analytical chemistry

Article

Cite This: Anal. Chem. 2019, 91, 5706−5714

pubs.acs.org/ac

Evaluation of Distance Metrics and Spatial Autocorrelation in Uniform Manifold Approximation and Projection Applied to Mass Spectrometry Imaging Data

Tina Smets,[*,†] Nico Verbeeck,[†,‡] Marc Claesen,[†,‡] Arndt Asperger,[¶] Gerard Griffioen,[§] Thomas Tousseyn,[∥] Wim Waelput,[⊥] Etienne Waelkens,[□] and Bart De Moor[†]

[†]STADIUS Center for Dynamical Systems, Signal Processing, and Data Analytics, Department of Electrical Engineering (ESAT), KU Leuven, 3001 Leuven, Belgium

[‡]Aspect Analytics NV, C-mine 12, 3600 Genk, Belgium

[¶]Bruker Daltonik GmbH, Fahrenheitstrasse 4, 28359 Bremen, Germany

[§]reMYND, Bio-Incubator, Gaston Geenslaan 1, 3000 Leuven, Belgium

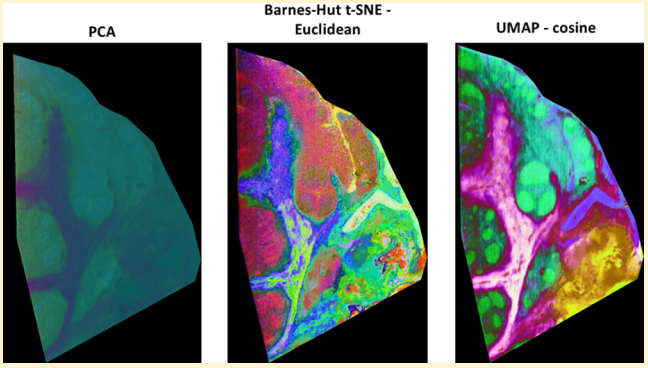
[∥]Department of Pathology, University Hospitals KU Leuven, 3001 Leuven, Belgium

[⊥]Department of Pathology, UZ-Brussel, 1000 Brussels, Belgium

[□]Department of Cellular and Molecular Medicine, KU Leuven, 3000 Leuven, Belgium

Supporting Information

ABSTRACT: In this work, uniform manifold approximation and projection (UMAP) is applied for nonlinear dimensionality reduction and visualization of mass spectrometry imaging (MSI) data. We evaluate the performance of the UMAP algorithm on MSI data sets acquired in mouse pancreas and human lymphoma samples and compare it to those of principal component analysis (PCA), *t*-distributed stochastic neighbor embedding (t-SNE), and the Barnes−Hut (BH) approximation of t-SNE. Furthermore, we compare different distance metrics in (BH) t-SNE and UMAP and propose the use of spatial autocorrelation as a means of comparing the resulting low-dimensional embeddings. The results indicate that UMAP is competitive with t-SNE in terms of visualization and is well-suited for the dimensionality reduction of large (>100 000 pixels) MSI data sets. With an almost fourfold decrease in runtime, it is more scalable in comparison with the current state-of-the-art: t-SNE or the Barnes−Hut approximation of t-SNE. In what seems to be the first application of UMAP to MSI data, we assess the value of applying alternative distance metrics, such as the correlation, cosine, and the Chebyshev metric, in contrast to the traditionally used Euclidean distance metric. Furthermore, we propose "histomatch" as an additional custom distance metric for the analysis of MSI data.



Mass spectrometry imaging (MSI) is a molecular-imaging technology that enables the direct study of the spatial distribution of biomolecular species in a tissue section.[1,2] MSI provides a very rich biochemical characterization of a sample; however, a single MSI experiment can lead to gigabytes of complex data. Furthermore, the number of pixels collected in an experiment, as well as the number of m/z bins measured, is ever-increasing with improving technology.

Because of the sheer volume and complexity of MSI data, there is a growing need for scalable dimensionality reduction techniques to extract the underlying trends from these data, in order to (i) facilitate human interpretation; (ii) reduce data size and complexity for additional data-analysis steps, such as clustering; and (iii) allow for visualization of these data. Dimensionality reduction techniques such as principal component analysis (PCA), non-negative matrix factorization (NMF), and probabilistic latent semantic analysis (pLSA)[3−7] have been successfully used to this end, with PCA probably being the most widely used technique. PCA relies on the determination of orthogonal eigenvectors along which the largest variances in the data are to be found. PCA works very well to approximate data by a low-dimensional subspace, which is equivalent to the existence of many linear relations among the projected data points.[8] In NMF, a data matrix, $\mathbf{X}$, is factored into two matrices, $\mathbf{W}$ and $\mathbf{H}$, such that by iteratively minimizing the residual squared distance between the products

Received: December 18, 2018
Accepted: April 15, 2019
Published: April 15, 2019

DOI: 10.1021/acs.analchem.8b05827
Anal. Chem. 2019, 91, 5706−5714

of these two factorized matrices, **W** and **H**, the resulting product resembles the original data matrix, **X**, as well as possible (**X** ≈ **WH**).[9] pLSA, on the other hand, relies on a statistical mixture model and aims to decompose the original data into the underlying latent variables via the iterative expectation-maximization (EM) procedure, resulting in a set of probability distributions for these discovered latent variables.[10]

A common issue in these methods is the selection of the number of components to be retained for dimensionality reduction. Each of the resulting components contains a part of the total information in the data, and the inclusion of an insufficient number of components will result in a loss of information and data features. This is especially relevant when trying to visualize the data in 2 or 3 dimensions. Another limitation specific to PCA is related to the fact that the interpretation is hampered by the possible presence of negative peaks in the pseudospectra, particularly in the context of MSI data, where negative values are not expected in the measurements. This disadvantage is alleviated in NMF and pLSA through the enforcement of non-negative constraints.

Although linear dimensionality reduction techniques, of which PCA, NMF, and pLSA are prominent examples, have shown satisfactory results in the past and have been instrumental in extracting information from MSI data, these models assume that there exist linear relationships among the variables. This is, however, a very strong assumption that most often cannot be imposed as many biological models are inherently nonlinear.

These linear methods thus face a limitation with regard to nonlinear pattern recognition, resulting in an incomplete capture of the underlying structure. As a result, nonlinear dimensionality reduction (NLDR) techniques have gained increasing popularity for the analysis of biological data, including MSI, with *t*-distributed stochastic neighbor embedding (t-SNE) being the leading example.[11,12]

t-SNE has been shown to be very valuable for MSI data analysis, not only for dimensionality reduction in itself but also for visualization purposes.[11,12] This is due to the fact that t-SNE is strong at maintaining the local distances among data points, and the number of features that can be captured in the reduced space is not restricted by the number of dimensions selected. Because t-SNE is able to embed all features into two or three dimensions, even if more than two or three features are present, the visualization of the data is greatly facilitated. This is an important advantage of t-SNE and similar nonlinear methods that is absent from methods such as PCA, NMF, pLSA, and clustering in general.

A major drawback of most NLDR methods, on the other hand, is their high computational cost. In t-SNE, the pairwise distance matrix needs to be calculated between pixels, which makes the method computationally hard and not suitable for large (>100 000 pixels) MSI data sets. The Barnes−Hut approach therefore seeks to approximate these calculations by using a nearest-neighbors approach but is nevertheless confronted with long runtimes for larger data sets.[13] To mitigate these computational constraints, Thomas et al.[14] have presented the application of autoencoders for unsupervised NLDR. This approach treats every pixel as a training example, such that the complete data set is not loaded in memory at once, and as long as the number of hidden neurons is smaller than the number of channels in the mass spectrum, this method will be computationally less expensive than t-SNE. Although greatly improving on standard t-SNE, the computation time of the weight matrix is still very large (20 h for a data set containing 409 × 404 pixels and 7036 m/z bins or 8 Gb in size).[14]

In this paper, we show the utility of uniform manifold approximation (UMAP) for the analysis of MSI data.[15] The UMAP algorithm seeks to find an embedding by searching for a low-dimensional projection of the data that has the closest possible equivalent fuzzy topological structure (see also the UMAP outline in Algorithm 1). The method can therefore be used in a similar fashion as t-SNE (i.e., for visualization purposes) but also for general NLDR. In order to make a comparison to the current state-of-the-art used for nonlinear dimensionality reduction and visualization, t-SNE, and linear dimensionality reduction using standard PCA, we propose the use of spatial autocorrelation. Spatial autocorrelation has previously been studied in the context of MSI by Cassese et al.[16] and refers to the observation that neighboring pixels in an MSI experiment generally more closely resemble each other than pixels that are far away from each other and are thus more likely to have a higher correlation between them. This is due to the fact that neighboring pixels are often located in the same tissue type and have a higher chance of having a similar function and thus chemical content.

One of the first steps in the application of NLDR algorithms is the calculation of distances or similarities between the pixels in the original high-dimensional space. This distance metric aims to describe how similar or dissimilar the spectra of each pixel and thus its chemical content are to those of the other pixels. In MSI research and many other fields, the Euclidean distance is very often used as the de facto distance metric; however, the choice of this distance measure can have a considerable impact on the results, as meaningful patterns might be lost due to the high-dimensionality of the data, and noise or outliers can become amplified.[17] In this work, we therefore compare the effects of different distance metrics, namely, the traditional Euclidean distance, the cosine distance, the Chebyshev distance, and the Pearson correlation coefficient between pixels. Furthermore, we highlight "histomatch" as an additional custom distance metric for the analysis of MSI data.[18,19]

To the best of our knowledge, this is the first time that spatial autocorrelation has been used for the relative comparison of the embeddings obtained through different dimensionality reduction methods, using a variety of distance metrics instead of relying on the widely applied Euclidean distance metric.

We demonstrate our method on MSI data collected from mouse pancreas and human lymph-node tissue (Table 1), comparing the resulting embeddings through spatial autocorrelation. Moreover, we evaluate the application of different distance metrics, including the proposed histomatch metric.

Table 1. Overview of Dimensionality Associated with the Used Datasets

sample	pixels	m/z bins
Pancreas M1	10 606	14 000
Pancreas M2	14 791	14 000
Lymphoma P1	572 173	8000
Lymphoma P2	552 701	8000

DOI: 10.1021/acs.analchem.8b05827
Anal. Chem. 2019, 91, 5706−5714

EXPERIMENTAL SECTION

Mass Spectrometry Imaging (MSI)-Data Acquisition and Processing. MSI was performed on mouse pancreatic tissue and on human-lymph-node samples. For all samples, MSI was done on a Bruker rapifleX MALDI-TOF mass spectrometer. For the mouse pancreatic tissue, cryosections of 7 μm thickness were prepared and mounted on ITO glass slides. Sinapinic acid (SA) was used as matrix and applied using a Bruker ImagePrep. The pixel size was set to 50 μm, and the recorded m/z range was 2−20 kDa in positive linear mode. The acquisition speed was 9 pixels/s with 1000 lasershots/pixel and a laser repetition rate of 10 kHz.

For the human-lymph-node samples, cryosections of 5 μm thickness were prepared and mounted on ITO glass slides. 2,5-Dihydroxybenzoic acid (2,5-DHB) was used as the matrix and applied using sublimation. The pixel size was set to 10 μm, and the recorded m/z range was 620−1200 Da in positive reflector mode. The acquisition was performed with 200 lasershots/pixel and a laser repetition rate of 10 kHz, resulting in an acquisition speed of 32 pixels/s.

Data Processing of Mouse-Pancreas and Human-Lymphoma Data Sets. *Data Modeling.* All data was normalized using total-ion count (TIC). Manifold learning approaches were used to embed high-dimensional data in a low-dimensional space for data visualization and investigation of nonlinear relations in the data; in addition, PCA was used for comparative reasons. The following three methods were used to map the data to three dimensions: (1) PCA, (2) t-SNE and the Barnes−Hut (BH) approximation thereof, and (3) UMAP. The t-SNE mapping to three dimensions was done using the default settings, as discussed by van der Maaten et al.,[12] apart from the different distance metrics that were evaluated. The UMAP mapping to three dimensions was performed using the Python implementation as provided by the paper's author, L. McInnes, according to the default settings (n_neighbors = 15, gamma = 1.0, n_epochs = None, alpha = 1.0, init = 'spectral', spread = 1.0, min_dist = 0.1, a = None, b = None, random_state = None, metric_kwds = , verbose = True; see https://github.com/lmcinnes/umap), except for the different distance metrics, which were evaluated. The distance metrics used here are Euclidean, correlation, cosine, and Chebyshev, for which the detailed formulas are shown in Figure S1. For UMAP, one additional distance metric was evaluated: histomatch. The cosine and histomatch distance metrics rely on the assumption that MSI data contain no negative values. Because of the high computational burden associated with the larger lymphoma data sets, the comparison of distance metrics was mainly carried out using the UMAP algorithm for this data. For the analyses where a t-SNE implementation was used on these large data sets an initial dimensionality-reduction step (n = 100) with PCA was required to make these analyses feasible. A more extensive comparison with regard to the distance metrics in t-SNE implementations has therefore been done using the smaller pancreas samples.

To run standard t-SNE, we relied on the Python code provided by L. van der Maaten (https://lvdmaaten.github.io/software/), and for the Barnes−Hut approximation of t-SNE, we relied on the Scikit-learn implementation, which uses the Barnes−Hut approximation by default. This implementation, as provided in Scikit-learn, was used for all analyses, unless it is explicitly stated that t-SNE was used (i.e., when a comparison is made for the standard t-SNE implementation, the Barnes−Hut approximation, and UMAP). All experiments were run on an Intel Xeon CPU E5-2660 v2 2.20 GHz machine with 10 cores and 128 Gb RAM.

UMAP Outline. From a general outline, UMAP uses local manifold approximations and assembles together their local fuzzy-simplicial-set representations to form a topological representation of the high-dimensional data. Given some low-dimensional representation of the data, a similar process can be used to form an equivalent topological representation. The layout of the data representation in the low-dimensional space is then optimized through the minimization of the cross-entropy between the two topological representations.[15] The general outline of the algorithm, as specified by McInnes et al., goes as follows:

Algorithm 1 Outline UMAP algorithm

```
function UMAP(X, n, d, min-dist, n-epochs)
    for all x ∈ X do
        fs-set[x] ← LocalFuzzySimplicialSet(X, x, n)
    top-rep ← ∪_{x∈X} fs-set[x]        ▷ We recommend the probabilistic t-conorm
    Y ← SpectralEmbedding(top-rep, d)
    Y ← OptimizeEmbedding(top-rep, Y, min-dist, n-epochs)
    return Y
```

For the detailed functions related to the construction of the local fuzzy simplicial sets, the determination of the spectral embedding, and the optimization of the embedding with regard to fuzzy-set cross-entropy, we refer to the original article by McInnes et al. from which this general outline has been adopted.[15]

An important component of the algorithm is the cost function for the optimization of the embedding through minimization of the fuzzy-set cross-entropy:

$$C_{\mathrm{UMAP}} = \sum_{i \neq j} v_{ij} \log\left(\frac{v_{ij}}{w_{ij}}\right) + (1 - v_{ij}) \log\left(\frac{1 - v_{ij}}{1 - w_{ij}}\right) \tag{1}$$

There is some resemblance to the Kullback−Leibler divergence (t-SNE cost function, eq 2) in the first part of the equation; however, it is important to note here that UMAP does not use the same definitions for v_{ij} and w_{ij}, wherein i and j refer to two objects in the high-dimensional (v_{ij}) and low-dimensional (w_{ij}) space.

$$C_{\text{t-SNE}} = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{2}$$

where v_{ij} refers to the local fuzzy-simplicial-set memberships defined in the high-dimensional space on the basis of the smooth nearest-neighbors distances, whereas w_{ij} refers to the low-dimensional similarities between i and j. The t-SNE cost function (eq 2), on the other hand, seeks to minimize the Kullback−Leibler divergence between the joint probability distribution in the high-dimensional space, p_{ij}, and the joint probability distribution in the low-dimensional space, q_{ij}. The fact that both p_{ij} and q_{ij} require calculations over all pairs of points imposes a high computational burden on t-SNE. Therefore, improvements in the efficiency of methods, such as the Barnes−Hut approximation and UMAP, focus on approximating these quantities. UMAP achieves an efficient approximate k-nearest-neighbor computation via the nearest-neighbor-descent algorithm[20] for which an empirical complexity of $O(p^{1.14})$ was reported, in comparison with the t-SNE of $O(p^2)$ and the BH t-SNE of $O(p \log p)$, wherein p refers to the number of pixels. Although the aforementioned time complex-

DOI: 10.1021/acs.analchem.8b05827
Anal. Chem. 2019, 91, 5706−5714

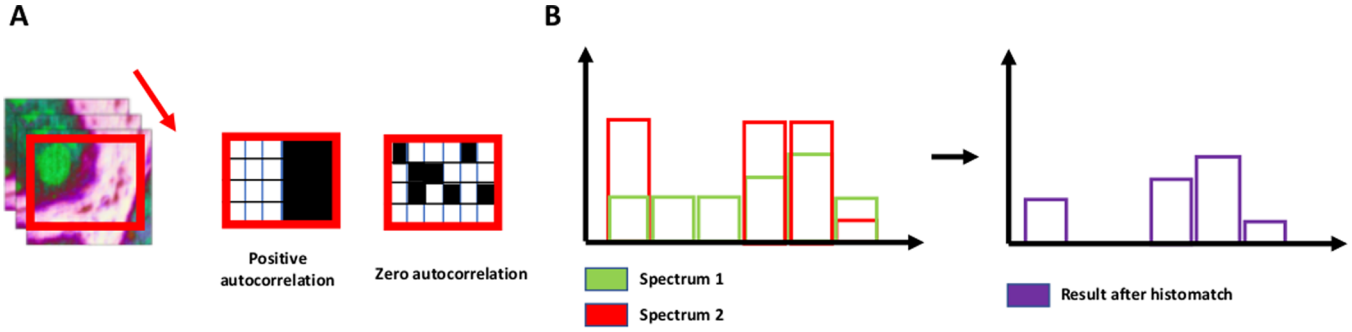


Figure 1. (A) Spatial autocorrelation. The resulting 3D embeddings, wherein each dimension corresponds to a color channel, obtained via dimensionality reduction are shifted diagonally according to a number of pixels in order to spatially correlate each pixel given its neighborhood. An example is given of a perfect positive autocorrelation consistent with a clear binary pattern. In the absence of such a pattern, as shown in the other example, zero autocorrelation is observed. Given the nature of our embeddings, the correlation is calculated according to each color channel using the RGB index, as shown by the given formula. (B) Process of histogram matching. The high-dimensional data are modified such that their histogram matches the histogram of the reference data.

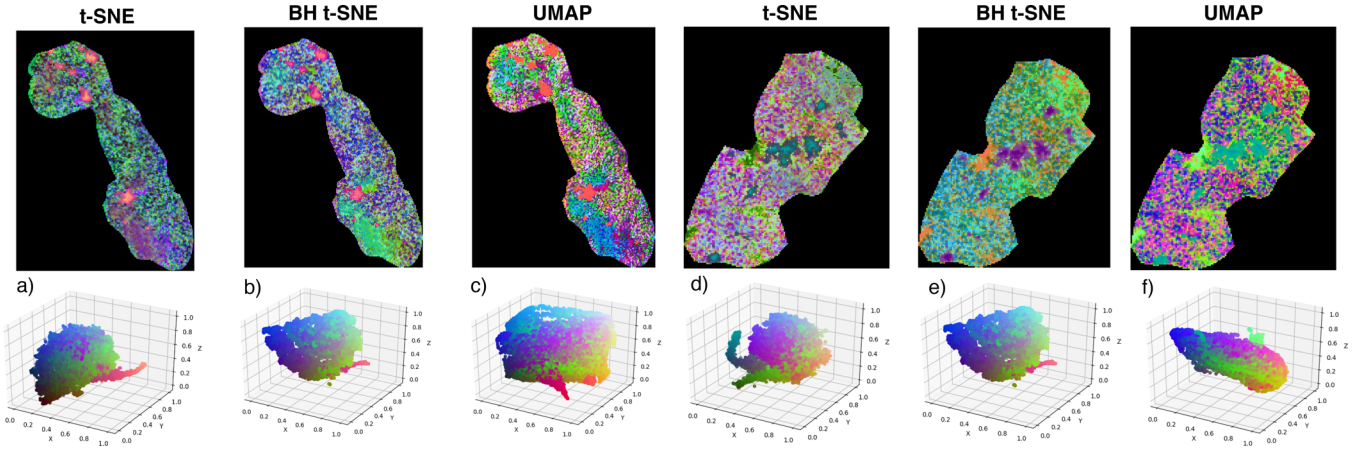


Figure 2. Comparison of t-SNE (a,d), Barnes−Hut (b,e), and UMAP (c,f) embeddings of two different pancreas-tissue samples, M1 (a−c) and M2 (d−f). Shown on top are the 3D embeddings of the tissues using hyperspectral visualization, with the corresponding scatter plots underneath. The Euclidean distance metric was used for all embeddings.

ities indicate better asymptotic scaling for BH t-SNE in comparison with UMAP, it is important to remind the reader that these complexities do not reflect the different constant time multipliers between BH t-SNE and UMAP. We thus want to emphasize that for any contemporary data set of common size (i.e., <5 M pixels), UMAP exhibits significantly shorter run times than BH t-SNE using the implementations we benchmarked. For a detailed outline of the complete mathematical foundations of the algorithm, we would like to refer the interested reader to the original manuscript and in particular to Appendix C for a detailed comparison between UMAP and t-SNE.[12,15]

Autocorrelation. Given the spatial nature of MSI data, a certain degree of spatial correlation is expected to occur. This means that neighboring pixels are more likely to be correlated to each other than pixels located at a further distance.[16] As shown in Figure 1, we have evaluated the obtained embeddings by subjecting them to a spatial-autocorrelation function. By diagonally shifting each embedding over a number of pixels (ranges of 1−15 and 1−50 pixels), we can determine the correlation of pixels to their neighbors according to incremental distance. Technically, the autocorrelation function is used to compute the dot product of the original image (hyperspectral visualization of tissue embedding in 3D), with the image shifted for increasing numbers of shifts in pixels. For each shift from the original image, the RGB autocorrelation index was calculated according to the following formula:

$$\mathrm{RGB}_{\mathrm{index}} = \sqrt{(r_{\mathrm{R}})^2 + (r_{\mathrm{G}})^2 + (r_{\mathrm{B}})^2} \tag{3}$$

The correlation vector, r, represents the correlation between the shifted and original image for the red (R), green (G), and blue (B) color components. Autocorrelation was performed using Python, and all results were normalized using min−max normalization.

Custom Distance Metric. The histomatch distance metric is inspired by the histogram matching algorithm, which forces the intensity distribution of an image to match the intensity distribution of a target.[18,19] The histomatch distance metric is specified according to the following formula:

$$d(x, y) = 1 - \sum_i \max(0, \min(x_i, y_i)) \tag{4}$$

where x and y are spectra from different pixels. A schematic illustration is given in Figure 1B.

Data Visualization. For all manifold learning approaches, the locations of pixels were translated to RGB color coding by varying the red, green, and blue intensities linearly on the three independent axes, such that the minimum value on an axis is represented by a color intensity of 0, and the maximum value

DOI: 10.1021/acs.analchem.8b05827
Anal. Chem. 2019, 91, 5706−5714

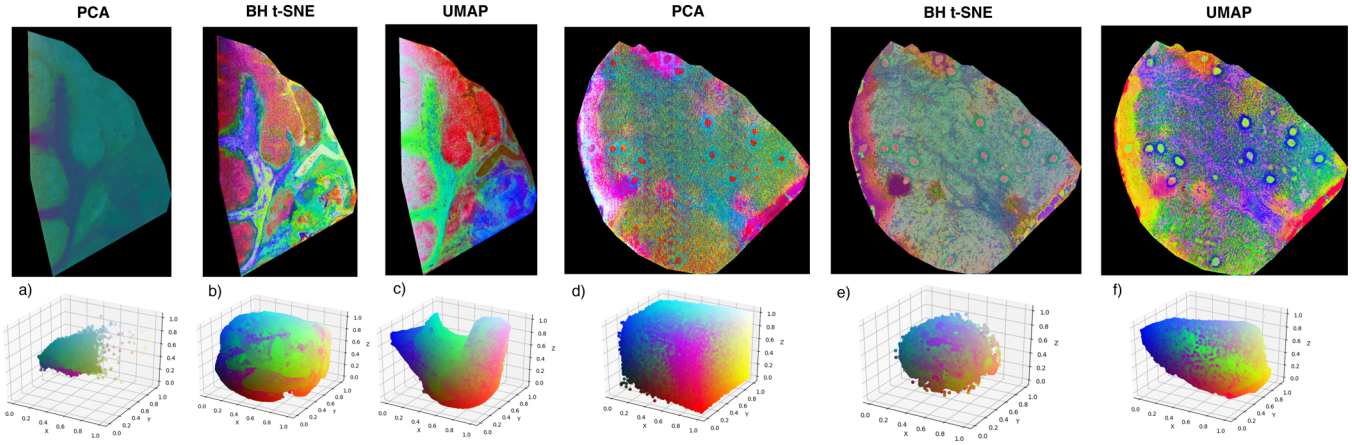


Figure 3. Comparison of PCA (a,d), Barnes–Hut (b,e), and UMAP (c,f) embeddings of two lymphoma-tissue samples, P1 (a–c) and P2 (d–f). Shown on top are the 3D embeddings of the tissues using hyperspectral visualization, with the corresponding scatter plots underneath. The Euclidean distance metric was used for all embeddings.

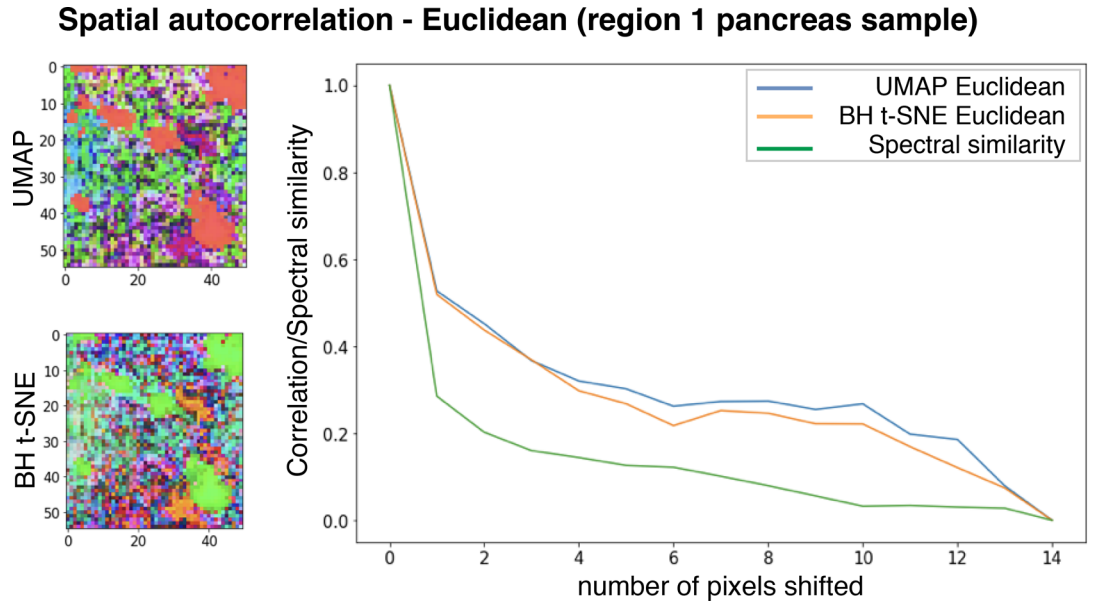


Figure 4. Spatial autocorrelation of the M2 pancreas embeddings shown as the correlation in the function of increasing distance or incremental number of pixels shifted. As expected, the correlation values are maximal for a zero shift and decrease as a function of the distance. Shown are the results for one region in the tissue when the Euclidean distance metric was used in the BH t-SNE (orange) and UMAP (blue) algorithms. Each graph also includes the spectral similarity between pixels according to the Euclidean distance metric (green), which serves as a benchmark indicator for the spatial autocorrelation. The same approach is shown in Figure S3a for a larger region. The graphs show clearly that in every case the spatial autocorrelation behaves similarly for both BH t-SNE and UMAP, which supports the fact that UMAP delivers embeddings of at least the same quality as BH t-SNE.

on an axis has an intensity of 255, which can be normalized to a scale of 0 to 1; this visualization is referred to as hyperspectral visualization.[11]

Spectral-Similarity or Chemical-Distance Information. To visualize the relationship between the spatial autocorrelation of the embedding and the chemical data, we followed a similar approach as above. This was done by calculating the mean distance between the MSI spectral or intensity data (according to the relevant distance metric used) rather than correlating the pixels of the resulting embeddings. Here, we assume that for pixels that are located closer to each other, the spectral similarity or chemical relatedness is more likely to be larger for these pixels than for the ones located further from each other. We use this assumption to support the relative comparison of the embeddings obtained on the same tissue section but via different algorithms (e.g., t-SNE versus UMAP) and in particular as a benchmark for the introduced autocorrelation measure. From here on, we will refer to this measure as spectral similarity.

RESULTS AND DISCUSSION

Enabling the Analysis of Large MSI Data Sets with UMAP. We compare the performance of the UMAP algorithm for the analysis of pancreas and lymph-node samples with the performances of t-SNE, the Barnes–Hut (BH) approximation, and PCA (Figures 2 and 3). The embeddings obtained via the different manifold learning approaches are visualized using the RGB color scheme such that the colors in the image depend on each pixel's location in the model space. Hence, similar colors

DOI: 10.1021/acs.analchem.8b05827
Anal. Chem. 2019, 91, 5706–5714

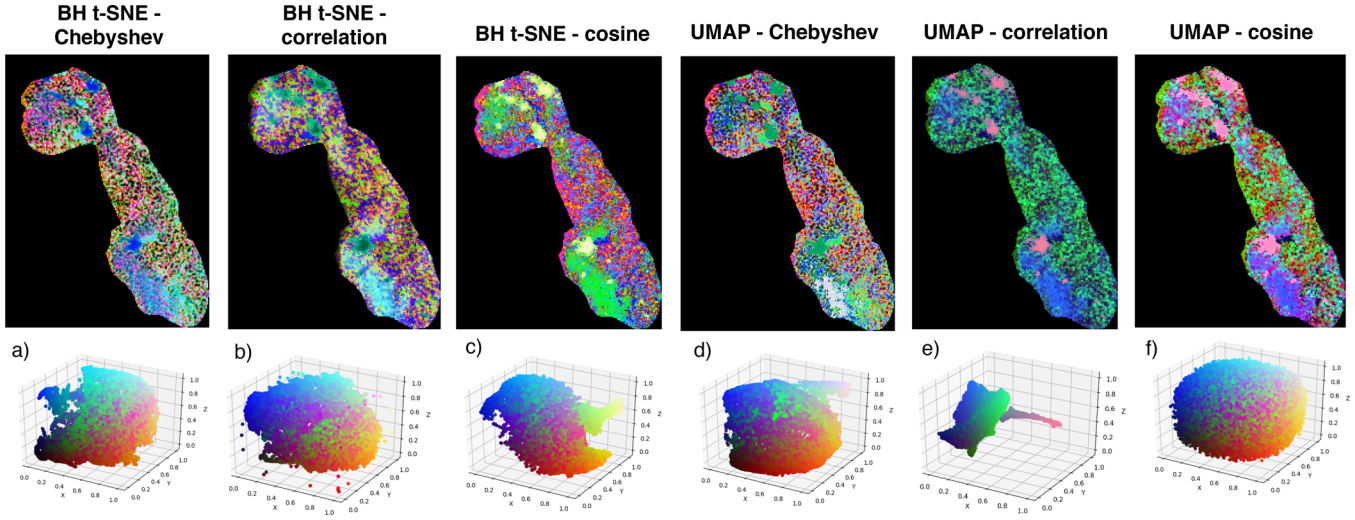


Figure 5. Comparison of the Chebyshev, correlation, and cosine distance metrics using BH t-SNE (a–c) and UMAP (d–f) for the M2 pancreas samples. Shown on top are the 3D embeddings of the tissues using hyperspectral visualization, with the corresponding scatter plots underneath.

or RGB values represent similar spectra or biochemical patterns.

UMAP strongly outperforms t-SNE as well as its faster counterpart, BH t-SNE, with almost a fourfold decrease in runtime. A comparison of the runtimes obtained by the different algorithms is given in Figure S2. Besides being computationally more efficient, UMAP is also capable of delivering embeddings with at least the same quality as t-SNE, as can be seen in Figures 2 and 3.

Figure 2 shows the hyperspectral visualization of the two analyzed pancreas samples for the 3D embeddings obtained using t-SNE, BH t-SNE, and UMAP. For this comparative analysis, all implementations were used with standard parameters. The original t-SNE implementation is not intended to be used with alternative distance metrics, which is why we used the Euclidean distance metric to make a cross-comparison possible. In Figures 3 and 4, an overview is given of the analyzed lymphoma samples. For the lymphoma data sets used in this paper (552 701 and 572 173 pixels × 8000 m/z bins, respectively), BH t-SNE required more than 2 weeks of computational time when another dimensionality reduction technique was not used in advance. In an ideal world, using PCA (or similar dimensionality reduction algorithms) in advance should not be necessary, because these methods are strong at preserving the global distance structure within the data, whereas methods such as t-SNE focus on the preservation of the local distance. In our case, the composition of biological tissue is very heterogeneous; hence, we are interested in preserving the local distances as well as possible. Ideally, however, we get the best of both worlds, and the global structure is preserved as well. The UMAP algorithm is said to be competitive with t-SNE in terms of visualization and to preserve more of the global structure in the data, while being more scalable toward large data sets.[15] We have indeed observed that for our larger data sets, UMAP is able to reduce 8000 dimensions for over 500 000 data points to 2 or 3 dimensions in approximately 6 h. In terms of memory usage for UMAP, the median measured across five runs was 1.7 Gb for the pancreas-tissue sample and 25 Gb for the lymphoma-tissue sample. This is comparable to the observed median (n = 5) memory usage for BH t-SNE (1.4 Gb) for the pancreas samples. Because of the high computational burden associated with BH t-SNE for the large lymphoma samples, an initial dimensionality reduction to 100 dimensions using PCA was needed. Therefore a fair comparison with UMAP in this regard was not possible.

Overall, UMAP shows good performance, and in comparison to PCA using NLDR techniques, as shown in Figure 3, has the major advantage of compressing the spatial and molecular information into three dimensions, resulting in detailed visualization.

Toward a Relative Comparison of the Embeddings Obtained by UMAP and t-SNE via Spatial Autocorrelation. Because it is difficult to compare the embeddings obtained through UMAP versus the other methods solely by visual inspection, we propose the use of spatial autocorrelation. This approach reflects the correlation between the values of a single variable strictly due to the proximity of these values in a geographical space by introducing a deviation from the assumption of independent observations of classical statistics.[21] For MSI data, it is commonly assumed that close neighboring pixels are likely to be more correlated to each other than to their more distant neighbors.[16] Our experiments empirically corroborate this assumption, because the correlation between a pixel and its neighboring pixels is more likely to decrease with increasing distance. Likewise, we assume that for pixels that are located closer to each other, the spectral similarity or chemical relatedness is more likely to be larger for these pixels than for ones located further from each other. We therefore use this assumption as a benchmark indicator for the autocorrelation measured.

In Figure 4, a comparison is given for the spatial autocorrelation of embeddings obtained on the basis of the pancreas M2 sample using UMAP and BH t-SNE. Shown is the graph for a selected region using the Euclidean distance metric. As expected, the correlation values are maximal for a zero shift and decrease as a function of the distance. In addition to the spatial autocorrelation, the spectral similarity is also shown as a benchmark indicator. The fact that the spatial autocorrelation shows similar behaviors for both BH t-SNE and UMAP further supports the fact that UMAP delivers embeddings of at least the same quality as BH t-SNE. This is also reflected in the additional results regarding the spatial autocorrelation as shown in Figures S3–S6.

DOI: 10.1021/acs.analchem.8b05827
Anal. Chem. 2019, 91, 5706–5714

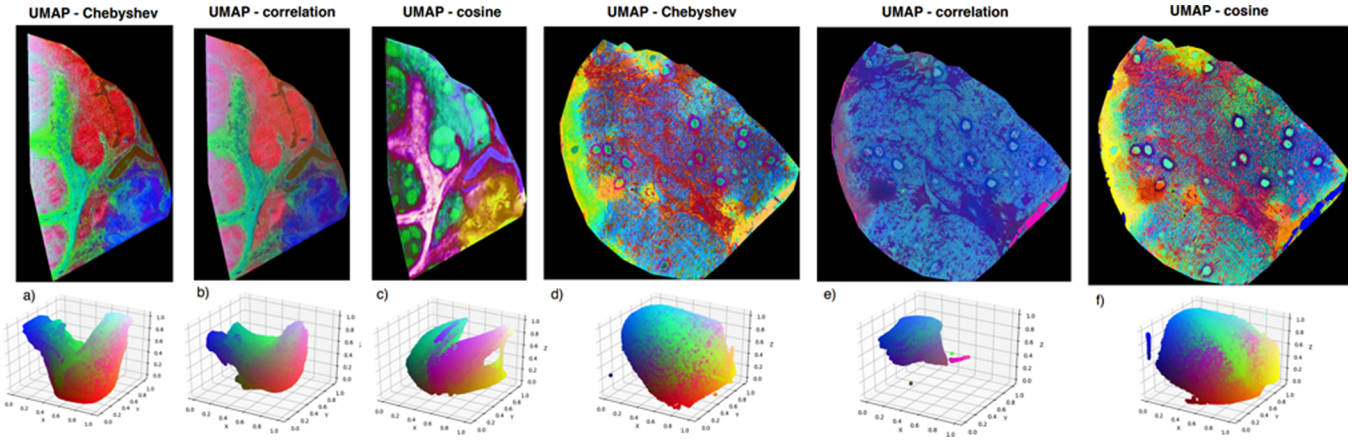


Figure 6. Comparison of the Chebyshev, correlation, and cosine distance metrics using UMAP for the P2 (a–c) and P1 (d–f) lymphoma samples. Shown on top are the 3D embeddings of the tissue using hyperspectral visualization, with the corresponding scatter plots underneath. (e,f) Importance of choosing a good distance metric. Using the correlation and cosine distance metrics, a series of outliers could be detected. Figure 7 shows the impact of removing these outliers on the hyperspectral visualization as a result of the improved colorspace utilization.

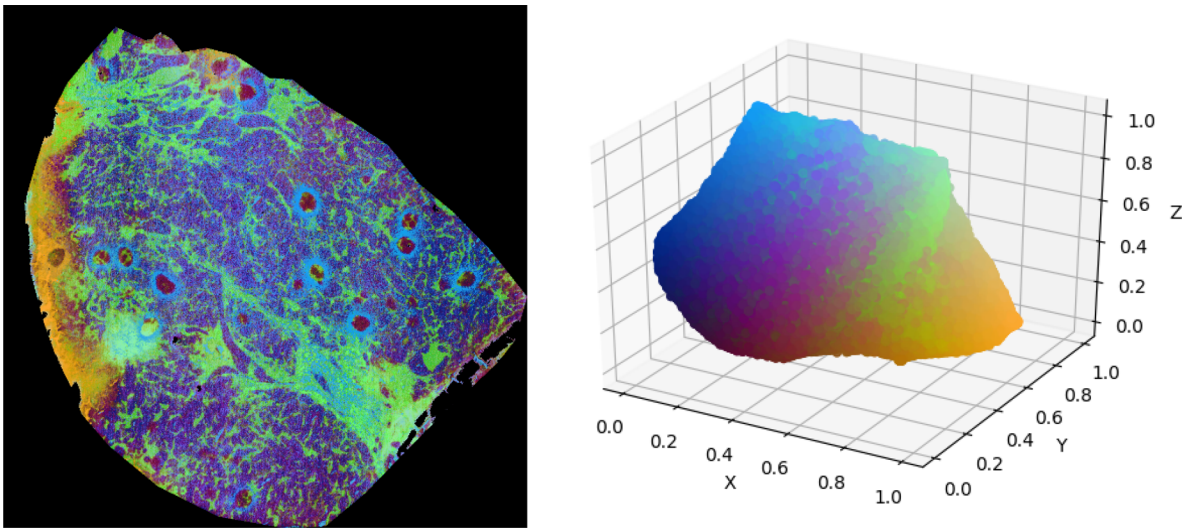


Figure 7. Hyperspectral visualization of the lymphoma P1 sample with the corresponding scatterplot. The improved visualization is due to the removal of outliers detected by using the correlation distance metric in the UMAP algorithm, as shown in Figure 6e. The removal of these outliers makes the utilization of the complete colorspace possible for the remaining points, resulting in an enhanced hyperspectral visualization of the data.

Underestimated Importance of the Distance Metric. Besides the value of NLDR methods for the analysis of MSI data, we can also question the suitability of the Euclidean distance measure in these analyses. As Aggarwal et al.[22] have noted previously, the choice of a particular distance metric may significantly improve the results of standard algorithms. Given that in the high-dimensional space, the data becomes sparse, and the concept of proximity or distance becomes less meaningful, the authors have shown that the Euclidean distance metric is not an ideal metric to be used in the high-dimensional space. Taking this information into account, we have evaluated the effect of using different distance metrics when applying BH t-SNE and UMAP to our data.

In Figure 5, a comparison is made between UMAP and BH t-SNE using the Chebyshev, correlation, and cosine distance metrics.

Figure 6 shows the embeddings for the lymphoma samples using the Chebyshev, correlation, and cosine distance metrics. For the P2 sample, it is clear to see how using the cosine metric strongly increases the level of detail observable in the resulting embedding. Moreover, as shown in Figure 6e,f, using the correlation and cosine metrics facilitates the detection of outliers, which take up a major part of the available colorspace. Therefore, their removal strongly improves the hyperspectral visualization because it enables the utilization of the complete colorspace. This is visualized in Figure 7, where the outliers detected by using the correlation distance metric were removed as an example.

It is important to note here that all data were TIC-normalized before the application of the dimensionality reduction methods because this step is commonly applied within the MSI field. However, to ensure that this normalization does not affect our observations, we have also included a comparison of the different distance metrics applied to the data without TIC normalization. These results are shown in Figures S9–S14. Overall, we can conclude the cosine distance metric, independently of whether the data was TIC-normalized or not, delivers good results across the different methods and tissues. This is in agreement with previous research by Winderbaum et al., who observed that *k*-means clustering of

DOI: 10.1021/acs.analchem.8b05827
Anal. Chem. 2019, 91, 5706–5714

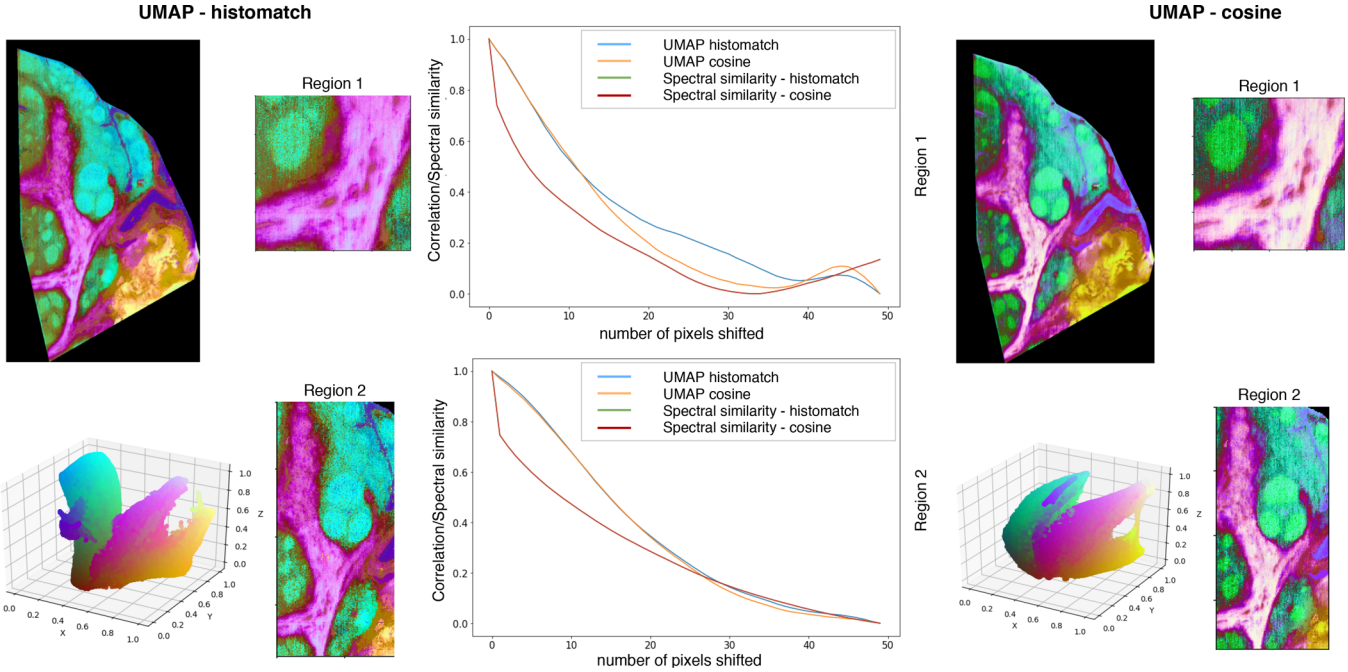


Figure 8. Hyperspectral visualization of P2-lymphoma-sample UMAP embedding using a custom distance metric called histomatch, with the corresponding scatterplot underneath, compared with the results obtained using the cosine distance metric. The graphs show the behaviors of the embeddings in terms of spatial autocorrelation for the two selected regions according to histomatch (blue) and the cosine distance metric (orange). The spatial autocorrelation is supported by the spectral similarity, which is shown in green for the histomatch distance and in red for the cosine metric. The spectral similarities measured according to the cosine and histogram metrics overlap completely, which is supported by Figure S16. Both the hyperspectral visualization and the spatial-autocorrelation plots resemble the results obtained with the cosine distance metric.

MSI data using the cosine distance led to superior results as opposed to the results from the Euclidean distance metric.[23]

Custom Distance-Metric-Histogram Matching. As shown above, the choice of a specific distance metric may significantly alter the resulting embeddings. We have therefore evaluated the performance of an additional distance metric called histomatch. Histogram matching is a method that is often used in computer vision and forms an approximation of correlation, making it an interesting candidate for the analysis of MSI data.

The resulting embeddings for a lymphoma sample obtained according to the histomatch metric, in comparison with those from the cosine metric, are shown in Figure 8; in Figure S15, the results are shown for the pancreas sample. As shown in Figure 8, the hyperspectral visualization of the P2 lymphoma sample is very similar to the one obtained using the cosine metric. This is also reflected in the spatial-autocorrelation results for regions 1 and 2. Therefore histomatch can be a valuable alternative distance metric for the analysis of MSI data. Our observations show that for different data sets, divergent distance metrics can shed alternative light on the same data, which makes experimenting with these metrics worthwhile.

CONCLUSION

We have shown that UMAP yields superior runtimes compared with t-SNE and Barnes−Hut t-SNE for the analysis of MSI data, while obtaining embeddings that are of at least the same quality as those obtained by (BH) t-SNE. In addition, we have demonstrated that spatial autocorrelation can be used for the relative comparison of the results obtained by different NLDR methods. Moreover, we have highlighted the importance of using different distance metrics for performing NLDR. Overall, we can conclude that for the analysis of MSI data, the correlation and cosine distance metrics achieve the best results, and going for the Euclidean distance metric as the standard might not be the best idea. Furthermore, we have presented histomatch as an additional distance metric for the analysis of MSI data. In conclusion, we have shown the value that UMAP, spatial autocorrelation, and different distance metrics can bring to the analysis of MSI data, which we hope will pave the way for future research in this area.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.8b05827.

Overview of the formulas for the applied distance metrics; runtimes associated with t-SNE, BH t-SNE, and UMAP; and additional results for the spatial-autocorrelation experiments for a comparison of the methods applied to the data without TIC normalization and for the histomatch distance metric (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: tina.smets@esat.kuleuven.be.

ORCID

Tina Smets: 0000-0003-1461-4989

Notes

The authors declare no competing financial interest.

DOI: 10.1021/acs.analchem.8b05827
Anal. Chem. 2019, 91, 5706−5714

■ ACKNOWLEDGMENTS

This work was supported by the Flemish Government through the Fonds de la Recherche Scientifique – FNRS and the Fonds Wetenschappelijk Onderzoek – Vlaanderen under EOS Project no 30468160, (SeLMA), PhD grants, IWT: PhD grants (SB/151622), VLAIO: Projects (COT.2018.018), PhD grants (HBC2017.0539), Industrial Projects, (HBC.2018.0405), KU Leuven Internal Funds, C16/15/059, C32/16/013, C24/18/022, iMinds-IMEC ICON project MSIPad, Professor De Moor holds the Chair for "Health Care System Quality & Accessibility" endowed by 'CM Health Insurance'.

■ REFERENCES

(1) Caprioli, R. M.; Farmer, T. B.; Gile, J. *Anal. Chem.* **1997**, *69*, 4751–4760.
(2) Römpp, A.; Spengler, B. *Histochem. Cell Biol.* **2013**, *139*, 759–783.
(3) Alexandrov, T. *BMC Bioinf.* **2012**, *13*, S11.
(4) Race, A. M.; Steven, R. T.; Palmer, A. D.; Styles, I. B.; Bunch, J. *Anal. Chem.* **2013**, *85*, 3071–3078.
(5) Siy, P. W.; Moffitt, R. A.; Parry, R. M.; Chen, Y.; Liu, Y.; Sullards, M. C.; Merrill, A. H.; Wang, M. D. Matrix factorization techniques for analysis of imaging mass spectrometry data. *Proceedings of the 8th IEEE International Conference on Bioinformatics and Bioengineering (BIBE 2008)*, Athens, Greece, Oct 8–10, 2008.
(6) Trindade, G.; Abel, M.-L.; Watts, J. *Chemom. Intell. Lab. Syst.* **2017**, *163*, 76–85.
(7) Hanselmann, M.; Kirchner, M.; Renard, B. Y.; Amstalden, E. R.; Glunde, K.; Heeren, R. M. A.; Hamprecht, F. A. *Anal. Chem.* **2008**, *80*, 9649–9658.
(8) Ma, S.; Dai, Y. *Briefings Bioinf.* **2011**, *12*, 714–722.
(9) Kaddi, C. D.; Wang, M. D. In *Health Informatics Data Analysis: Methods and Examples*; Xu, D., Wang, M. D., Zhou, F., Cai, Y., Eds.; Springer International Publishing: Cham, 2017; pp 37–49.
(10) Aggarwal, C. C. *Machine Learning for Text*, 1st ed.; Springer Publishing Company, 2018.
(11) Fonville, J. M.; Carter, C. L.; Pizarro, L.; Steven, R. T.; Palmer, A. D.; Griffiths, R. L.; Lalor, P. F.; Lindon, J. C.; Nicholson, J. K.; Holmes, E.; Bunch, J. *Anal. Chem.* **2013**, *85*, 1415–1423 PMID: 23249247 .
(12) van der Maaten, L.; Hinton, G. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
(13) van der Maaten, L. *arXiv* **2013**, 1301.3342.
(14) Thomas, S. A.; Race, A. M.; Steven, R. T.; Gilmore, I. S.; Bunch, J. Dimensionality reduction of mass spectrometry imaging data using autoencoders. *Proceedings of the 2016 IEEE Symposium Series on Computational Intelligence (SSCI 2016)*, Athens, Greece, Dec 6–9, 2016.
(15) McInnes, L.; Healy, J.; Melville, J. *arXiv* **2018**, 1802.03426v2.
(16) Cassese, A.; Ellis, S. R.; Ogrinc Potocnik, N.; Burgermeister, E.; Ebert, M.; Walch, A.; van den Maagdenberg, A. M. J. M.; McDonnell, L. A.; Heeren, R. M. A.; Balluff, B. *Anal. Chem.* **2016**, *88*, 5871–5878.
(17) McCune, B.; Grace, L.; Urban, D. *Analysis of Ecological Communities*; MjM Software Design: Gleneden Beach, OR, 2002.
(18) Castleman, K. R. *Digital Image Processing*; Prentice Hall Press: Upper Saddle River, NJ, 1996.
(19) Gonzalez, R. C.; Woods, R. E. *Digital image processing*; Prentice Hall: Upper Saddle River, NJ, 2008.
(20) Dong, W.; Moses, C.; Li, K. Efficient K-nearest Neighbor Graph Construction for Generic Similarity Measures. *Proceedings of the 20th International Conference on World Wide Web*, Hyderabad, India, March 28 to April 1, 2011; ACM: New York, NY, 2011; pp 577–586.
(21) Griffith, D. A.; Wong, D. W. S.; Whitfield, T. *J. Reg. Sci.* **2003**, *43*, 683–710.
(22) Aggarwal, C. C.; Hinneburg, A.; Keim, D. A. On the Surprising Behavior of Distance Metrics in High Dimensional Spaces. In *Database Theory—ICDT 2001*, Proceedings of the 8th International Conference, London, U.K., Jan 4–6, 2001; Springer, 2001; pp 420–434.
(23) Winderbaum, L. J.; Koch, I.; Gustafsson, O. J. R.; Meding, S.; Hoffmann, P. *Ann. Appl. Stat.* **2015**, *9*, 1973–1996.

DOI: 10.1021/acs.analchem.8b05827
Anal. Chem. 2019, 91, 5706–5714