analytical chemistry

ACS AUTHORCHOICE

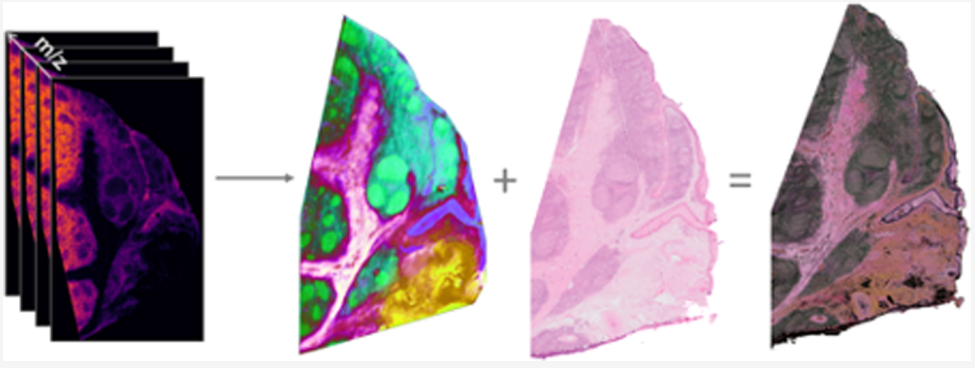
pubs.acs.org/ac Article

Correspondence-Aware Manifold Learning for Microscopic and Spatial Omics Imaging: A Novel Data Fusion Method Bringing Mass Spectrometry Imaging to a Cellular Resolution

Tina Smets,* Tom De Keyser, Thomas Tousseyn, Etienne Waelkens, and Bart De Moor

Cite This: *Anal. Chem.* 2021, 93, 3452−3460 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: High-dimensional molecular measurements are transforming the field of pathology into a data-driven discipline. While hematoxylin and eosin (H&E) stainings are still the gold standard to diagnose diseases, the integration of microscopic and molecular information is becoming crucial to advance our understanding of tissue heterogeneity. To this end, we propose a data fusion method that integrates spatial omics and microscopic data obtained from the same tissue slide. Through correspondence-aware manifold learning, we can visualize the biological trends observed in the high-dimensional omics data at microscopic resolution. While data fusion enables the detection of elements that would not be detected taking into account the separate data modalities individually, out-of-sample prediction makes it possible to predict molecular trends outside of the measured tissue area. The proposed dimensionality reduction-based data fusion paradigm will therefore be helpful in deciphering molecular heterogeneity by bringing molecular measurements such as mass spectrometry imaging (MSI) to the cellular resolution.

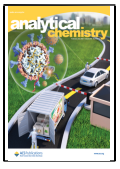
INTRODUCTION

Pathologists have been relying on morphology-based methods for decades to study and diagnose diseases. While such staining approaches enable the assessment of one or two markers in a single tissue slide, spatial transcriptomic and proteomic studies make it possible to evaluate many thousands of molecules simultaneously. The number of studies gathering high-dimensional omics measurements keeps growing in an effort to understand the complex interactions taking place in biological systems. These studies have moved from focusing on single components (e.g., gene) to encompassing the entire genome, and even evaluating complementary omics measurements in parallel (e.g., transcriptomics, proteomics, etc.).[1,2]

Increasingly, these components are being evaluated in terms of their spatial organization as well. A prominent example is the field of mass spectrometry imaging (MSI), which is capable of detecting thousands of endogenous (small metabolites, lipids, peptides, and proteins) or exogenous (drugs and drug metabolites) species in their spatial context.[3] Another important example is the spatial transcriptomics field that spatially resolves the distribution of gene expression profiles to improve our molecular understanding of tissues.[4]

Often, alongside these molecular measurements other imaging modalities, such as high-resolution microscopy images, are being collected as well. Different modalities obtained from the same sample can provide relevant complementary information that cannot be obtained from a single modality.[5] Typically, histological or microscopic images (e.g., haematoxylin and eosin (H&E) stainings) are overlaid with the molecular profiles obtained by MSI. These microscopy images give insight into the correlation between the structure and (pathological) function of cells and tissues that is complementary to the molecular information obtained using

Received: November 11, 2020
Accepted: January 21, 2021
Published: February 8, 2021



https://dx.doi.org/10.1021/acs.analchem.0c04759
Anal. Chem. 2021, 93, 3452−3460

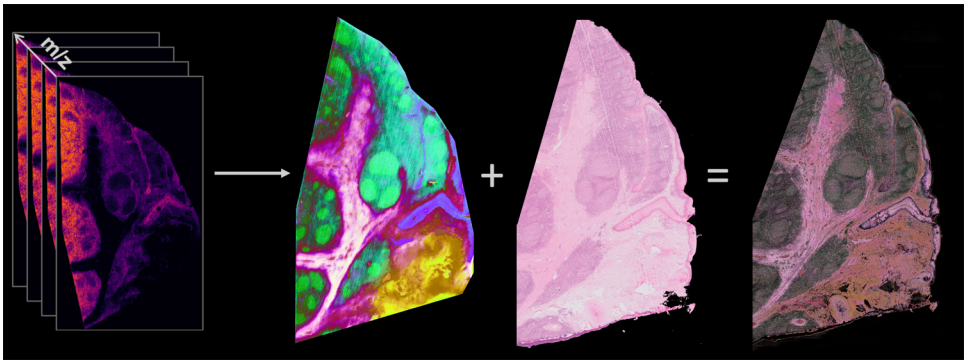


Figure 1. Conceptual overview. Correspondence-aware manifold learning is able to fuse the complete molecular feature space of high-dimensional omics data with their corresponding microscopy image. An example is shown for MSI data where the data set is first reduced to three dimensions and this representation is fused with the microscopy data such that all molecular trends are visualized at a much higher resolution.

molecular imaging. Creating these overlays, however, is challenging because it is difficult to register two images obtained at different spatial resolutions. It is therefore important to overcome these challenges and truly integrate these heterogeneous data sources, a concept referred to as data fusion. By generating a single view or image from a set of source images, we can get a more complete picture of the complex interdependencies present in biological phenomena. In this article, we present a novel data fusion method called "correspondence-aware manifold learning" (CAML) that builds on recent developments in the dimensionality reduction field.

Nonlinear dimensionality reduction methods such as t-distributed stochastic neighbor embedding (t-SNE)[6] are often used for the visualization of high-dimensional biological data.[7] Not only are these methods capable of detecting nonlinear trends, they can also capture the complete feature space when reducing data to two or three components, which is not always the case for methods such as, for example, principal component analysis (PCA).[8] Recently, uniform manifold approximation and projection (UMAP)[9] was introduced to this family of methods with major improvements in terms of scalability, enabling the analysis of large spatial omics data such as MSI.

In an earlier work, we have shown how the hyperspectral visualizations obtained using UMAP can reflect the molecular trends present in an entire tissue sample.[10] Connecting these molecular trends to histological information is essential to support clinical settings. These images, obtained from the same subject or tissue sample but acquired in different ways, are expected to show some level of correspondence. The same anatomical structures can be displayed in the microscopic image and can also be reflected by the molecular trends. Where the microscopic information will typically have a lower "chemical resolution" but a very high spatial resolution, the molecular trends are complementary in this regard, as they offer very rich chemical information but at a lower spatial resolution. With correspondence-aware manifold learning we are now able to visualize these molecular trends at a higher spatial resolution by fusing the molecular and microscopic data (Figure 1). Moreover, using out-of-sample prediction we can predict the distribution of these molecules for regions not measured by the molecular measurements such that we can enrich a complete microscopy slide with the biological signals available. We demonstrate our approach for the fusion of representative spatial omics and optical microscopy data.

METHODS

UMAP Outline. UMAP creates a topological structure that represents the high-dimensional data by assembling approximations of local manifolds and assembles an equivalent topological structure for a low-dimensional representation of the data. It then optimizes the low-dimensional representation to the high-dimensional data by minimizing the cross-entropy between the two topological structures.[9] The algorithm innovates by its mathematical foundations to make some assumptions about the data. An important assumption often used in manifold approximation is a uniform distribution of the data on the manifold.[11] For real world data, this is usually not the case. UMAP addresses this problem by creating local Riemannian manifold approximations on which the data is assumed to be uniformly distributed and patching them together into a fuzzy simplicial set representation of the data.

UMAP uses the fuzzy set cross-entropy to compare the two fuzzy simplicial set representations, (X,v) for the high-dimensional data and (X,w) for the low-dimensional data, for which X is the carrier set and v and w are membership functions upon X. A low-dimensional embedding can be optimized with respect to the cross-entropy loss with v and w as catalysts for attraction and repulsion, respectively

$$C_{\mathrm{UMAP}}((X, v), (X, w)) \sum_{x \in X} v(x) \log\left(\frac{v(x)}{w(x)}\right) + (1 - v(x)) \log\left(\frac{1 - v(x)}{1 - w(x)}\right) \quad (1)$$

For more information, we refer to Section 2.3 Definition 10 in the original UMAP paper.[9]

In general UMAP fits well within the family of algorithms such as t-SNE[6] or LargeVis.[12] These algorithms rely on different mathematical principles although their implementations have lots of common ground. Like t-SNE and LargeVis, manifold approximations are implemented as weighted k-neighbor graphs. It is explained in detail in ref 9 that the main equations from these algorithms also share similarities. Any of these algorithms would suit the data fusion method explained here.

In an earlier work, we have shown the strong visualization capabilities of UMAP for MSI data, making it an excellent choice as a general purpose algorithm for high-dimensional omics data.[10] UMAP is therefore used and extended to fit the desired data fusion goals. We use UMAP as a dimensionality

https://dx.doi.org/10.1021/acs.analchem.0c04759
Anal. Chem. 2021, 93, 3452−3460

reduction algorithm for high-dimensional omics data and adapt UMAP to fuse the resulting low-dimensional representation with high-resolution imaging.

Capturing Correspondence. Our goal is to capture spatial correspondence between high-resolution and high-dimensional data and leverage this information into the manifold learning process. Let us use the matrix $A_{n\times p}$ to denote the flattened high-resolution data and $B_{m\times q}$ for the low-resolution spectral data. The correspondence between these two data sets is recorded in the matrix $C_{n\times m}$ such that

$$C_{ij} = \begin{cases} 1 & \text{if } A_i \text{ corresponds to } B_j \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Finding matching pairs can be achieved with registration techniques. Geometric transformation algorithms can estimate a projection between the two coordinate spaces using only a small set of matching pixel pairs. The output is a transformation matrix that can be applied to warp all other pixels from one image to the other. In this work, we relied on the Python Scikit-image library for image processing to perform registration and obtain a transformation matrix through affine transformation. More specifically, we identified only four landmark points in both images to estimate the transformation matrix. We have observed that CAML is robust against small registration deviations.

Correspondence-Aware Manifold Learning (CAML). Correspondence-aware manifold learning (CAML) creates a fused representation of two data sets such that its corresponding instances lie close to each other in the fused representation. Specifically, CAML models the fusion task as an optimization problem that aims to (1) preserve local distances within the first data set and to (2) minimize distances of the corresponding instances with the second data set.

Consider the task of fusing the high-resolution data matrix $A_{n\times p}$ and high-dimensional data matrix $B_{m\times q}$ based on a correspondence matrix $C_{n\times m}$. In this case, $n > m$ such that multiple instances in A correspond to a single instance in B. The information of the correspondence matrix $C_{n\times m}$ is reconstructed as a mapping γ between the index sets of both data sets.

Definition 1. Define γ: $I \rightarrow J \cup \{0\}$, the correspondence map for matrices $A_{n\times p}$ and $B_{m\times q}$ and their respective index sets $I = \{1, 2, ..., n\}$ and $J = \{1, 2, ..., m\}$, as

$$\gamma(i) = \begin{cases} j & \text{if } A_i \text{ corresponds to } B_j \\ 0 & \text{otherwise} \end{cases}$$

As a consequence, $\gamma(i) = 0$ when there are no corresponding instances for A_i in B.

We capture the concept of distance between the corresponding points as an interplay of attraction and repulsion. CAML aims to minimize the repulsion between the corresponding points. We formally define the repulsion between A and B:

Definition 2. Consider matrices $A_{n\times d}$ and $B_{m\times d}$. Let γ be the correspondence map between the index sets of A and B. Define Ψ: $\mathbb{R}^d \rightarrow \mathbb{R}$, the repulsive strength between these matrices, as

$$\Psi(A_i) = \|A_i - B_{\gamma(i)}\|_2 \cdot \min\{1, \gamma(i)\}$$

Note that when no corresponding instance is found, then $\gamma(i) = 0$ and also $\Psi(A_i) = 0$.

Definition 2 defines repulsion by comparing instances of the two data sets, which is done by choosing a shared dimension d for both data sets. In our case, $d = p = 3$, similar to the RGB color space. The second data set can then be reduced to d dimensions using any dimensionality reduction algorithm, resulting in the two data sets $A_{n\times d}$ and $B_{m\times d}$.

Given the two data sets $A_{n\times d}$ and $B_{m\times d}$ and their correspondence map γ, CAML can be formulated as a constrained optimization problem for a manifold learning cost function C_{MAN}.

$$\min_A C_{\text{MAN}}(A) \text{subject to} \Psi(A_i) = 0,\ i \in \{1, 2, ..., n\} \quad (3)$$

This equality-constrained problem can be transformed into the following quadratic penalty function, which concludes the CAML cost function

$$Q(A; \mu) = C_{\text{MAN}}(A) + \frac{\mu}{2}\sum_{i=1}^{n} \Psi^2(A_i) \quad (4)$$

The first term in eq 4 focuses on preserving local distances within the data. In our case, we use UMAP to optimize A with respect to the fuzzy set cross-entropy defined in eq 1. The mapping to fuzzy set representations is done by UMAP. The second term penalizes the repulsion between the corresponding instances in the two data sets. Because the penalty term in $Q(A; \mu)$ is smooth, we can use unconstrained optimization methods to find a solution.

The penalty parameter μ controls the balance between the two terms. A high value for μ increases the importance of the correspondence information while a low value for μ focuses on the manifold projection of A. Figure S1 provides an overview of running the algorithm using different values for μ.

The optimal value for μ can also be learned using iterative methods, however, optimizing $Q(A; \mu)$ is costly depending on the underlying manifold learning algorithm. Better means of evaluating low-dimensional embeddings have been published instead of calculating $Q(A; \mu)$ directly. We used the following as an alternative

$$E(A; \mu) = 1 - \left((1-T)^2 + \sum_{i=1}^{n} \Psi^2(A_i)\right) \quad (5)$$

We want to capture both the quality of the manifold embedding $A_{n\times d}$ and the similarity between the low-resolution image $B_{m\times d}$. The first concept can be tackled using the measure of trustworthiness.[13,14] Trustworthiness evaluates to what extent the local structure within the data is retained in a manifold embedding. For the second concept, we already defined repulsion to calculate the similarity between the corresponding points of images. Equation 5 thus expresses the trade-off between the trustworthiness T of the embedding and the repulsion Ψ of the corresponding points in a value between 0 and 1. A represents the fused result after solving the optimization problem from eq 3. A high value for $E(A; \mu)$ corresponds to a high trustworthiness and a low repulsion. Figure S1 presents the values for $E(A; \mu)$ for each of the embeddings.

Currently CAML has been implemented as an extension of the UMAP algorithm because of its general applicability. The implementation is based on the model implementation of UMAP by the original author. Although CAML expands the algorithm, to our knowledge the implementation does not impose additional theoretical complexities on UMAP and does

https://dx.doi.org/10.1021/acs.analchem.0c04759
Anal. Chem. 2021, 93, 3452−3460

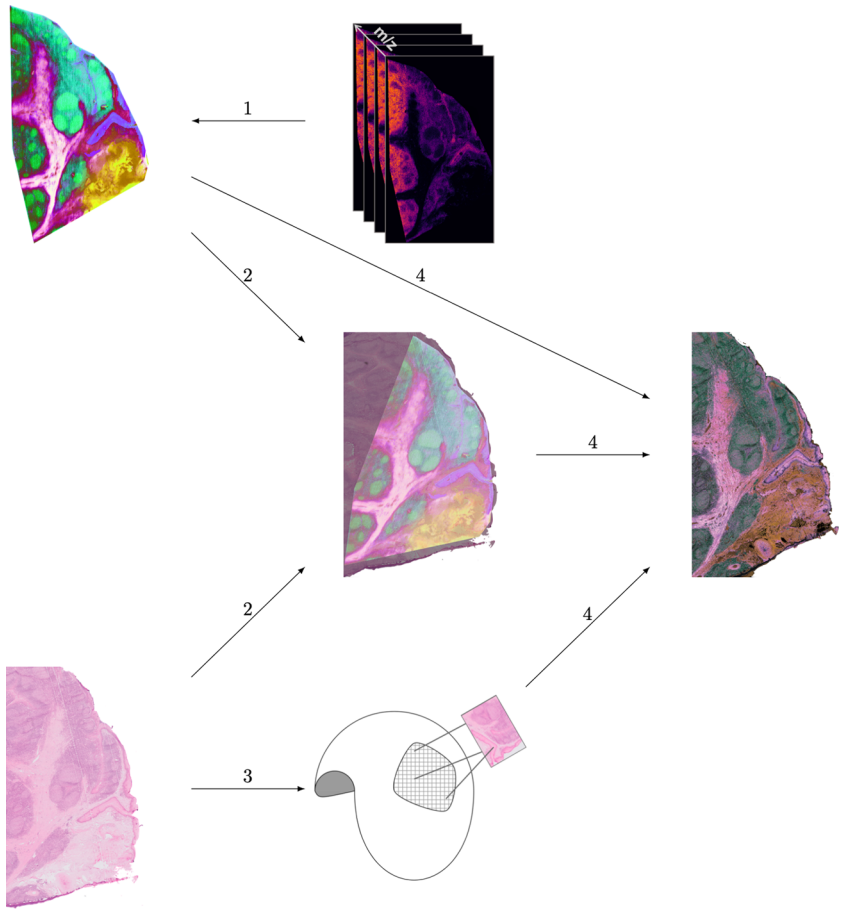


Figure 2. Method overview. (1) Molecular data represented by the matrix $B_{m\times q}$ is reduced to a matrix $B_{m\times 3}$ target embedding. (2) Pixel coordinates of the molecular and the microscopy image undergo a registration step such that we obtain a correspondence matrix. (3) Subsequently, the microscopy image is subjected to a dimensionality reduction step, wherein each pixel is evaluated as a function of its correspondence to target embedding. (4) Specifically, the projection step in the dimensionality reduction method is constrained based on target embedding, causing pixels in the microscopy image to receive a similar color based on the reduced target embedding of the molecular data. This approach not only enables the transfer of information obtained from a complete high-dimensional molecular data set to a single microscopy slide but can also be used to transfer the information from a single feature or molecular image to the microscopy slide.

not remove performance improvement made to the algorithm. The UMAP algorithm also makes it possible to embed data based on an existing embedding. Leveraging this data transformation option together with image slicing releases the computational burden imposed by very large images.

All images are converted to the LAB color space prior to fusing. The LAB color space is used to approximate a uniform color space (i.e., a color space in which same-size changes in the color coordinates correspond to same-size recognizable changes in the visible color tones and color saturation).[15]

Only the dimensionality reduction step for the lymphoma MSI data set has been done on an Intel Xeon CPU E5-2660 v2 2.20 GHz machine with 10 cores and 128 GB RAM. All other experiments have been done on a MacBook Pro with a 2.8 GHz Intel Core i7 CPU and 16 GB RAM.

Data Measurements. For the human lymph node sample, cryosections of 5 μm thickness were prepared and mounted on indium tin oxide (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. For the mouse brain spatial transcriptomics samples, H&E images and count matrices were downloaded from https://www.spatialresearch.org/resources-published-datasets/ licensed under the Creative Commons Attribution license. The dimensions of the data sets are as follows: for the lymphoma data set 500 000 pixels × 8000 m/z values and for the spatial transcriptomics data 281 pixels × 16 416 transcripts.

RESULTS

With correspondence-aware manifold learning we:

(i) project the high-dimensional molecular features to a low-dimensional space,
(ii) capture spatial correspondence between the high-resolution microscopic image and the high-dimensional molecular measurements of the same tissue sample, and
(iii) perform correspondence-aware manifold projection using both data modalities to obtain a fused image reflecting both modalities in one visualization.

The general methodology is depicted in Figure 2. The high-dimensional molecular data is first reduced to three dimensions. After a registration step, this hyperspectral visualization is used as a constraint to transform the microscopy image, resulting in a fusion of the molecular information with the microscopy data. By modeling the

https://dx.doi.org/10.1021/acs.analchem.0c04759
Anal. Chem. 2021, 93, 3452−3460

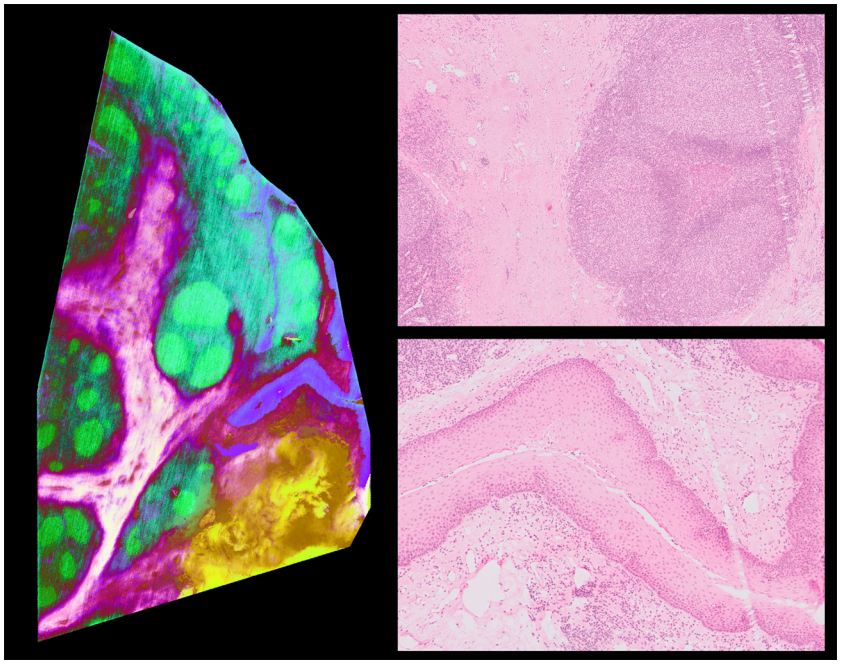


Figure 3. Low-dimensional representation of a lymphoma MSI data set and the corresponding H&E image. Shown on the left: the low-dimensional representation of a human lymphoma MSI data set (500.000 pixels × 8000 m/z features, 10 μm resolution). The different colors reflect the molecular trends present in the data. On the right, two parts of the corresponding microscopy image are shown.

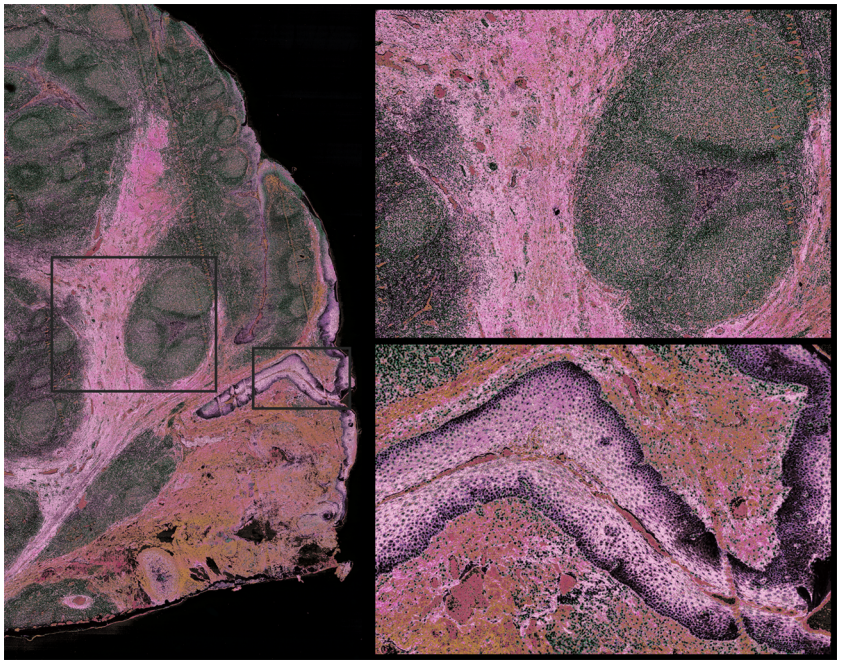


Figure 4. Fusion of mucosa-associated lymphoid tissue of the tonsil MSI data set and the corresponding H&E image. On the left, the data fusion result for the molecular and microscopy images is shown. On the right, the fusion details are shown for the two regions.

manifold and piecewise transforming the full-resolution microscopy image while taking into account the properties of the molecular data, we are able to visualize the molecular image at a much higher resolution. As such we leverage the complementarity of the high spatial resolution offered by optical microscopy with the high-dimensional but lower spatial resolution molecular imaging data.

We demonstrate the correspondence-aware manifold learning approach for data fusion of molecular measurements and their corresponding microscopy images. We show:

(i) the prediction of molecular trends at a higher spatial resolution through data fusion,

(ii) the prediction of molecular trends outside of the tissue area measured with out-of-sample prediction, and

(iii) the general applicability of the method.

Correspondence-Aware Manifold Learning for Data Fusion of Molecular and Microscopy Images. In Figure 3, we show the low-dimensional representation of reactive lymphoid tissue in a human tonsil MSI data set (500.000

https://dx.doi.org/10.1021/acs.analchem.0c04759
Anal. Chem. 2021, 93, 3452−3460

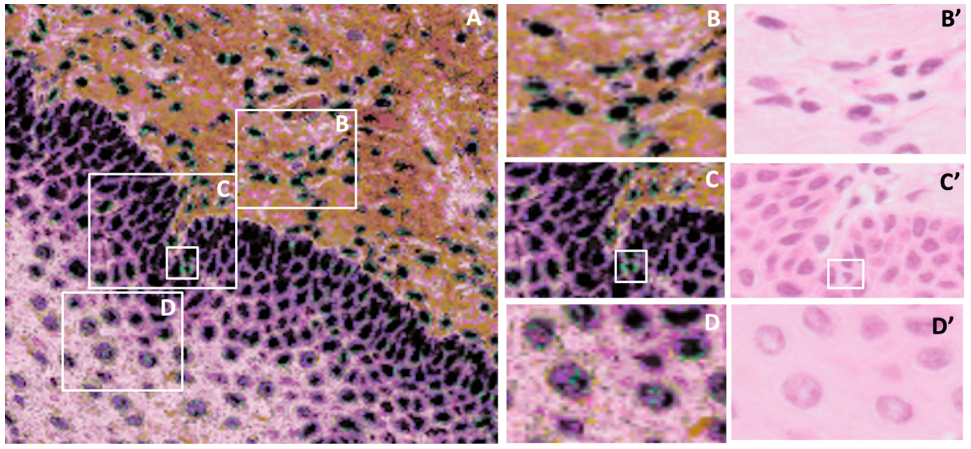


Figure 5. Details of the fused data. Panel (A) shows the extracellular matrix and stroma comprising glycoproteins and proteoglycans, as well as plasma cells on top and the multilayered squamous epithelium below. Panel (B) shows a detailed view of a cluster of plasma cells. The nuclei of these cells received a black color after data fusion, while the green color seems to correspond to the abundant cytoplasm. In active plasma cells a high density of the golgi apparatus is required for the synthesis of immunoglobulins, which explains the larger amount of cytoplasm present. A closer look in panel (C) reveals that these plasma cells infiltrate the epithelium, as they display a kind of integration with the present keratinocytes (purple color) such that they do not overlay or compress these cells. Panel (D) shows that the keratinocyte cells further away from the basement membrane have become larger in comparison to the ones closer to this basal layer, which can be explained by the progressive maturation of these cells. These findings are supported by the corresponding H&E stainings in panels (B′)−(D′). Note that these results are provided at the cellular level: individual cells with their particular nuclei and cytoplasm are shown. This demonstrates the power of data fusion to surpass the regional or subregional level of interpretation offered via MSI by bringing the molecular information to the cellular level.

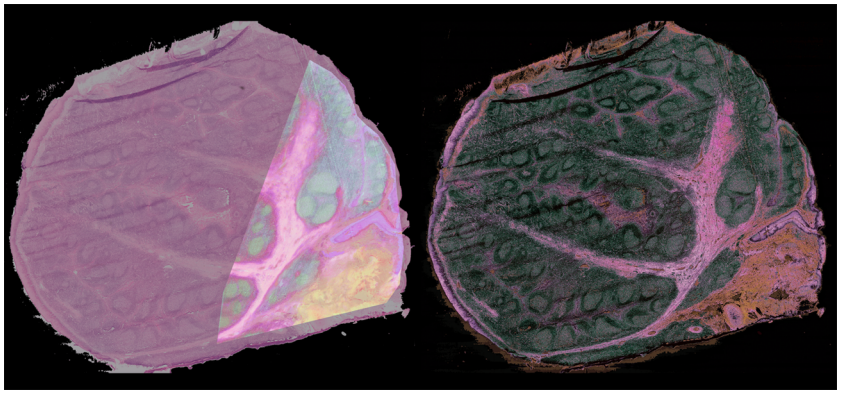


Figure 6. Example of out-of-sample prediction. On the left, an overlay of the region in the lymphomoid tissue measured by MSI with the microscopy slide is shown. On the right, the result of out-of-sample prediction shows the fused result for the complete microscopy or H&E image.

pixels × 8000 m/z features, measured at 10 μm resolution). This hyperspectral visualization represents the complete 8000 m/z feature space compressed into three dimensions, such that each color is connected to a molecular trend present in the data. This molecular, hyperspectral visualization is subsequently used to perform data fusion with the corresponding microscopy image. In Figure 4, the fused results show that the molecular trends in the data can be visualized at a much higher resolution.

The goal of performing data fusion is to exploit the complementarity present between the different data modalities such that the resulting visualization goes beyond the information offered by a single modality. In Figure 5, a fused detail is shown of the extracellular matrix and stroma comprising glycoproteins and proteoglycans with below the multilayered squamous epithelium (panel A). Panel B shows a detailed view of a cluster of plasma cells. The nuclei of these cells received a black color after data fusion, while the green color seems to correspond to the cytoplasm. In active plasma cells, a high density of the golgi apparatus is required for the synthesis of immunoglobulins, which explains the larger amount of cytoplasm present. A closer look reveals that these plasma cells infiltrate the epithelium (panel C) and display a kind of integration with the present epithelial cells (purple color) such that they do not overlay or compress these keratinocytes. Panel D shows that the keratinocyte cells further away from the basement membrane have become larger in comparison to the ones closer to this basal layer, which can be explained by the progressive maturation process taking place inside of the squamous epithelium. This maturation process is associated with changes in the composition of the cytoplasm with mainly an increase in the number of cytokeratines, which are part of the cytoskeleton. These findings are supported by the corresponding H&E stainings in panels B′−D′, and show the power of data fusion to surpass the regional or subregional level of interpretation offered via MSI by bringing the molecular information to the cellular level.

In the Supporting Figures, additional examples highlight the potential of data fusion. In Figure S2, two blood vessels are shown where we can see that a venule on the left is surrounded

https://dx.doi.org/10.1021/acs.analchem.0c04759
Anal. Chem. 2021, 93, 3452−3460

by a thin layer of endothelial cells, while the arteriole on the right is lined by a layer of smooth muscle cells. In light pink, we can also perceive some collagen fibers. Figure S3 demonstrates a secondary B-cell follicle where the reactive germinal center is surrounded by a lymphocyte corona, highlighted by an orange and green dashed line, respectively. Figure S3 shows how data fusion can support us in distinguishing artifacts from true biological signals. The epithelium contains a small, rounded structure where the number of cells is increased. While this structure is visible in the H&E image, it becomes more pronounced upon data fusion. It is regarded as an artifact created during the cutting process.

All findings are supported by the corresponding H&E images. Moreover, these results highlight the potential of our method to distinguish individual cells in their microenvironment instead of being limited to the interpretation of regional or subregional molecular trends measured by MSI. This could be very valuable when evaluating, for instance, the invasiveness of individual tumor cells and their interaction with the tumor microenvironment.

Out-of-Sample Prediction. In addition to performing data fusion for the area covered by the MSI measurements, we can predict the molecular distributions for the entire microscopy image through out-of-sample prediction, as shown in Figure 6. This approach enables the interpretation of a much larger microscopic area based on a limited amount of molecular information. This can be a valuable asset given the high costs associated with molecular measurements or the limited amount of tissue that is often available.

Data Fusion for Spatial Transcriptomics Data. We illustrate the general applicability of the method for spatial omics data. In Figure S5, we show the hyperspectral visualization of the low-dimensional representation obtained for a spatial transcriptomics mouse brain measurement and the corresponding microscopy image with the fused result on the right. We show these results to highlight the potential of this method for other spatial omics technologies. Notwithstanding the low spatial resolution of the molecular measurement (281 pixels × 16.416 features), we can see that the green colored cell nuclei are embedded in a purple background of the tissue center. Given the low spatial resolution, we want to emphasize the restricted potential for biological interpretation. However given the fast technical improvements that are being made in terms of spatial resolution, we believe the proposed method holds a lot of potential for data fusion of spatial omics measurements. In this light, in Figure S5, we can also observe that the highlighted regions show a clear correspondence of the green dots with the cell nuclei, as stained by hematoxylin, across the associated H&E image.

DISCUSSION

Scalable and powerful dimensionality reduction methods have become indispensable to deal with the growing number of high-dimensional data sets. Nonlinear dimensionality reduction methods, such as t-SNE and UMAP, have brought and continue to bring significant value for the biomedical sciences in this regard.[16,17] Due to their strong visualization capabilities these methods have become a standard for the analysis of high-dimensional data sets.[18] And while the number of high-dimensional and spatial omics measurements keeps on growing, the computational methods capable of fusing the multimodal measurements acquired from the same sample are lagging behind. In this work, we present a novel data fusion method that is able to compress and fuse the complete molecular and microscopic feature spaces toward a combined image. This work builds on the framework of nonlinear dimensionality reduction methods to enable the fusion of molecular and microscopic data obtained from the same tissue sample. We demonstrate our results according to the UMAP framework, but the same principle could be applied starting from similar methods such as, for example, t-SNE. By constraining the projection step based on the molecular target information, we are able to transform the corresponding pixels in the microscopy image accordingly, resulting in a fused representation presenting the molecular information at a much higher spatial resolution.

In Figure 4, we highlight the potential of our method for the integration of MSI data with the corresponding microscopy images. While we show that this enables us to improve the resolution of the MSI data, these colors reflect in fact an underlying group of biomolecules. As such, in the previous work, we have shown that it is possible to prioritize and identify those molecules associated with a molecular trend or color.[19] This could support researchers in finding correlations between underlying biological actors and their histopathological architecture. Recent work has shown that it is possible to correlate single-cell morphological features based on microscopic images with molecular information.[20] Given that the proposed method is capable of retaining the cellular morphology in the fused results, we believe it holds potential for this area of study as well. Moreover, in Figure 5 we show that the proposed method can leverage MSI measurements to study tissues at the cellular level such that we can move beyond the regional or subregional insights and evaluate the presence of aberrant cells in their microenvironment. This will be of growing importance with technological advancements in terms of spatial resolution and also with the increasing demand to integrate multimodal data measurements. In this regard, we have also included a spatial transcriptomics sample as an example of a rapidly evolving domain with a lot of potential.[21] While the molecular measurements are at the moment still of a lower resolution, it is yet possible to show that the method is widely applicable and will be able to offer more value with increasing spatial resolution. An additional advantage data fusion has to offer is the ability to better distinguish artifacts from true biological signals (Figure S4). These advantages will be useful not only in the domain of molecular imaging but in general when dealing with other imaging technologies such as, for example, magnetic resonance imaging (MRI), computed tomography (CT), positron emission tomography (PET), etc.

An earlier work in this domain has been focused on the modeling of linear relationships between the different modalities. Van de Plas et al. have focused on the fusion of matrix-assisted laser desorption ionization (MALDI) MSI data with optical images of H&E stainings through the mapping between two modalities based on linear regression models.[22] And recently, a novel method was introduced by Race et al.[23] based on dimensionality reduction through non-negative matrix factorization. While providing good results, given the nonlinear nature of biological data, taking into account these complex relationships is preferable. This is in particular the case for MALDI MSI data which can suffer from artifacts caused by the nonlinear ionization process or for instance by ion suppression. Therefore, the method proposed in this work performs data fusion through a nonlinear dimensionality reduction approach. In addition to the benefit of being able to take more complex

https://dx.doi.org/10.1021/acs.analchem.0c04759
Anal. Chem. 2021, 93, 3452−3460

interactions into account, the data fusion result is not hindered by an imperfect multimodal registration, which can also be seen in Figure 6. Given that registration is often a time-consuming and difficult step, this constitutes an important advantage. To illustrate the general applicability of the method, we have also performed a study on a multimodal MNIST data set, an extension of the well-known MNIST database of handwritten digits. In Figure S6, we show that we can perform data fusion on single multimodal digits and we can also perform out-of-sample prediction based on this initial trained data fusion model, added with additional experimental verification for the MSI data in Figures S7 and S8.

In conclusion, data fusion facilitates the combination of complimentary data sources to obtain insights that would not be obtained from a single modality alone. Given the growing interest toward spatial multiomics studies, this method will be valuable to enable the mapping of molecular measurements to the underlying tissue architecture at the cellular resolution. Moreover, given the large costs associated with state-of-the-art molecular measurements, the out-of-sample prediction can expand the amount of information obtained from conducted experiments. Finally, due to its broad applicability we hope that the proposed paradigm will be valuable to researchers coming from different domains.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.0c04759.

> Further details regarding the optimization of the μ parameter; the fused MSI lymphoma data; additional examples for a spatial transcriptomic data set; and figures regarding conceptual verification and illustration of the CAML approach (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Tina Smets − *STADIUS Center for Dynamical Systems, Signal Processing, and Data Analytics, Department of Electrical Engineering (ESAT), KU Leuven, 3001 Leuven, Belgium;* orcid.org/0000-0003-1461-4989; Email: tina.smets@esat.kuleuven.be

Authors

Tom De Keyser − *Independent Researcher, 3010 Leuven, Belgium*

Thomas Tousseyn − *Department of Pathology, University Hospitals Leuven, 3000 Leuven, Belgium*

Etienne Waelkens − *Department of Cellular and Molecular Medicine, KU Leuven, 3000 Leuven, Belgium*

Bart De Moor − *STADIUS Center for Dynamical Systems, Signal Processing, and Data Analytics, Department of Electrical Engineering (ESAT), KU Leuven, 3001 Leuven, Belgium*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.0c04759

Notes

The authors declare no competing financial interest.

[⊥]B.D.M.: Fellow IEEE, SIAM.

■ ACKNOWLEDGMENTS

The authors would like to thank Arndt Asperger from Bruker Daltonics for infrastructural support. This work was supported by KU Leuven: Research Fund (Projects C16/15/059, C3/19/053, C32/16/013, C24/18/022), Industrial Research Fund (Fellowship 13-0260) and several Leuven Research and Development bilateral industrial projects, Flemish Government Agencies: FWO (EOS Project No. 30468160 (SeLMA), SBO Project S005319N, Infrastructure Project I013218N, TBM Project T001919N; Ph.D. Grants (SB/1SA1319N, SB/1S93918, SB/151622)). This research received funding from the Flemish Government (AI Research Program). B.D.M. and T.S. are affiliated to Leuven. AI-KU Leuven Institute for AI, B-3000 Leuven, Belgium. VLAIO (City of Things (COT.2018.018), Ph.D. Grants: Baekeland (HBC.20192204) and Innovation Mandate (HBC.2019.2209), Industrial Projects (HBC.2018.0405)), European Commission: This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No. 885682); (EU H2020-SC1-2016-2017 Grant Agreement No. 727721: MIDAS); T.T. holds a mandate for Fundamental and Translational Research from the "Stichting tegen Kanker" (2014-083; 2019-091) and is supported by the "Stichting Me to You (https://www.stichtingmetoyou.be/nl/)".

■ REFERENCES

(1) Barsoum, I.; Tawedrous, E.; Faragalla, H.; Yousef, G. M. *Diagnosis* **2019**, *6*, 203−212.

(2) Manzoni, C.; Kia, D. A.; Vandrovcova, J.; Hardy, J.; Wood, N. W.; Lewis, P. A.; Ferrari, R. *Briefings Bioinf.* **2018**, *19*, 286−302.

(3) Buchberger, A. R.; DeLaney, K.; Johnson, J.; Li, L. *Anal. Chem.* **2018**, *90*, 240−265.

(4) Burgess, D. J. *Nat. Rev. Genet.* **2019**, *20*, 317.

(5) Siegel, T. P.; Hamm, G.; Bunch, J.; Cappell, J.; Fletcher, J. S.; Schwamborn, K. *Mol. Imaging Biol.* **2018**, *20*, 888−901.

(6) van der Maaten, L.; Hinton, G. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.

(7) Kobak, D.; Berens, P. *Nat. Commun.* **2019**, *10*, No. 5416.

(8) Pearson, K. *London, Edinburgh Dublin Philos. Mag. J. Sci.* **1901**, *2*, 559−572.

(9) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, *arXiv (Machine Learning)*, December 6, 2018, 1802.03426, ver. 2. https://arxiv.org/abs/1802.03426v3 (accessed 2021-01-21).

(10) Smets, T.; Verbeeck, N.; Claesen, M.; Asperger, A.; Griffioen, G.; Tousseyn, T.; Waelput, W.; Waelkens, E.; De Moor, B. *Anal. Chem.* **2019**, *91*, 5706−5714.

(11) Belkin, M.; Niyogi, P. *Neural Comput.* **2003**, *15*, 1373−1396.

(12) Moon, K. R.; van Dijk, D.; Wang, Z.; Gigante, S.; Burkhardt, D. B.; Chen, W. S.; Yim, K.; van den Elzen, A.; Hirn, M. J.; Coifman, R. R.; et al. *Nat. Biotechnol.* **2019**, *37*, 1482−1492.

(13) Venna, J.; Kaski, S. *Neural Networks* **2001**, 485−491.

(14) Venna, J.; Kaski, S. In *Visualizing Gene Interaction Graphs With Local Multidimensional Scaling*. European Symposium on Artificial Neural Network, 2006; pp 557−562.

(15) Koschan, A.; Abidi, M. A. *Digital Color Image Processing*; Wiley-Interscience, 2008.

(16) Mahfouz, A.; van de Giessen, M.; van der Maaten, L.; Huisman, S.; Reinders, M.; Hawrylycz, M. J.; Lelieveldt, B. P. *Methods* **2015**, *73*, 79−89.

(17) Diaz-Papkovich, A.; Anderson-Trocmé, L.; Gravel, S.; et al. *PLoS Genet.* **2019**, *15*, No. e1008432.

(18) Becht, E.; McInnes, L.; Healy, J.; Dutertre, C.-A.; Kwok, I. W.; Ng, L. G.; Ginhoux, F.; Newell, E. W. *Nat. Biotechnol.* **2019**, *37*, 38−44.

(19) Smets, T.; Waelkens, E.; De Moor, B. *Anal. Chem.* **2020**, *92*, 5240−5248.
(20) Ščupáková, K.; Dewez, F.; Walch, A. K.; Heeren, R. M.; Balluff, B. *Angew. Chem.* **2020**, *132*, 17600−17603.
(21) Ståhl, P. L.; et al. *Science* **2016**, *353*, 78−82.
(22) Van de Plas, R.; Yang, J.; Spraggins, J.; Caprioli, R. M. *Nat. Methods* **2015**, *12*, 366−372.
(23) Race, A. M.; Rae, A.; Vorng, J.-L.; Havelund, R.; Dexter, A.; Kumar, N.; Steven, R. T.; Passarelli, M. K.; Tyler, B. J.; Bunch, J.; Gilmore, I. S. *Anal. Chem.* **2020**, *92*, 10979−10988.

https://dx.doi.org/10.1021/acs.analchem.0c04759
Anal. Chem. 2021, 93, 3452−3460