fig. 5c). This

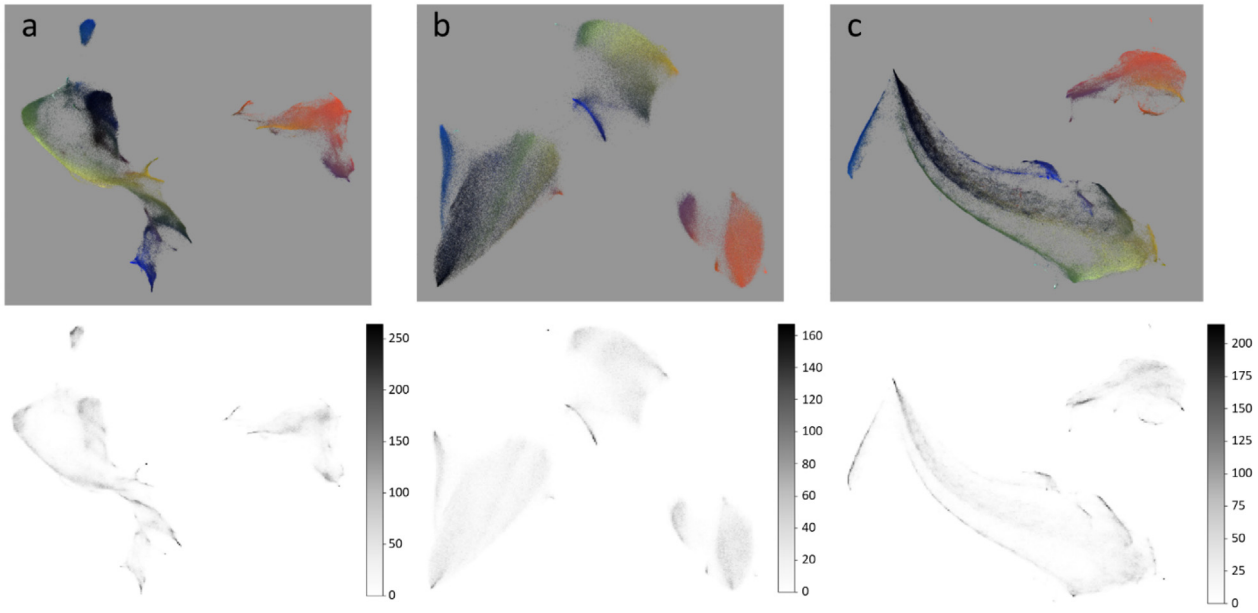


Fig. 6. UMAP 2D scatterplots and associated density maps for the Gauguin historical painting normalized dataset using (a) cosine, (b) Chebyshev and (c) Euclidean distance metrics. A number of neighbors of 10 and 15% of the total pixels were used to create the embeddings.

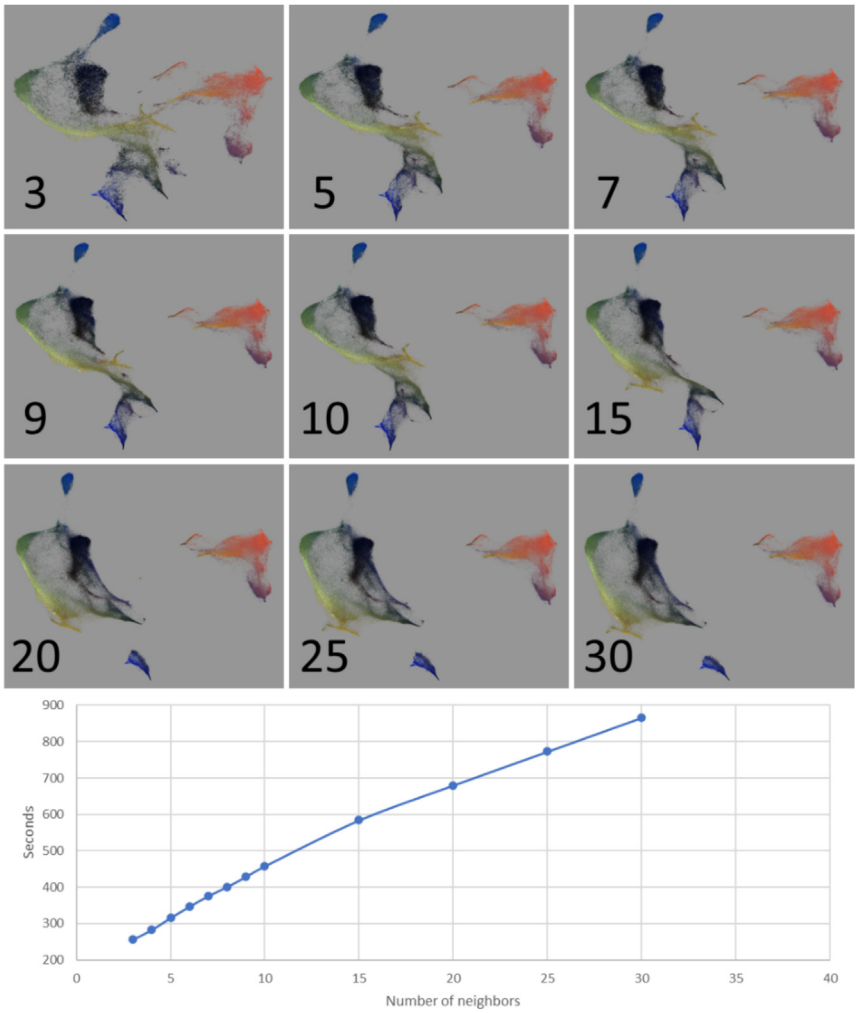


Fig. 7. UMAP color scatterplots obtained with using the cosine distance metric with number of neighbors varying from 3 to 30 and variation of their associated runtime (in seconds). Due to the large size of the historical dataset, only a subset of 15% of the total pixels were selected.

is easily explained by the fact that cosine measures the angle between vectors so differences in vector magnitude have no influence (Table S1). However, the resulting embeddings when working on normalized data with the Chebyshev and Euclidean metrics (Fig. 6b,c) appears simpler with few groupings than in the non-normalized dataset (Fig. 5b,d) Even so, the density maps associated with the normalized dataset do not appear to present denser clusters. For normalized Chebyshev, dense groupings drop down to ca. 160 pixels compared to ca. 220 pixels for the non-normalized data with most groupings being ca. 40 to 60 pixels. For Euclidean, we also observe a drop in the density of the denser groupings (ca. 220 vs. ca. 350 for non-normalized data) with a large portion of the groupings in the 100–125 pixels range. Therefore, data normalization does not appear to be helpful regarding density of the groupings when working with Euclidean or Chebyshev distance metric. Based on these considerations we conclude that the cosine distance metric produces the most consistent and densest reduction of hyperspectral data obtained on art materials.

3.5. Influence of the number of neighbors

In the literature [49,54,73], it is found that the higher the number of neighbors (n_neighbors) used for the UMAP algorithm, the more global structure is preserved, whereas a smaller number will force the algorithm to preserve more local structure (and perform similarly to the t-SNE algorithm). Also, the higher n_neighbors number will have a corresponding impact on processing time. Therefore, it is important to find an n_neighbors high enough to balance global vs. local structure but low enough to allow the data reduction to be realized in a reasonable amount of time and yield denser groupings. Here, we tested the influence of the number of neighbors from 3 to 30 on the Gauguin dataset (Fig. 7).

From the resulting embeddings, the number of neighbors used for the algorithm appears to have little influence on their 2-dimensional representation when using values larger than 5 or 7. Using such values, all colors (and therefore variables) are well grouped. Nonetheless, for low n_neighbors (<5), the clusters can be seen as not yet optimal, as points are found loosely scattered around the groupings. Higher n_neighbors (15 and above) have however a limited influence on the scatterplot, only creating slightly tighter groupings, while increasing considerably the processing time (Fig. 7). Therefore, while a n_neighbors of 5 would be suitable for most datasets both in term of embedding and running time (ca. 300 sec), given the limited processing time for n_neighbors of 10 (ca. 450 sec), it was decided that such a value would be a good compromise between global vs. local structure preservation and processing time for the more complex datasets obtained from paintings.

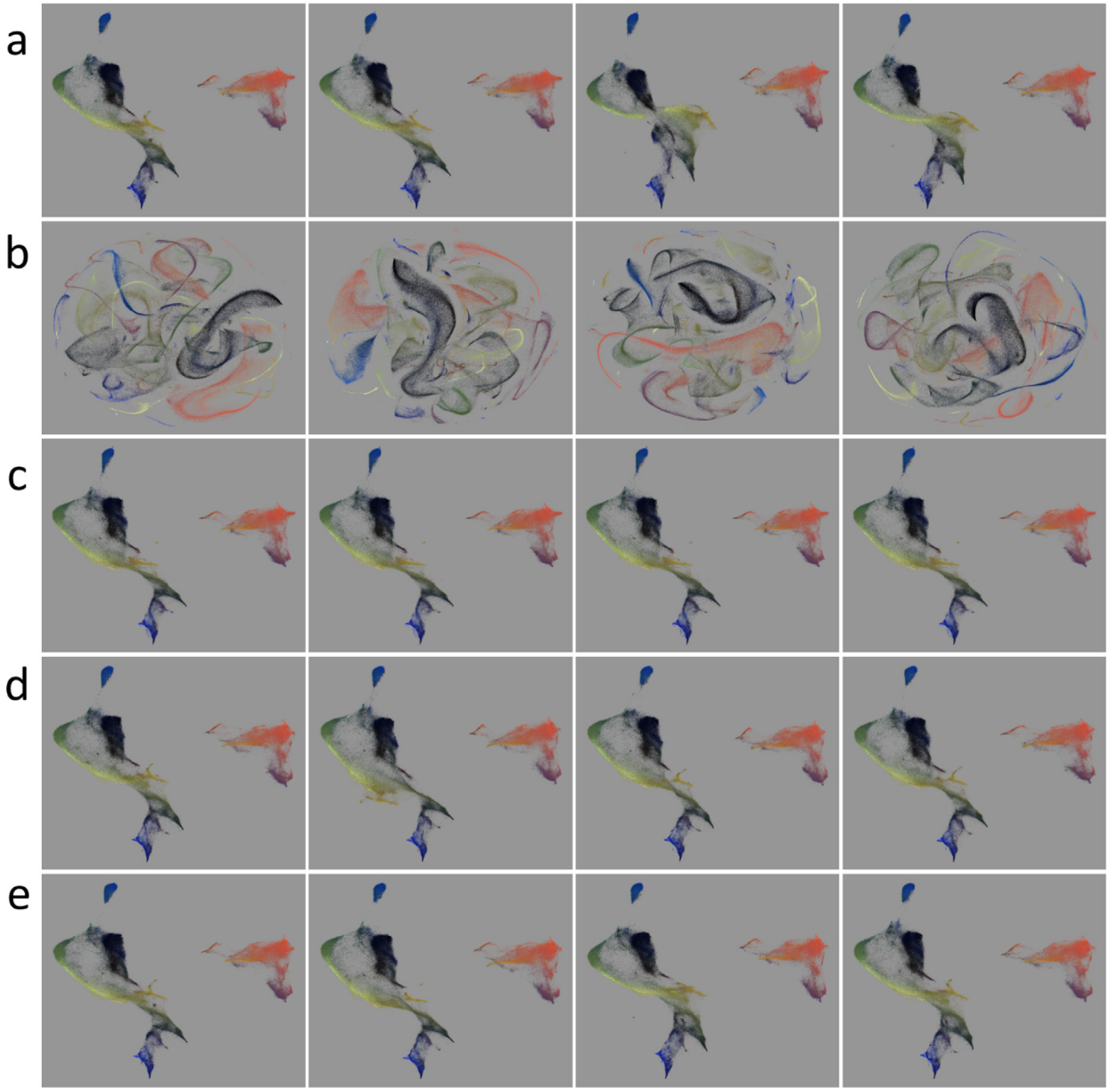


Fig. 8. 2D scatterplots for four successive UMAP or t-SNE runs using identical or different 15% sub-sampling of the datasets and selecting or not the UMAP random seed state in order to test the embedding reproducibility. Reproducibility is measured using the Structural Similarity Index Measure (SSIM). (a) UMAP, no random state, identical dataset, SSIM: 0.843 ± 0.010, (b) t-SNE, identical dataset, SSIM: 0.578 ± 0.004, (c) UMAP, random state, identical dataset, SSIM: 1, (d) UMAP, random state, different datasets, SSIM: 0.297 ± 0.005, (e) UMAP, no random state, different datasets, SSIM: 0.286 ± 0.014.

3.6. Reproducibility

Reproducibility of an experiment over time is very important. This can be required when reanalyzing an object past its original investigation. Therefore, reproducibility of the embedding obtained when using UMAP was investigated.

UMAP embeddings of a single dataset over several repetitions will present variations. This is due to the stochastic nature of the UMAP algorithm which makes use of randomness both to speed up approximation steps, and to aid in solving hard optimization problems [70,73]. Nonetheless, its strong mathematical foundations ensure a robust, interpretable and stable algorithm [48]. As a result, the variance between runs is relatively small but different runs still present variations (Fig. 8). The similarity between the various embeddings can be quantified using Structural Similarity Index Measure (SSIM), which value is equal to 1 for identical embeddings.

The SSIM over the four runs using the same random selection of 15% of the data cube pixels without using the UMAP random seed state (Fig. 8a) has been calculated at 0.843 ± 0.010 using a Python implementation of the algorithm. This reproducibility is much higher than the one observed for four successive runs of the same 15% sub sampling of pixels using t-SNE (Fig. 8b), for which the SSIM is of 0.578 ± 0.004. This highlights the poorly reproducible embeddings of the t-SNE method compared to UMAP, also described in literature [47,65,74]. To ensure that results can be reproduced exactly, the UMAP algorithm allows the user to set a random seed state (the variable referred to as "random_state" [68]). Fixing the random seed state to a given value (42 in the case of this paper), all runs of the same dataset under the same experimental conditions are found to be identical (Fig. 8c). Their structural similarity index over four successive runs of the same dataset was calculated to be 1, proving the exact reproduction of the embedding. However, it is important to keep in mind that,

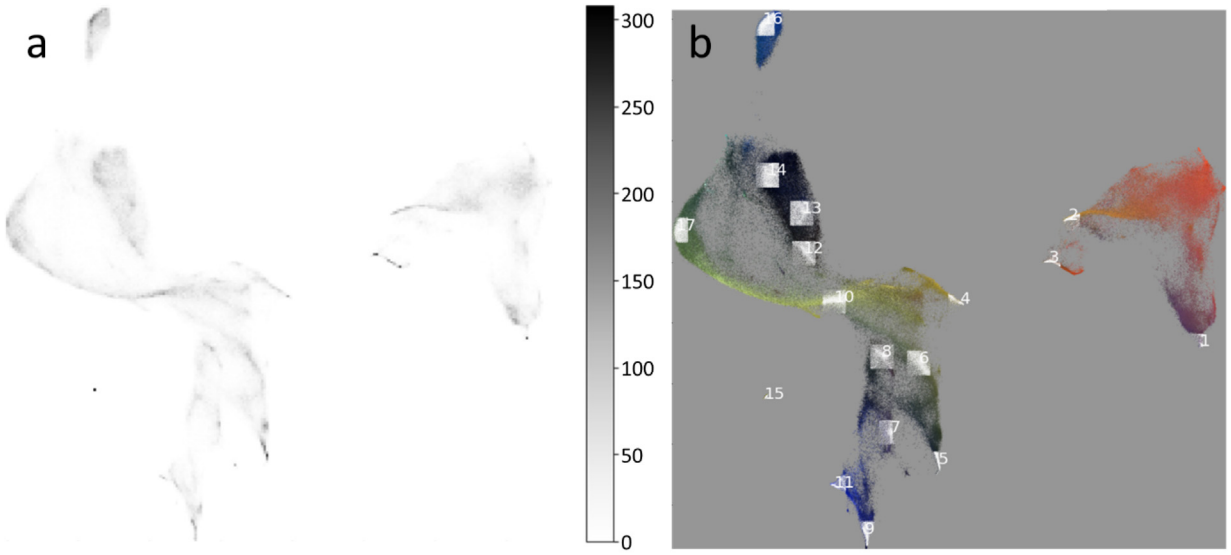


Fig. 9. 2D histogram(a) and corresponding density-based selected areas (b) for the Gauguin dataset using a threshold value of 0.3 and a minimum distance of 15.

Table 1
Number of identified endmembers upon variation of the threshold and minimum distance values used for the 2D histogram algorithm. 15% of the total available pixels of the Gauguin dataset were used to compute the embedding with a number of neighbors of 10 used for the UMAP calculation.

Minimum distance Threshold	5	10	15	20	25	30
0.1	72	33	20	16	12	11
0.15	63	32	20	16	12	11
0.2	57	31	20	16	12	11
0.25	48	30	19	15	12	11
0.3	38	25	17	14	12	11
0.35	29	21	16	13	12	11
0.4	20	14	14	12	11	10
0.45	16	12	11	9	8	8
0.5	14	10	10	8	7	7
0.55	13	9	9	7	7	7
0.6	10	7	7	5	5	5
0.65	7	6	6	5	5	5
0.7	6	5	5	4	4	4

when working with large datasets and subsampling randomly a certain percentage of pixels in order for the data reduction to run in an acceptable amount of time, the random seed state will not be useful as it is very unlikely that two randomly selected sub datasets will be identical and would therefore lead to embeddings presenting variations (Fig. 8d-e). This is confirmed by the structural similarity index of 0.297 ± 0.005 (fixed random seed state using four different sub sampling of the data, Fig. 8d), which is comparable to the one obtained for a random selection of pixels without the use of random seed state (0.286 ± 0.014, Fig. 8e). Despite a low structural similarity index, the 2D representations of their scatterplots do not present as much global variations as the ones observed for the t-SNE scatterplots (Fig. 8b). This is explained by the UMAP ability to retain the global structure of the original dataset, which is not the case with t-SNE.

Furthermore, the use of the random seed state in order to reach a perfect reproducibility of the result will increase the processing time for the embedding. In the case of the Gauguin dataset, using 15% of the total available pixels, the runtime without the use of the random seed state was 3 min 58 sec (±2 sec) versus 6 min 30 sec (±9 sec) when the random seed state was used. This is a not negligible increase in processing time (+60%), which may prove unnecessary if reproducibility is not required.

3.7. Integration of the reduction and visualization steps with 2D histogram density mapping and non-negative least square fitting

Data reduction and visualization are only the first steps in the exploration and interpretation of hyperspectral data obtained on works of art. The ultimate goal is to identify and visualize the distribution of pigments or mixture thereof. Therefore, we have integrated the data reduction using UMAP with density-based cluster identification to extract the corresponding endmembers, equivalent to pigments or the most commonly used mixtures of pigments employed by the artist. This is performed based on the density of the groupings obtained through a 2D histogram (Fig. 9a). The detection of the densest hotspots is dependent on the threshold and minimum distance. The threshold and minimum distance values to use to extract the endmembers may fluctuate from an embedding to another. This is left to the user's discretion based on the number of endmembers expected and the localization of the various hotspots on the colored scatterplot. However, it has been found that the smaller the values for the threshold and minimum distance, the larger the number of endmembers identified (Table 1). For a low threshold value (0.1–0.2) and a large minimum distance (25–30), the algorithm cannot detect an adequate number of hotspots and is underestimating the number of endmembers. On

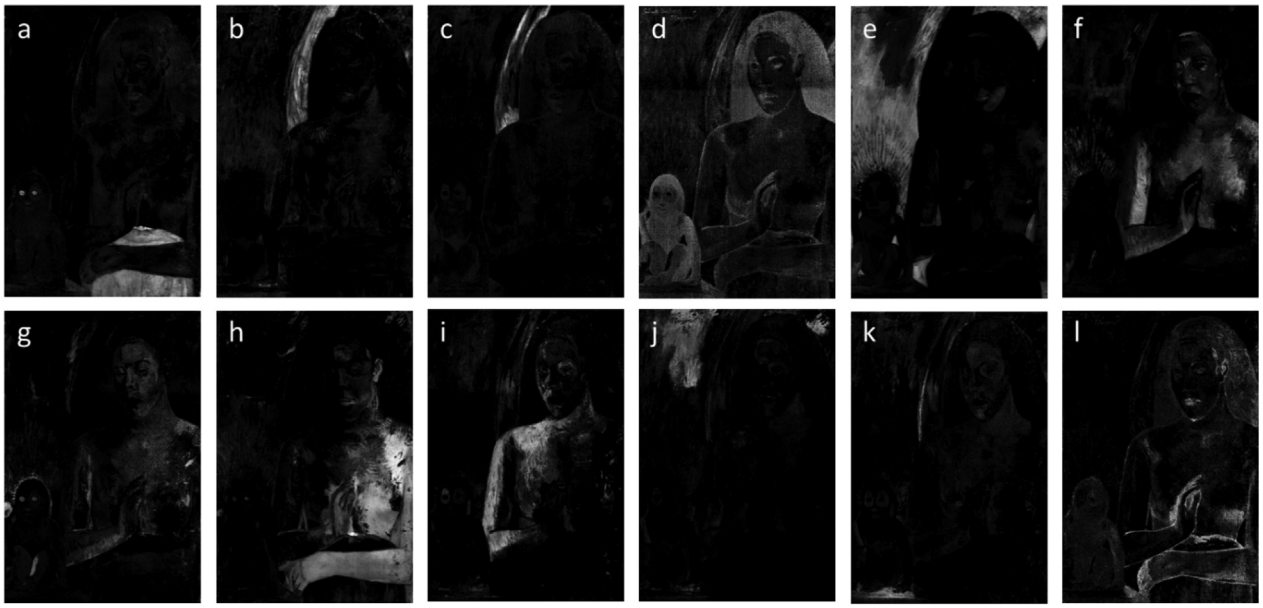


Fig. 10. Distribution maps for the various pigments and mixtures thereof using UMAP, 2D histogram cluster identification and non-linear least square fitting. (a) Prussian blue, (b) ultramarine blue, (c) cobalt blue, (d) unidentified black, (e) vermilion, (f) iron oxide, (g) cadmium yellow), (h) mixture of Prussian blue and unidentified yellow, (i) mixture of ultramarine blue and unidentified yellow, (j) and (k) mixture of ultramarine blue and vermilion to yield purple hues and, (l) Prussian blue. (a-k) were obtained using 2D histogram threshold values and minimum distance of 0.3 and 15 respectively whereas l was obtained using 0.3 and 10 as threshold and minimum distance values.

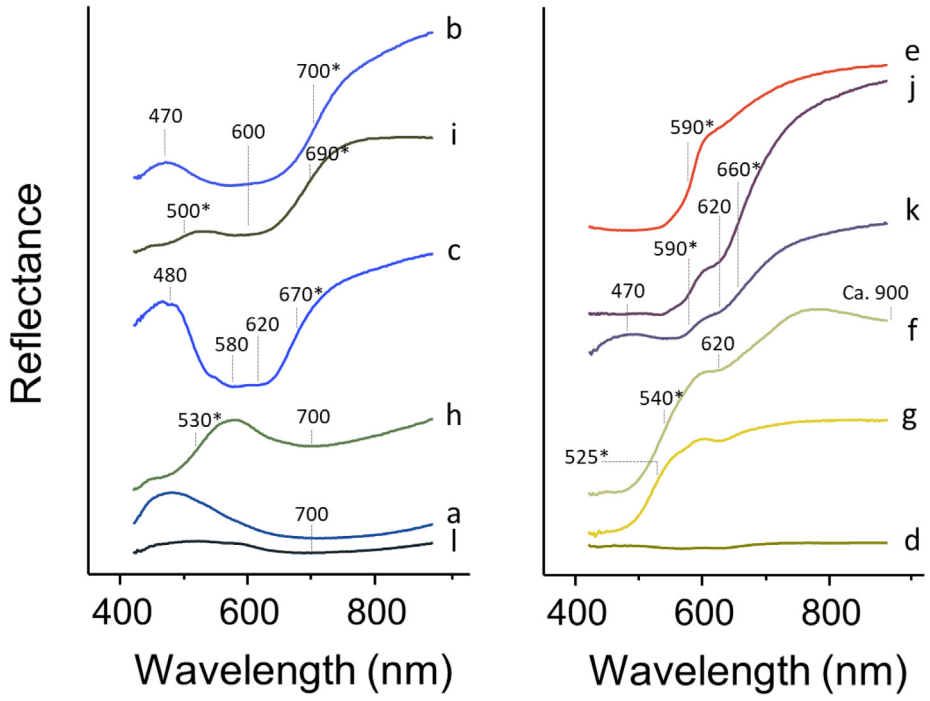


Fig. 11. Reflectance spectra associated with the distribution maps presented in Fig. 9. The values indicated correspond to the inflection points (marked with an asterisk) and maximum absorptions, characteristic for each pigment. Identification is done by comparing the spectra and these values to published databases [1,3,77,78].

the other hand, for a similar threshold and low minimum distance values (5–10) the algorithm is over evaluating the number of endmembers as 30–70 is much above the expected number of pigments or mixtures of pigments expected. The opposite observation can be done for high values of minimum distance and threshold where the algorithm is underestimating the number of endmembers. Therefore, it is important to choose values for threshold and minimum distance that will yield reasonable numbers of endmembers. Once threshold and minimum distance parameters are set, the algorithm scans the entire data cube with a given bounding box size (set to 0.4 here). This step will identify the hotspots in the scatterplot for which all points lie within the bounding box. An example of the resulting hotspots selection using a threshold value of 0.3 and a minimum distance of 15, yielding 17 endmembers, is given in Fig. 9b.

After the endmembers are identified based on the densest regions on the plot, weighted combinations of these endmembers are fitted to each pixel using non-negative least squares (nnls)

Table 2
Summary of the multispectral imaging, MA-XRF results and pigment tentative identification for Gauguin's *Poèmes Barbares* painting.

Color	Distribution map(s)	Spectral features		MA-XRF	Tentative pigment identification
		Maximum absorption (nm)	Inflexion points (nm)		
Blue 1	a	600–900	n/a	Pb, Sn	Prussian blue Lead white Carmine lake precipitated on tin substrate
Blue 2	b	600	700	n/a	Ultramarine blue
Blue 3	c	480, 580, 620	670	Co	Cobalt blue
Red	e	n/a	600	Hg	Vermilion
Yellow 1	f	620, 900	540	Fe, Mn	Iron oxide (umber-type)
Yellow 2	g	n/a	525	Cd	Cadmium yellow
Green 1	h	600–900	530	Fe	Prussian blue Iron oxide yellow
Green 2	i	600	500, 690	Fe	Ultramarine blueIron oxide yellow
Purple	j, k	620	590, 660	Hg	Ultramarine blueVermilion

algorithm [75,76] from the Python SciPy library. These weights, or concentrations, shows the contribution of each endmember to a given pixel and, when spatially addressed, form endmember distribution maps (Fig. 10). Using 0.5 threshold and minimum distances of 15 pixels allow the investigator to identify the main components of the composition (Fig. 10a-k). However, running the same embedding with smaller values for the minimum distance may create further sub hotspots and therefore highlight finer details such as the dark outlines of the arm, hands, ear, mouth and chin of the female figure (Fig. 10l), details that were not visualized in the previous run using higher values. This proves how crucial the selection of the values for the threshold and the minimum distance or the importance of running several instances of the same embedding with variations of these values.

By comparing the overall shape, inflection points and maximum absorbances the reflectance spectra associated with the distribution maps (Fig. 11) with published reflectance spectroscopy databases [1,3,77] and with the help of the elemental distribution maps obtained using MA-XRF (Fig. 4), identification of pigments used by Gauguin in his composition became possible. A summary of the results is found in Table 2.

The blue from the garment of the female figure did not yield any elements characteristic for blue pigments using MA-XRF (Fig. 4), which often indicate the use of ultramarine blue. This pigment was however ruled out based on visual observation and infrared photography (not shown), for which the pigment used proved to be too absorbent to be ultramarine blue. As a result, Prussian blue was assumed to be blue pigment used in that area despite the lack XRF response for iron. HSI allowed to confirm the use of Prussian blue in in garment of the female figure (Fig. 10a), characterized by its important absorbance in the 600–900 nm range (Fig. 11a). MA-XRF also highlighted lead (Pb, Fig. 4-Pb) and tin (Sn, Fig. 4-Sn) in this area of the painting. Tin oxide has been identified as a common substrate for red carmine lake in 19th century paintings [79–81]. Therefore, Sn identified with MA-XRF is most likely associated with the substrate of a carmine lake, also identified through micro analyses (data not shown), while Pb most likely indicates the use of lead white mixed with the pigments to create a lighter shade. Prussian blue was also identified in the eyes of Ta'aroa (the Tahitian deity who is the creator of the universe), highlights of the wing, and shadows of the lips, hair, and hands of the female figure. The presence of iron in these areas went undetected in MA-XRF as suggested by the iron (Fe) elemental map (Fig. 4-Fe). This may be due to the dynamic range and variations of concentration between the Prussian blue in the blue garment and the other iron-based pigments found in the body of the female figure. Prussian blue having a high tinting strength, only small quantities of the pigment are needed to produce the deep blue colors often found in artworks. Such low concentration may therefore go undetected in XRF analysis [82]. The blue found in the wing (Fig. 10b) was identified as ultramarine blue, characterized by the lack of elements in the MA-XRF and its characteristic reflectance curve with reflectance at 470 nm, maximum absorbance around 600 nm and inflection point around 700 nm [3,78]. Bright shades of blue are observed in the highlights of the wing (Fig. 10c). Based on the maximum absorbance at 480, 580, 620 nm and the inflection point at 670 nm (Fig. 11c), this shade of blue may have been realized with cobalt blue ($CoO \cdot Al_2O_3$). This is supported by the presence of cobalt (Co) in these areas using MA-XRF (Fig. 4–Co). The cobalt blue characteristic band at ca. 480 nm is very dim in the reflectance spectrum but, it has been shown that in dark shades of cobalt blue, this band can go undetected [77]. As suggested by the MA-XRF, cobalt blue is also found throughout the dark areas of the deity's body and female figure's hair (Fig. 4–Co). This is suggested by the distribution map presented in Fig. 10d. However, with these areas being very dark, it becomes challenging to make identification based on the reflectance spectra, which will appear flat, close to 0 and for which characteristic features will be dimmed (Fig. 11d). While the presence of tin (Sn) in these dark areas (Fig. 4-Sn) may also suggests the use of cerulean blue ($CoO \cdot n\ SnO_2$), the use of such pigment is very unlikely and micro-invasive analyses (not shown) revealed that the Sn in the dark areas of the female figure's hair and deity's body was associated with the tin oxide substrate of the carmine organic lake. In these dark areas, Gauguin does not seem to employ a pure black pigment but rather employs an optical black effect achieved by mixing cobalt blue, carmine lake and emerald green, as suggested by the Co, Sn, copper (Cu) and arsenic (As) MA-XRF maps (Fig. 4–Co, Fig. 4-Sn, Fig. 4-As and Fig. 4-Cu).

The mercury (Hg) map obtained using MA-XRF (Fig. 4-Hg) along with the steep inflection point at 600 nm (Fig. 11e) strongly suggest the use of vermilion in the orange/red areas of the picture (Fig. 10e).

Identification of yellows is always challenging using HSI alone as most present a similar sigmoid curve with inflection point around 550–600 nm. Iron oxide, however, can easily be differentiated from the other yellow pigments due to its characteristic reflectance curve (Fig. 11f) [3]. This allowed this identification of iron oxide pigments in the yellow areas of the proper right arm and body of the female figure (Fig. 10f). The co-localization of the iron (Fe) and manganese (Mn) maps obtained in MA-XRF (Fig. 4-Fe and 4-Mn) further suggests the use of a umber-type iron oxide pigment (Fe_2O_3 ($\cdot H_2O$) + $MnO_2 \cdot (n\ H_2O)$ + Al_2O_3). Along with iron oxide, cadmium yellow is also identified in the yellow areas of the arm and body of the female figure, in the radiation and inner thigh of the deity, and in the fruit in the deities hand (Fig. 10g). The identification is based on the 525 nm inflection point (Fig. 11g) along with the cadmium (Cd) MA-XRF distribution map (Fig. 4-Cd). The location of iron oxide and cadmium yellow in the

yellow area of the arm suggests that both pigments were used in mixture. This show the difficulty associated with the identification of pigments using non-invasive imaging techniques.

The majority of greens appear to have been obtained using different mixtures of yellows and blues. The arm and the proper left side of the body (Fig. 10h) can be identified as a mixture of yellow and Prussian blue. The yellow used can be potentially identified as cadmium yellow due to the 530 nm inflection point whereas Prussian blue is characterized by its large absorption in the 600–900 nm range with maximum absorbance around 700 nm (Fig. 11h). Nonetheless, the use of cadmium yellow is not fully supported by the MA-XRF data. Instead, the Fe map (Fig. 4-Fe) suggests the use of an iron oxide pigment mixed with the Prussian blue. Another green can be observed in the female figure's right-hand side shoulder, right-hand side arm and part of the face (Fig. 10i). Based on the maximum absorbance around 600 nm and two inflection points around 500 and 690 nm (Fig. 11i), the mixture is hypothesized as being yellow and ultramarine blue. Here again, the Fe MA-XRF map (Fig. 4-Fe) suggests the use of an iron oxide pigment rather than cadmium or chrome yellow. The co-presence of Cu and As in the greenish tones of the female figure's skin (Fig. 4-Cu and Fig. 4-As) suggests the use of emerald green. However, this pigment could not be identified through HSI.

Finally, two shades of purples were identified through the UMAP and 2D histogram processes. They are present in the red background (Fig. 10j) and in the table and purple highlight of the wing (Fig. 10k). Both present similar reflectance spectra with inflection point at 590 and 660 nm along with maximum absorption at 620 nm (Fig. 11j-k). Furthermore, the purple found in the table where the deity stands presents a reflection band at 470 nm, often associated with ultramarine blue (Fig. 11k). Therefore, it is very likely that the purple hues have been obtained by mixing ultramarine blue (470 nm, inflection point around 660 nm) and vermilion (inflection point at 590 nm), the latter supported by the Hg XRF map (Fig. 4-Hg). The 470 nm feature observed for spectrum k would be due to the deeper blue hue of the purple associated with this composition.

4. Conclusion

With this article, we showed that Uniform Manifold Approximation and Projection (UMAP) is a solid and reliable alternative to current data reduction techniques used in the field of cultural heritage for hyperspectral data obtained in the visible range. It yields superior runtimes compared with t-SNE and the embeddings produced present less but much tighter clusters than t-SNE projections, characteristic of a better preservation of the global vs. local structure balance. Such balanced embeddings are easier to interpret and will allow a better understanding of the artists' creative processes.

In addition, we evaluated various parameters such as the distance metric, the number of neighbors used for the data reduction, the influence of the normalization of the data. We were able to conclude that the cosine distance metric was the most appropriate in terms of data reduction, visualization, creation of tight clusters and was not influenced by the normalization of the data cube. While higher number of neighbors are said to help maintain the global structure of the original data, we concluded that, in the cases of artistic material hyperspectral data in the visible range, higher number of neighbors have a limited influence but increase drastically the processing time. As a result, we found that a number of neighbors of 10 was a good compromise for global and local structure conservation and acceptable processing time.

When UMAP is associated with density cluttering recognition and non-negative least square fitting of the data, the notebook provided and presented in this article allows to extract, identify, and localize pigments or mixture of, also called end embers, answering one of the main goals of hyperspectral imaging analysis of works of art.

While this approach is not completely novel, it is the first time that data reduction, endmembers identification and extraction as well as visualization can be done through a single free user interface. With the growing interest and development of UMAP, we hope this research will pave the way for future research in this area.

CRediT authorship contribution statement

Marc Vermeulen: Methodology, Software, Formal analysis, Investigation, Data curation, Writing - original draft. **Kate Smith:** Conceptualization, Writing - review & editing. **Katherine Eremin:** Conceptualization, Writing - review & editing. **Georgina Rayner:** Conceptualization, Writing - review & editing. **Marc Walton:** Conceptualization, Methodology, Software, Writing - review & editing, Supervision, Funding acquisition.

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 paper.

Acknowledgments

This collaborative initiative is part of NU-ACCESS's broad portfolio of activities, made possible by generous support of the Andrew W. Mellon Foundation as well as supplemental support provided by the Materials Research Center, the Office of the Vice President for Research, the McCormick School of Engineering and Applied Science and the Department of Materials Science and Engineering at Northwestern University. The authors gratefully acknowledge Emeline Pouyet and Gianluca Pastorelli (formerly NU-ACCESS) for the acquisition of the MA-XRF and hyperspectral data on Gauguin's *Poèmes Barbares*. Finally, the authors thank Giovanni Verri (Art Institute of Chicago) for his feedback on the Jupyter Notebook containing these scripts.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.saa.2021.119547.

References

[1] M. Vermeulen, E.M.K. Müller, M. Leona, Non-Invasive Study of the Evolution of Pigments and Colourants Use in 19th Century Ukiyo-e, Arts of Asia 50 (2020).
[2] D. Tamburini, J. Dyer, Fibre optic reflectance spectroscopy and multispectral imaging for the non-invasive investigation of Asian colourants in Chinese textiles from Dunhuang (7th-10th century AD), Dyes Pigm. 162 (2019) 494–511.
[3] M. Aceto, A. Agostino, G. Fenoglio, A. Idone, M. Gulmini, M. Picollo, P. Ricciardi, J.K. Delaney, Characterisation of colourants on illuminated manuscripts by portable fibre optic UV-visible-NIR reflectance spectrophotometry, Anal. Meth. 6 (2014).
[4] C. Biron, A. Mounier, J.P. Arantegui, G.L. Bourdon, L. Servant, R. Chapoulie, C. Roldán, D. Almazán, N. Díez-de-Pinos, F. Daniel, Colours of the « images of the floating world ». non-invasive analyses of Japanese ukiyo-e woodblock prints (18th and 19th centuries) and new contributions to the insight of oriental materials, Microchem. J. 152 (2020).
[5] E. Casanova-Gonzalez, M.A. Maynez-Rojas, A. Mitrani, I. Rangel-Chavez, M.A. Garcia-Bucio, J.L. Ruvalcaba-Sil, K. Munoz-Alcocer, An imaging and spectroscopic methodology for in situ analysis of ceiling and wall decorations in Colonial missions in Northern Mexico from XVII to XVIII centuries, Heritage Sci. 8 (2020) 14.

[6] D. Tamburini, C.R. Cartwright, J. Adams, The scientific study of the materials used to create the Tahitian mourner's costume in the British Museum collection, J. Cult. Heritage 42 (2020) 263–269.
[7] M. Bacci, Fiber optics applications to works-of-art, Sens. Actuator B-Chem. 29 (1995) 190–196.
[8] M. Bacci, R. Bellucci, C. Cucci, C. Frosinini, M. Picollo, S. Porcinai, B. Radicati, Fiber optics reflectance spectroscopy in the entire VIS-IR range: A powerful tool for the non-invasive characterization of paintings, in: P.B. Vandiver, J.L. Mass, A. Murray (Eds.), Materials Issues in Art and Archaeology VII, Materials Research Society, Warrendale, 2005, pp. 297–302.
[9] M. Bacci, A. Casini, C. Cucci, M. Picollo, B. Radicati, M. Vervat, Non-invasive spectroscopic measurements on the Il ritratto della figliastra by Giovanni Fattori: identification of pigments and colourimetric analysis, J. Cult. Heritage 4 (2003) 329–336.
[10] M. Bacci, D. Magrini, M. Picollo, M. Vervat, A study of the blue colors used by Telemaco Signorini (1835–1901), J. Cult. Heritage 10 (2009) 275–280.
[11] B. Rosenzweig, E. Carretti, M. Picollo, P. Baglioni, L. Dei, Use of mid-infrared fiber-optic reflectance spectroscopy (FORS) to evaluate efficacy of nanostructured systems in wall painting conservation, Appl. Phys. A-Mater. Sci. Process. 83 (2006) 669–673.
[12] G. Dupuis, M. Menu, Quantitative characterisation of pigment mixtures used in art by fibre-optics diffuse-reflectance spectroscopy, Appl. Phys. A-Mater. Sci. Process. 83 (2006) 469–474.
[13] L. Appolonia, D. Vaudan, V. Chatel, M. Aceto, P. Mirti, Combined use of FORS, XRF and Raman spectroscopy in the study of mural paintings in the Aosta Valley (Italy), Anal. Bioanal. Chem. 395 (2009) 2005–2013.
[14] M. Leona, J. Winter, Fiber Optics Reflectance Spectroscopy: A Unique Tool for the Investigation of Japanese Paintings, Stud. Conserv. 46 (2001) 153–162.
[15] M. Bacci, F. Baldini, R. Carla, R. Linari, A COLOR ANALYSIS OF THE BRANCACCI CHAPEL FRESCOES, Appl. Spectrosc. 45 (1991) 26–31.
[16] M. Vermeulen, L. Burgio, N. Vandeperre, E. Driscoll, M. Viljoen, J. Woo, M. Leona, Beyond the connoisseurship approach: creating a chronology in Hokusai prints using non-invasive techniques and multivariate data analysis, Heritage Sci. 8 (2020) 62.
[17] K. Keune, J. Mass, F. Meirer, C. Pottasch, A. van Loon, A. Hull, J. Church, E. Pouyet, M. Cotte, A. Mehta, Tracking the transformation and transport of arsenic sulfide pigments in paints: synchrotron-based X-ray micro-analyses, J. Anal. At. Spectrom. 30 (2015) 813–827.
[18] G. Marchioro, C. Daffara, PCA-based method for managing and analyzing single-spot analysis referenced to spectral imaging for artworks diagnostics, MethodsX 7 (2020).
[19] S.V.J. Berbers, D. Tamburini, M.R. van Bommel, J. Dyer, Historical formulations of lake pigments and dyes derived from lac: A study of compositional variability, Dyes Pigm. 170 (2019).
[20] G. Capobianco, M.P. Bracciale, D. Sali, F. Sbardella, P. Belloni, G. Bonifazi, S. Serranti, M.L. Santarelli, M.C. Guidi, Chemometrics approach to FT-IR hyperspectral imaging analysis of degradation products in artwork cross-section, Microchem. J. 132 (2017) 69–76.
[21] J.A. Lee, M. Verleysen, Nonlinear Dimensionality Reduction, Springer Publishing Company, Incorporated, 2007.
[22] M. Albrecht, O. de Noord, S. Meloni, A. van Loon, R. Haswell, Jan Steen's ground layers analysed with Principal Component Analysis, Heritage Science 7 (2019) 53.
[23] P. Nabais, M.J. Melo, J.A. Lopes, T. Vitorino, A. Neves, R. Castro, Microspectrofluorimetry and chemometrics for the identification of medieval lake pigments, Heritage Science 6 (2018) 13.
[24] G. Sciutto, P. Oliveri, S. Prati, M. Quaranta, S. Lanteri, R. Mazzeo, Analysis of paint cross-sections: a combined multivariate approach for the interpretation of mu ATR-FTIR hyperspectral data arrays, Anal. Bioanal. Chem. 405 (2013) 625–633.
[25] N. Navas, J. Romero-Pastor, E. Manzano, C. Cardell, Raman spectroscopic discrimination of pigments and tempera paint model samples by principal component analysis on first-derivative spectra, J. Raman Spectrosc. 41 (2010) 1486–1493.
[26] G. Musumarra, M. Fichera, Chemometrics and cultural heritage, Chemometrics Intell. Lab. Syst. 44 (1998) 363–372.
[27] M.K. Donais, M. Alrais, K. Konomi, D. George, W.H. Ramundt, E. Smith, Energy dispersive X-ray fluorescence spectrometry characterization of wall mortars with principal component analysis: Phasing and ex situ versus in situ sampling, J. Cult. Heritage 43 (2020) 90–97.
[28] I. Jolliffe, Principal Component Analysis, in: M. Lovric (Ed.), International Encyclopedia of Statistical Science, Springer, Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 1094–1096.
[29] L. Cséfalvayová, M. Strlič, H. Karjalainen, Quantitative NIR Chemical Imaging in Heritage Science, Anal. Chem. 83 (2011) 5101–5106.
[30] C. Cucci, J.K. Delaney, M. Picollo, Reflectance Hyperspectral Imaging for Investigation of Works of Art: Old Master Paintings and Illuminated Manuscripts, Accounts Chem. Res. 49 (2016) 2070–2079.
[31] A. Orlando, M. Picollo, B. Radicati, S. Baronti, A. Casini, Principal Component Analysis of Near-Infrared and Visible Spectra: An Application to a XIIth Century Italian Work of Art, Appl. Spectrosc. 49 (1995) 459–465.
[32] N. Rohani, E. Pouyet, M. Walton, O. Cossairt, A.K. Katsaggelos, Nonlinear Unmixing of Hyperspectral Datasets for the Study of Painted Works of Art, Angew. Chem.-Int. Edit. 57 (2018) 10910–10914.
[33] P. Kubelka, New contributions to the optics of intensely light-scattering material, part ii: Non-homogenous layers, J. Optical Soc. 44 (1954) 330.
[34] P. Kubelka, New contributions to the optics of intensely light-scattering material, part i, J. Optical Society 38 (1948) 448.
[35] P. Kubelka, F. Munk, Ein Beitrag Zur Optik Der Farbanstriche, Zeitschrift für Technische Physik 12 (1931) 593–601.
[36] A. Diaz-Papkovich, L. Anderson-Trocme, C. Ben-Eghan, S. Gravel, UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts, PLoS Genet. 15 (2019) 24.
[37] V.C. Rodrigues, J.C. Soares, A.C. Soares, D.C. Braz, M.E. Melendez, L.C. Ribas, L.F. S. Scabini, O.M. Bruno, A.L. Carvalho, R.M. Reis, R.C. Sanfelice, O.N. Oliveira, Electrochemical and optical detection and machine learning applied to images of genosensors for diagnosis of prostate cancer with the biomarker PCA3, Talanta 222 (2021).
[38] J.A. Carter, L.M. O'Brien, T. Harville, B.T. Jones, G.L. Donati, Machine learning tools to estimate the severity of matrix effects and predict analyte recovery in inductively coupled plasma optical emission spectrometry, Talanta 223 (2021).
[39] B.M. Devassy, S. George, P. Nussbaum, Unsupervised Clustering of Hyperspectral Paper Data Using t-SNE, Journal of Imaging 6 (2020).
[40] E. Pouyet, N. Rohani, K. Katsaggelos Aggelos, O. Cossairt, M. Walton, Innovative data reduction and visualization strategy for hyperspectral imaging datasets using t-SNE approach, Pure Appl. Chem. (2018) 493.
[41] M. Alfeld, S. Pedetti, P. Martinez, P. Walter, Joint data treatment for Vis–NIR reflectance imaging spectroscopy and XRF imaging acquired in the Theban Necropolis in Egypt by data fusion and t-SNE, C.R. Phys. 19 (2018) 625–635.
[42] L. van der Maarten, G. Hinton, Visualizing Data using t-SNE, J. Machine Learning Res. 9 (2008) 2579–2605.
[43] A. Gisbrecht, A. Schulz, B. Hammer, Parametric nonlinear dimensionality reduction using kernel t-SNE, Neurocomputing 147 (2015) 71–82.
[44] J. Wu, J. Wang, H. Xiao, J. Ling, Visualization of High Dimensional Turbulence Simulation Data using t-SNE, in: 19th AIAA Non-Deterministic Approaches Conference.
[45] G.C. Linderman, M. Rachh, J.G. Hoskins, S. Steinerberger, Y. Kluger, Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data, Nat. Methods 16 (2019) 243–245.
[46] J. Tang, J. Liu, M. Zhang, Q. Mei, Visualizing Large-scale and High-dimensional Data, in: Proceedings of the 25th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, Montréal, Québec, Canada, 2016, pp. 287–297.
[47] E. Becht, L. McInnes, J. Healy, C.A. Dutertre, I.W.H. Kwok, L.G. Ng, F. Ginhoux, E. W. Newell, Dimensionality reduction for visualizing single-cell data using UMAP, Nat. Biotechnol. 37 (2019) 38.
[48] L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, in: arXiv e-prints, 2018, pp. arXiv:1802.03426.
[49] A. Coenen, A. Pearce, Understanding UMAP, in: Google PAIR.
[50] E. Becht, C.-A. Dutertre, I.W.H. Kwok, L.G. Ng, F. Ginhoux, E.W. Newell, Evaluation of UMAP as an alternative to t-SNE for single-cell data, bioRxiv (2018).
[51] A. Mazher, Visualization Framework for High-Dimensional Spatio-Temporal Hydrological Gridded Datasets using Machine-Learning Techniques, Water 12 (2020) 15.
[52] M. Picollo, C. Cucci, A. Casini, L. Stefani, Hyper-Spectral Imaging Technique in the Cultural Heritage Field: New Possible Scenarios, Sensors 20 (2020) 2843.
[53] L. Wander, A. Vianello, J. Vollertsen, F. Westad, U. Braun, A. Paul, Exploratory analysis of hyperspectral FTIR data obtained from environmental microplastics samples, Anal. Methods 12 (2020) 781–791.
[54] G. Franch, G. Jurman, L. Coviello, M. Pendesini, C. Furlanello, MASS-UMAP: Fast and Accurate Analog Ensemble Search in Weather Radar Archives, Remote Sens. 11 (2019) 25.
[55] F. Pont, M. Tosolini, J.J. Fournie, Single-Cell Signature Explorer for comprehensive visualization of single cell signatures across scRNA-seq datasets, Nucleic Acids Res. 47 (2019) 9.
[56] T. Smets, N. Verbeeck, M. Claesen, A. Asperger, G. Griffioen, T. Tousseyn, W. Waelput, E. Waelkens, B. De Moor, Evaluation of Distance Metrics and Spatial Autocorrelation in Uniform Manifold Approximation and Projection Applied to Mass Spectrometry Imaging Data, Anal. Chem. 91 (2019) 5706–5714.
[57] H. Yanagisawa, T. Yamashita, H. Watanabe, Manga Character Clustering with DBSCAN using Fine-Tuned CNN Model, in: Q. Kemao, K. Hayase, P.Y. Lau, W.N. Lie, Y.L. Lee, S. Srisuk, L. Yu (Eds.), International Workshop on Advanced Image Technology, Spie-Int Soc Optical Engineering, Bellingham, 2019.
[58] S. Sakaue, J. Hirata, M. Kanai, K. Suzuki, M. Akiyama, C.L. Too, T. Arayssi, M. Hammoudeh, S. Al Emadi, B.K. Masri, H. Halabi, H. Badsha, I.W. Uthman, R. Saxena, L. Padyukov, M. Hirata, K. Matsuda, Y. Murakami, Y. Kamatani, Y. Okada, Dimensionality reduction reveals fine-scale structure in the Japanese population with consequences for polygenic risk prediction, Nat. Commun., 11 (2020) 11.
[59] Y. Jiale, Z. Ying, Visualization method of sound effect retrieval based on UMAP, in: 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chongqing, China, 2020, pp. 2216-2220.
[60] D. Wu, J.Y. Poh Sheng, G.T. Su-En, M. Chevrier, J.L. Jie Hua, T.L. Kiat Hon, J. Chen, Comparison Between UMAP and t-SNE for Multiplex-Immunofluorescence Derived Single-Cell Data from Tissue Sections, bioRxiv (2019).
[61] 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, A. Cardona, Fiji: an open-source platform for biological-image analysis, Nat. Methods 9 (2012) 676–682.
[62] S. Preibisch, S. Saalfeld, P. Tomancak, Globally optimal stitching of tiled 3D microscopic image acquisitions, Bioinformatics 25 (2009) 1463–1465.
[63] B. McCune, J.B. Grace, D.L. Urban, Analysis of Ecological Communities, MjM Software Design, Gleneden Beach, OR, 2002.
[64] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine Learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[65] M. Sanchez-Rico, J.M. Alvarado, A Machine Learning Approach for Studying the Comorbidities of Complex Diagnoses, Behav. Sci. 9 (2019) 14.
[66] L. Oakley, S. Zaleski, B. Males, O. Cossairt, M. Walton, Improved spectral imaging microscopy for cultural heritage through oblique illumination, Heritage Science 8 (2020) 27.
[67] R.S. Berns, Digital color reconstructions of cultural heritage using color-managed imaging and small-aperture spectrophotometry, Color Res. Appl. 44 (2019) 531–546.
[68] L. McInnes, UMAP API Guide, in, 2018, pp. https://umap-learn.readthedocs.io/en/latest/api.html#umap-api-guide.
[69] M. Vermeulen, M. Leona, Evidence of early amorphous arsenic sulfide production and use in Edo period Japanese woodblock prints by Hokusai and Kunisada, Heritage Science 7 (2019).
[70] D. Probst, J.L. Reymond, Visualization of very large high-dimensional data sets as minimum spanning trees, J. Cheminformatics 12 (2020) 13.
[71] [71] E. Amid, M.K. Warmuth, TriMap: Large-scale dimensionality reduction using triplets, arXiv preprint arXiv:1910.00204, (2019).
[72] N. Rohani, E. Pouyet, M. Walton, O. Cossairt, A.K. Katsaggelos, Pigment Unmixing of Hyperspectral Images of Paintings Using Deep Neural Networks, Int. Conf. Acoust. Spee (2019) 3217–3221.
[73] [73] L. McInnes, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, in, 2018.
[74] D.Y. Orlova, L.A. Herzenberg, G. Walther, Science not art: statistically sound methods for identifying subsets in multi-dimensional flow and mass cytometry data sets, Nat. Rev. Immunol. 18 (2018) 77.
[75] W. Feng, R. Shi, C. Zhang, T. Yu, D. Zhu, Lookup-table-based inverse model for mapping oxygen concentration of cutaneous microvessels using hyperspectral imaging, Opt. Express 25 (2017) 3481–3495.
[76] B. Rankin, J. Meola, D. Perry, J. Kaufman, Methods and challenges for target detection and material identification for longwave infrared hyperspectral imagery, SPIE, 2016.
[77] IFAC/CNR, Fiber Optics Reflectance Spectra (FORS) of Pictorial Materials in the 270-1700 nm range, in, 2020.
[78] M. Aceto, A. Agostino, G. Fenoglio, M. Picollo, Non-invasive differentiation between natural and synthetic ultramarine blue pigments by means of 250–900 nm FORS analysis, Anal. Methods 5 (2013).
[79] A. Burnstock, I. Lanfear, K.J. Berg, L. Carlyle, M. Clarke, E. Hendriks, J. Kirby, Comparison of the fading and surface deterioration of red lake pigments in six paintings by Vincent van Gogh with artificially aged paint reconstructions, in, 2005.
[80] I. Schaefer, K. Lewerentz, C.v. Saint-George, Painting light: the hidden techniques of the Impressionists, Skira ; Distributed in North America by Rizzoli International Publications, Milano, Italy : New York, NY, 2008.
[81] M. van bommel, M. Geldof, E. Hendriks, An Investigation of Organic Red Pigments used in Paintings by Vincent Van Gogh (November 1885 to February 1888), in: ArtMatters : Netherlands technical studies in art., Waanders, Zwolle, 2005, pp. 111-137.
[82] L. Glinsman, The application of X-ray fluorescence spectrometry to the study of museum objects, in, University of Amsterdam, Amsterdam, 2004.