Bioinformatics, 2019, 1–8
doi: 10.1093/bioinformatics/btz505
Advance Access Publication Date: 21 June 2019
Original Paper

Data and text mining

Deep representation learning for domain adaptable classification of infrared spectral imaging data

Arne P. Raulf[1,2], Joshua Butke[1,2], Claus Küpper[1,2], Frederik Großerueschkamp[1,2], Klaus Gerwert[1,2] and Axel Mosig [1,2]*

[1]Center for Protein Diagnostics (ProDi) and [2]Chair of Biophysics, Department for Biology and Biotechnology, Ruhr-Universität Bochum, 44801 Bochum, Germany

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on December 21, 2018; revised on May 28, 2019; editorial decision on June 5, 2019; accepted on June 13, 2019

Abstract

Motivation: Applying infrared microscopy in the context of tissue diagnostics heavily relies on computationally preprocessing the infrared pixel spectra that constitute an infrared microscopic image. Existing approaches involve physical models, which are non-linear in nature and lead to classifiers that do not generalize well, e.g. across different types of tissue preparation. Furthermore, existing preprocessing approaches involve iterative procedures that are computationally demanding, so that computation time required for preprocessing does not keep pace with recent progress in infrared microscopes which can capture whole-slide images within minutes.

Results: We investigate the application of stacked contractive autoencoders as an unsupervised approach to preprocess infrared microscopic pixel spectra, followed by supervised fine-tuning to obtain neural networks that can reliably resolve tissue structure. To validate the robustness of the resulting classifier, we demonstrate that a network trained on embedded tissue can be transferred to classify fresh frozen tissue. The features obtained from unsupervised pretraining thus generalize across the large spectral differences between embedded and fresh frozen tissue, where under previous approaches separate classifiers had to be trained from scratch.

Availability and implementation: Our implementation can be downloaded from https://github.com/arnrau/SCAE_IR_Spectral_Imaging.

Contact: axel.mosig@bph.rub.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In recent years, the application of label-free infrared microscopy to histopathological tissue samples has paved the way for *spectral histopathology* (Bird *et al.*, 2012; Kallenbach-Thieltges *et al.*, 2013), which has proven to be a reliable approach to assess the disease status of histological sections. Infrared microscopy measures samples with a resolution of few μm and provides an infrared spectrum representing the biochemical tissue status at each pixel location. It has been shown that the pixel spectra obtained from infrared microscopes are highly representative for different tissue components as well as for disease status. As illustrated in Figure 1, this allows supervised classifiers to infer the tissue component or disease status from an infrared pixel spectrum, which has proven successful for several types of cancer ranging from colon carcinoma (Kallenbach-Thieltges *et al.*, 2013; Kuepper *et al.*, 2016) to lung (Bird *et al.*, 2012; Großerueschkamp *et al.*, 2015) and bladder (Großerueschkamp *et al.*, 2017) cancer.

It is commonly observed that besides biomedically relevant molecular signatures, data obtained from highly sensitive bioanalytical techniques contain technological or biological artifacts, background signal and other confounders that mask those features that are relevant towards disease status. For infrared microscopy, such

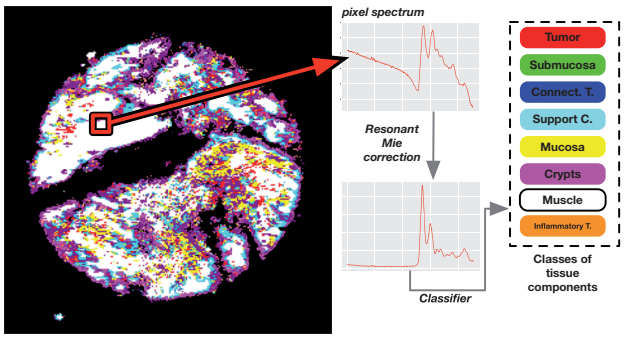

Fig. 1. Principle of spectral histopathology: The infrared spectrum from each pixel position is preprocessed using a physical model and subsequently classified into the respective tissue component. Our newly proposed approach aims to use deep neural networks to classify the uncorrected spectrum, where preprocessing is substituted with an unsupervised deep representation learning step

background signals are particularly severe and complex in nature (Bassan *et al.*, 2010), and several approaches have been proposed to disentangle them from diagnostically relevant signals. The bandwidth of proposed approaches range from methods based on physical models (Bassan *et al.*, 2010; Konevskikh *et al.*, 2018; Romeo and Diem, 2005) to statistical approaches utilizing principal components (Marcsisin *et al.*, 2012). While these methods have contributed to the successful application of spectral histopathology in clinical studies, preprocessing infrared spectra remains subject of active investigation in the community (Konevskikh *et al.*, 2018).

Our contribution breaks with those existing approaches and takes a machine learning perspective on the problem. To elaborate, preprocessing infrared spectra can be viewed as a *representation learning* (Bengio *et al.*, 2013) problem: raw infrared spectra are difficult to classify, so that they need to be transformed into a representation that is more accessible for classification or interpretation. Recent progress on learning such representations in an unsupervised manner suggests that the resulting representations are often more suitable for classification than previous, problem domain specific 'feature engineered' representations (Bengio *et al.*, 2013). The success of unsupervised representation learning is often coupled with the availability of large amounts of data, which are commonly accessible in spectral histopathology. Even a single image commonly contains tens of millions of spectra (Kallenbach-Thieltges *et al.*, 2013) that can be measured within minutes (Kuepper *et al.*, 2018). In other words, the substantial recent progress in the field of representation learning bears great promises for spectral histopathology that we investigate in this contribution.

The aims of our present contribution go beyond the mere assessment of recent progress in representation learning to spectral histopathology. Specifically, we propose an approach to assess the robustness of a learned representation. We achieve this by transferring a classifier that is based on a learned representation to a related domain. In our specific case, we transfer a classifier trained on data obtained from formalin-fixed tissue to data from fresh frozen tissue, which is accompanied by significant changes in the infrared spectra. If the classifier turns out to be transferable, the learned representation can be considered highly robust against sample variability and heterogeneity. As robustness is of predominant importance when classifying biomedical sample material, we consider the domain transfer approach as a major contribution of our present paper, which can be employed to assess the robustness of not just infrared spectral classifiers, but also of classifiers for data obtained from other bioanalytical techniques.

2 Background

2.1 Spectral histopathology

In order to assign disease relevant classes to infrared microscopic pixel spectra, different studies in spectral histopathology have employed a range of different classification approaches. One common approach (Bird *et al.*, 2012; Großerueschkamp *et al.*, 2015; Kallenbach-Thieltges *et al.*, 2013) is to use pixel spectra classifiers to obtain a segmentation of the tissue sample into different physiologically or pathologically relevant components. The segmented image then serves as a basis for a diagnostic characterization, for instance by determining the relative abundance of cancerous or otherwise disease relevant pixels (Yosef *et al.*, 2017). Some studies (Bird *et al.*, 2012; Kallenbach-Thieltges *et al.*, 2013) suggest that resolving other tissue components along with the distinction into pathological versus healthy regions is helpful or even necessary to reliably characterize the disease status. Remarkably, all aforementioned infrared microscopy based studies involve preprocessing of the spectra, typically in the form of physical models that either remove resonant Mie scattering (Bassan *et al.*, 2010) or dispersion 'artefacts' (Romeo and Diem, 2005).

Until recently, most spectral histopathology studies utilized *Fourier transform infrared* (FTIR) microscopes, where the infrared spectrum is derived by Fourier transforming the signal obtained from an interferometer. Very recently, FTIR microscopy has been challenged by quantum-cascade laser (QCL) microscopes, where the spectrum is obtained from frequency tunable quantum cascade lasers. QCL-based microscopes exceed the measurement speed of FTIR-based systems by almost two orders of magnitude (Großerueschkamp *et al.*, 2017), so that infrared images of complete slides of histological sections can be captured within minutes. At the same time, however, the infrared spectrum is limited to a smaller spectral range, which in particular affects the spectral baseline that is important for resonant Mie correction.

2.2 Infrared spectroscopy

To understand the challenges in spectral histopathology, it is worthwhile to introduce some background on infrared spectroscopy. Infrared spectra are well-known to be a highly characteristic fingerprint of the molecular vibrations and hence of the molecular composition of biological samples. In particular, the so-called fingerprint region between 1500 and 500 cm^{-1} is known to be highly specific for the molecular decomposition of the sample. Furthermore, the *amide I* and *amide II* bands are commonly observed as dominant peaks between wavenumbers 1700–1600 and 1600–1500 cm^{-1}, respectively; these amide bands are highly characteristic of protein secondary structure.

Infrared spectrometers in general and infrared microscopes specifically will yield spectra for molecular characterization between 4000 and 900 cm^{-1} due to the used MCT (mercury cadmium telluride) detectors at a typical wavenumber resolution of 1–2 cm^{-1}. It is common practice in infrared microscopy (Bird *et al.*, 2012; Großerueschkamp *et al.*, 2015; Kallenbach-Thieltges *et al.*, 2013; Kuepper *et al.*, 2016) to limit the spectrum to the region of roughly 1800–900 cm^{-1} after preprocessing using physical models, which usually involve the complete spectrum.

2.3 Adjusting spectral background in tissue spectra

Scattering components in infrared spectra of biological samples were first reported in (Mohlenhoff *et al.*, 2005), which was later recognized as Mie scattering (Miljković *et al.*, 2012) and led to the first correction models (Kohler *et al.*, 2008). The explanation of the scattering

background signal was further improved by identifying it as a dispersion of the refractive index due to absorption (Bassan *et al.*, 2009). This model finally led to the now widely used correction procedure proposed in (Bassan *et al.*, 2010).

2.4 Essential traits of infrared spectra

Infrared spectra of biological samples are typically dominated by the amide I and amide II peak due to the high content in protein. As illustrated in Supplementary Figure S1, uncorrected raw spectra exhibit a high degree of variance across the whole spectrum. Only after eliminating resonant Mie scattering, the variance is largely reduced and those differences that are characteristic for the molecular composition and hence disease status become more pronounced. While these differences seem subtle at first sight, the numerous success stories of spectral histopathology show that they are highly significant and discriminative.

Raw spectra at first sight appear enigmatic and almost inaccessible to machine learning. On the one hand, mean spectra of different classes of tissue components that belong to different tissue components or undergo different sample preparation are highly correlated (Supplementary Fig. S7), while on the other hand the overwhelming degree of variance largely overshadows those rather subtle differences that one observes when comparing mean spectra.

The combination of subtle differences covered by high variance background signal explain the importance and the success of physical correction approaches, most notably the seminal work on resonant Mie separation (Bassan *et al.*, 2010). To put the present contribution into context, we investigate to what extent deep representation learning can untangle variances in a way that infrared spectra can be classified without preprocessing.

2.5 Representation learning and domain adaptation

Our work is motivated by groundbreaking progress in the field of representation learning surveyed in (Bengio *et al.*, 2013), which gained much of its momentum from related breakthroughs in the field of deep neural networks and convolutional neural networks (Rifai *et al.*, 2011). Specifically, we employ *stacked autoencoders*, a schematic example of which is depicted in Supplementary Figure S2. The process of training an autoencoder yields a lower dimensional representation of the input data in the last layer of the encoder. This representation can be thought of as an in some sense near-optimal non-linear embedding of the original data in a lower dimensional embedding. Note that the process of training an autoencoder also yields a decoder, i.e. the reverse mapping from the embedding to the original feature space. While the decoder rarely is of immediate practical use, applying the concatenation of the encoder with the decoder to an original data point will essentially reconstruct a very close approximation of the data point. The expectation towards the representation in the last layer of the encoder is that variances within the input data are being untangled according to features that have been obtained from the non-linear transformations that result from previous layers of the autoencoder.

2.6 Transfer learning

An essential question arising from the high accuracies often achieved by extremely parameter rich deep neural networks is whether the network is overfitting the training data rather than having identified truly discriminative features between the classes during supervised training. Observing strong validation measures in conventional cross-validation schemes is certainly a necessary, but not sufficient criterion. A stronger criterion will be to test how the classifiers perform on previously unseen types of data. In the domain of clinical data, one can identify several levels of what constitutes previously unseen. In Guo *et al.* (2017), it was suggested that validation should be performed at the highest possible level of replication in order to avoid overfitting. In clinical studies, different levels of replication are conceivable such as changes in the measurement device, changes in sample preparation, or even multi-center studies (Zech *et al.*, 2018).

The concept of robustness is closely related to the *transferability* of classification models: If a classifier generalizes when trained on one specific task, it should be feasible to further generalize classification towards a second task similar to the first task. In a related contribution (Guo *et al.*, 2018), it was shown that model transfer significantly improves classification of Raman microscopic images across four different microscopes. We investigate transferability in a somewhat broader setting: First, we train a deep neural network on formalin fixed paraffin embedded (FFPE, henceforth referred to as *embedded*) histopathological samples of colon tissue. Then, we investigate a transfer of this classifier to infrared images of fresh frozen tissue samples (henceforth referred to as *fresh tissue*). In related previous studies (Kuepper *et al.*, 2016), classifiers were built independently for embedded and fresh tissue, respectively, as model transfer appeared infeasible. In our contribution, we assess transfer learning approaches (Pan *et al.*, 2010) to facilitate the transfer of an FFPE trained neural network to fresh tissue.

3 Approach

Our computational approach is summarized in Figure 2. It is based on first obtaining a gold standard segmentation based on conventionally corrected spectra classified by a conventional, previously established classifier. The only purpose of this pre-segmentation is to obtain a sufficient amount of training data for the second *deep learning* stage. The deep learning stage is in turn divided into two steps: An unsupervised *pre-training* is succeeded by supervised fine-tuning into the final network **pt-MLP**.

As it has been demonstrated (Chen and Lin, 2014), these neural network based approaches largely benefit from the availability of large amounts of data. This sets apart our approach in a fundamental way from previous approaches, which rely on manual annotations. As there are inherent difficulties in obtaining suitable annotations on larger numbers of histopathological samples, conventional approaches are favorable towards classifiers that generalize well on small amounts of training data. In this sense, our present contribution aims to establish neural network based approaches that scale with the large amounts of data that are typically available in clinical studies involving infrared microscopy.

Beyond the accuracy of classifier **pt-MLP**, a key question to be assessed is whether **pt-MLP** generalizes well to unseen datasets, or whether it rather overfits the training data. In order to assess the capability of **pt-MLP** to generalize, we perform *transfer learning*. Specifically, we employ a second set of colon cancer related tissue samples. This set of tissue samples and the image spectra obtained from it differ substantially from the first set: first, this dataset has been acquired from fresh tissue rather than paraffin embedded tissue. Second, the samples were obtained as full sections rather than as tissue microarrays. The substantial differences in the image spectra are illustrated in Supplementary Figure S1. A gold standard pre-segmentation for producing training data was applied in a similar fashion as for the first dataset.

For convenience, Supplementary Table S1 provides an overview of the different supervised classifiers and their role in this study.

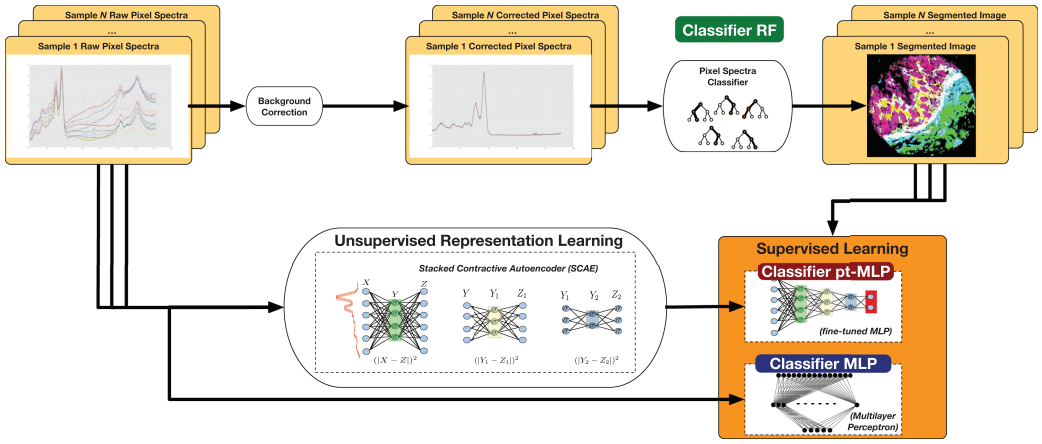

Fig. 2. Workflow for obtaining ground truth results, performing unsupervised pre-training and finally supervised fine-tuning as described in Section 3

4 Materials and methods

4.1 Sample material

We employed two datasets of infrared images of histopathological samples related to colon cancer that have been investigated in previous studies.

The first dataset from (Kuepper *et al.*, 2016) consists of infrared microscopic images of embedded tissue microarray (TMA) samples. We employed two such TMA slides purchased from *US Biomax Inc., MD, USA*, referred to by their IDs *CO1002b* and *CO722*, respectively, each of which consists of 100 circular spots of tissue sample from more than 60 different patients. Each spot has a diameter of roughly 1 mm. Along with this dataset, we also utilized the random forest classifier established previously in (Kuepper *et al.*, 2016), which classifies resonant Mie corrected infrared pixel spectra into 19 different classes representing 13 types of tissue components and subclasses, as depicted in Supplementary Figure S3. This random forest classifier is referred to as classifier **RF** in Figure 2 and throughout the rest of this manuscript. We will refer to this dataset of formalin-fixed and paraffin-embedded (FFPE) tissue microarray spots as the *FFPE dataset*. The FFPE dataset was available and utilized both in the form of uncorrected spectra and resonant Mie corrected spectra preprocessed by the approach from (Bassan *et al.*, 2010). The FFPE dataset was subdivided into three parts, one part for unsupervised pre-training, the second subset for supervised fine-tuning, and the third subset was withheld for validation. The second subset for supervised fine-tuning was further subdivided into a training set and a test set, so that training and test accuracy can be determined throughout the training process of supervised fine-tuning. The validation subset was strictly separated from the other two parts and was neither involved in pre-training nor in supervised learning.

Our second dataset was acquired from fresh-frozen histopathological colon tissue samples. Each sample is represented by one infrared image covering a whole tissue section roughly 2 cm^2 in size. The dataset involves seven such whole-slide images labeled as *Fresh 1–Fresh 7* involving seven different patients. For this dataset, we utilized a corresponding classifier **RF2** that was established previously. One of the seven samples (see Supplementary Table S2 for an overview) was partially used to recruit training data for transfer learning as described in Section 4. The training data for classifier **RF2** have been obtained fully independent of classifier **RF**. Throughout this manuscript, we will refer to this dataset as the *fresh tissue dataset*.

Following common practice (see Section 2), we limited the wavenumber range to the range between 1815 and 950 cm^{-1} so that each pixel spectrum is represented as a 450 dimensional vector. In both datasets, the FFPE as well as the fresh dataset, pixel spectra with low signal intensity were filtered out and marked as background based on previously described practice (Kallenbach-Thieltges *et al.*, 2013; Kuepper *et al.*, 2016). The filtering is performed on uncorrected raw spectra, so that it does not affect our approach being independent from the physical model based resonant Mie correction.

The datasets were divided into a pre-training set, a training set for supervised finetuning and validation dataset. The pre-training set for training the autoencoders involves 2.2 million spectra from the FFPE dataset *CO722* covering spectra from 25 tissue microarray spots from 25 spots from 10 different patients. For the finetuning set, a subregion of 1.3 million spectra (20 spots, 16 different patients) from dataset *CO1002b* was selected. Validation was conducted on 3.51 million spectra from 24 spots (22 different patients) from dataset *CO1002b*.

4.2 Obtaining ground truth segmentations

In order to obtain uncorrected spectra with labels for supervised training, we applied classifier **RF** which was previously established in (Kuepper *et al.*, 2016) to all resonant Mie corrected spectra from the FFPE dataset. Mie correction was performed on the wavenumber region 2300–950 cm^{-1} using the approach from (Bassan *et al.*, 2010) using one iteration. Since the uncorrected counterpart is immediately available for each corrected pixel spectrum, this allowed us to assign a class to each uncorrected spectrum. As illustrated in Figure 2, we use the resulting assignment between uncorrected spectra and tissue components as ground truth for the training dataset for our deep neural networks. Note that the classification outcome of **RF** cannot be assumed to be 100% correct on a per-pixel basis and the assignment thus obtained constitutes a *gold standard* in the sense of the best-possible per-pixel annotation rather than a ground truth. Despite this somewhat curtailing factor, we will refer to the training labels obtained from **RF** as ground truth.

4.3 Learning regularized representations through autoencoders

Formally, an autoencoder is constituted by a neural network that represents a mapping $A : \mathbb{R}^d \to \mathbb{R}^d$, i.e. a network whose input and output layers consist of d neurons each. In its most basic form, an autoencoder involves one hidden layer with $M < d$ neurons. A sequence of $(d, M_1), (M_1, M_2), \ldots, (M_{K-1}, M_K)$ autoencoders can be cascaded in a straightforward manner as illustrated in Supplementary Figure S2 and

detailed in Supplementary Section S.1. In (Vincent *et al.*, 2010), such *stacked autoencoders* have been proposed and successfully established as highly effective regularizers on several datasets.

Our stacked autoencoder preprocesses spectra represented as a vector featuring absorbances at $d = 450$ wavenumbers. The stacked autoencoder involved six hidden layers of sizes $M_1, \ldots, M_6 = 450, 900, 450, 100, 100, 100$. The stacked autoencoder was trained in an unsupervised fashion following (Vincent *et al.*, 2008) on 2 220 000 spectra obtained from 25 spots of TMA slide *CO1002b*. Each autoencoder was initialized following (Glorot and Bengio, 2010). For training autoencoders, mean squared error with an additive regularization term was used as loss function, as detailed in Supplementary Section S.1.

4.4 Supervised finetuning for classification of pixel spectra

We followed the approach by Rifai *et al.* (2011) and employed the autoencoder described in Section 4 as an unsupervised pretraining procedure to improve a subsequent supervised learning step. To this end, a *softmax* output layer with one output neuron for each of the 19 classes was added to the six encoding layers of the stacked autoencoder. The resulting network topology was initialized with the parameters of the stacked autoencoder for the first six layers, and random values for the input weights of the output layer. The last three hidden layers were treated as drop out layers (Srivastava *et al.*, 2014) with a drop out rate of 50%. This network was trained on the ground truth provided by classifier **RF** as shown in Figure 2 to obtain supervised classifier **pt-MLP** using *RMSProp* optimization running for 15 000 epochs using categorical cross entropy as a loss function.

As a reference to assess the performance of **pt-MLP**, we trained a conventional multilayer perceptron **MLP** based on the same topology as network **pt-MLP**. As in classifier **pt-MLP**, the last three hidden layers were implemented as drop out layers with a dropout rate of 50%. We initialized all parameters in the network randomly and trained it against the same ground truth using *RMSprop* for optimization running 15 000 iterations, also using categorical cross entropy as a loss function.

4.5 Transfer learning for domain adaptation

In order to assess the generalization capability of the unsupervised pretraining procedure and the resulting classifier **pt-MLP**, we performed transfer learning from FFPE to fresh tissue samples. Ground truth on fresh tissue for transfer learning was obtained from a previously established classifier **RF2**, which was trained on resonant Mie corrected spectra in fresh tissue. To adapt to the different classes of tissue components annotated in FFPE versus fresh tissue (see Supplementary Fig. S3), the output layer was substituted by a randomly initialized softmax layer. The transfer learning approach is illustrated in Supplementary Figure S5.

We divided dataset *Fresh 1* into a training dataset and a test dataset for transfer learning, and performed 15 000 epochs of *RMSProp* training. As the purpose of training **tl-MLP** is to demonstrate the generalization capability of a representation learned by **pt-MLP**, the region used for training was explicitly chosen to be small and with limited variability, so that we chose only 1.3 Million spectra (corresponding to roughly 5×5 mm of sample) from only one single sample.

Datasets *Fresh 2*–*Fresh 7* were used for validation. Determining accuracies for validation neglects spectra that were masked out as background in the ground truth annotation, as detailed in Supplementary Section S.2.

5 Results

5.1 Pretraining with SCAE and supervised finetuning

We performed pretraining as described in Section 4 on 2.2 million spectra from the FFPE dataset *CO722*. The deep learning classifier **pt-MLP** was obtained by finetuning as described in Section 4 on 1.3 million spectra from dataset *CO1002b*.

Figure 3 demonstrates the generalization capability on a held-back TMA dataset. The per-pixel accuracy of classifier **pt-MLP** reconstructing the ground truth segmentation of classifier **RF** achieved a validation accuracy of 96% (83% after not counting the highly abundant background class, see Supplementary Figs S10–S12), while the test accuracy during training was 93%. The unexpected gain in accuracy between test and validation dataset may be explained by the heterogeneous sample quality of the TMA samples. Switching the pretraining from a stacked contractive autoencoder to a plain stacked autoencoder by dropping the regularizing term of Frobenius norm of the Jacobi matrix during training, the validation accuracy dropped slightly to 95%, while the training accuracy dropped to 90%. In other words, using the gap between the accuracies of training set and validation as an indicator of the generalization error, the stacked contractive autoencoder achieves a lower generalization error than the plain stacked autoencoder.

The accuracy of **pt-MLP** compares to a slightly lower accuracy of 94% when training a network with the same topology on resonant Mie corrected data using backpropagation in the same manner as **MLP** was trained on uncorrected data. For the accuracies of 96

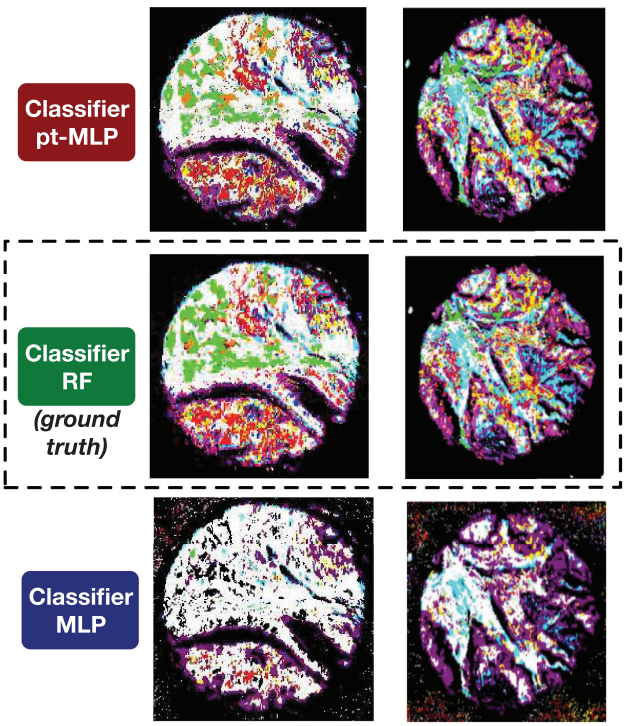

Fig. 3. Classification of FTIR raw data of two FFPE-embedded Tissue Microarray spots (TMA) from the fully independent validation dataset. The agreement between the pre-trained **pt-MLP** (first row) and the ground truth provided by **RF** (middle row) is very high, and differences recognizable only in small details. The non-pretrained classifier MLP, on the other hand, exhibits systematic misclassification, which is most notably of crypts in the tumor-free spot (right, index color pink) and the tumor and the submucosa class (red and green), which are systematically misclassified as muscle (white) in the tumor spot (left). Classification results of further spots are displayed in Supplementary Figure S4

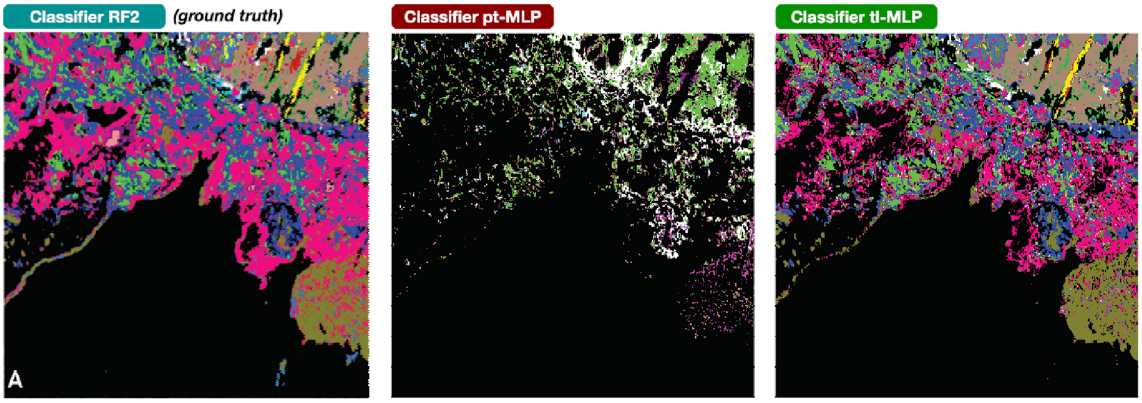

Fig. 4. Classification of FTIR raw data with and without transfer learning on independent validation dataset *Fresh 2*. *Left:* ground truth obtained from classifier **RF2** on Mie-scattering corrected FTIR-microspectroscopy imaging; *Middle:* prediction obtained from the FFPE-based classifier **pt-MLP**, which fails to identify most of the tissue components in fresh tissue and achieves an accuracy of only 53%. *Right:* prediction of the transfer learned deep learning classifier **tl-MLP**. Results for dataset *Fresh 3* are shown in Supplementary Figure S6

and 94%, it must be kept in mind that the ground truth of classifier **RF** will not be perfect. To assess this further in a qualitative manner, we inspected the segmentations of a spot that according to annotation is free of cancer, as shown in Supplementary Figure S9. As it turns out, **pt-MLP** recognizes less false-positive tumor positions in this spot than the ground-truth classifier **RF**. This indicates that even higher accuracies will not be plausible without overfitting to imperfect ground truth.

5.2 Transfer-learning segmentations of fresh tissue

To assess the transferability of classifiers from FFPE tissue to fresh tissue, we first visualized spectral differences between FFPE tissue and fresh tissue, which is illustrated for two classes of tissue components in Supplementary Figure S1. In particular at the level of uncorrected raw spectra, these differences are substantial, so that applying classifiers trained on FFPE spectra to fresh tissue can be expected to lead to very limited success. To examine this in practice, we applied the FFPE-trained classifier **pt-MLP** to fresh tissue. As ground truth, we used the segmentation obtained from a previously established random forest classifier **RF2**. As shown in Figure 4, **pt-MLP** performs poorly in identifying the tissue structure, achieving an accuracy of only 53%. Two classes, namely submucosa and muscle, were at least partially detected correctly by **pt-MLP**.

Finally, we performed transfer learning from FFPE tissue spectra to fresh tissue spectra by training classifier **tl-MLP** using **pt-MLP** as a starting point. As training data for transfer learning, we used uncorrected fresh tissue spectra labeled with the output classifier **RF2** as ground truth, as illustrated in Supplementary Figure S5. A validation of the result is shown in Figure 4 and Supplementary Figure S6. Accuracies of validating the transfer learned classifier **tl-MLP** on five whole slide images (Supplementary Fig. S13) range between 40 and 76%. For four out of the five validation spots, overall tissue structure is reconstructed by **tl-MLP**, although in some cases with systematic bias. Yet, considering the deliberately low amount of training data and the fact the accuracies are pixel accuracies in images, it is reasonable to claim that the overall tissue structure is reconstructed in most cases.

6 Discussion

In our study, we demonstrated that unsupervised pre-training of infrared spectra facilitates highly accurate classification of spectral histopathology imaging data. Unsupervised pre-training takes the role of spectral pre-processing, which previously has been tackled on the grounds of physical models in combination with conventional classifiers. Thus, we have demonstrated that a purely data-driven approach can take the role that has previously been taken by a physical model when classifying pixel spectra.

A natural question that arises in this context, and in neural networks in general, is to characterize and interpret what the neural network actually learned. Our results allow the conclusion that the network *implicitely* identified variances that are eliminated by correction procedure underlying the resonant Mie model by Bassan *et al.* (2010), because classifier **pt-MLP** can reproduce the classification of resonant Mie corrected spectra. This is remarkable since this implies that the network has learned to disentangle the complex interference between molecular spectrum and the scattering background signal, which is neither additive nor linear. It will be worthwhile to assess our approach in comparison to recent improvements of resonant Mie correction Konevskikh *et al.* (2018).

Since the resonant Mie model is neither explicitly nor implicitly involved in the procedure of training the network, the question arises whether the network may have learned a much more general representation of infrared pixel spectra from histopathological samples. While beyond the scope of our current contribution, this question of model interpretation points to an interesting and relevant new direction for future research, namely to correlate the output of the stacked autoencoder with different physical model-based correction procedures. More specifically, one could investigate how well a spectrum corrected by a physical model can be reconstructed from the representation learned by the neural network. This could be realized by establishing a neural network that learns to approximate a given correction procedure, using the representation learned by the unsupervised pre-training as a starting point. If such a network could reconstruct the result of given physical model-based preprocessing procedure, one would obtain a more *explicit* proof that the network has learned a certain preprocessing function.

In general, the features learned by the stacked autoencoder and the subsequent finetuned classifier remain a black box which calls for being understood better. This *model interpretation* problem is particularly relevant to understand whether the classifier uses information rather equivalent to the information used in resonant Mie corrected spectra, or whether scattering information is being used that has been removed in resonant Mie corrected spectra. However,

this will require suitable approaches for model interpretation. While an abundant list of approaches such as *saliency maps* (Simonyan *et al.*, 2013) or *class activation maps* (Zhou *et al.*, 2016) have been proposed for CNNs, they rely on topological features of CNNs such as *max pooling* layers which are not applicable to the fully connected networks presented in this work, so that the interpretation of our trained networks remains an open issue. To at least gain some basic insight into the learned models, Supplementary Figure S8 visualizes the weight matrices at the six hidden layers of the trained networks, indicating a strong effect of both the finetuning and the transfer learning on the weights in the network.

As we have argued, there is strong evidence that the stacked autoencoder did learn a meaningful representation of infrared pixel spectra. An obvious next question is how far this representation will *generalize*: how much variance can be added and what type of variance can be added before networks derived from the representation will lose their classification capability. Our approach to determine generalization capability was to investigate the capability of the network to perform *domain adaptation*: If the network can adapt from the domain of embedded tissue to the domain fresh tissue, the underlying representation must be sufficiently abstract and thus generalizable. The high accuracies we observe in the transfer learned networks for fresh tissue strongly indicate that indeed the representation is sufficiently general. It is quite remarkable that despite the moderate effort used for transfer learning—we used the same number of epochs for transfer learning **tl-MLP** as for fine-tuning the original network **pt-MLP** while the last hidden layer of **tl-MLP** had to be fully re-initialized due to the differences in ground truth classes—we obtain very high accuracies.

As the generalization capability is strong enough to adapt an FFPE classifier to fresh tissue, the question emerges how strongly the network will generalize. There is a broad bandwidth of conceivable sources of variance over which it is desirable to obtain a general spectral representation: Besides the variation in tissue type (FFPE versus fresh) investigated here, there may be variation in the type of microscope (FTIR versus QCL), variation in the substrate, or variation in the organ of origin (colon versus tissue form other organs), to mention only a few. A key question to be addressed will be which variation needs to be included in the data for pre-training the stacked autoencoder, and to which variation will the resulting network be capable of generalizing. As we have shown in our work, not all variation needs to be reflected in the pre-training data: Although we did not use any fresh tissue for pre-training, the network turned out to be transferable to fresh tissue anyway. An ultimate goal may be to train an universal preprocessing network that generalizes broadly across the aforementioned sources of variance. We limit our claims of the pre-training performed in the present study for the network to generalize across tissue type, i.e. across FFPE and fresh tissue. This constitutes a major progress over the previously employed physical model based preprocessing, whose generalization capabilities are inherently limited if present at all.

Finally, generalization capability of classifiers against different sources of variance certainly is a key aspect for robust classifiers in life science research in general, and biomarker discovery in particular. As our approach to combine representation learning with domain adaptation involves no assumptions specific to infrared microscopy, this observation bears promise well beyond infrared microscopy. In fact, baseline correction or other forms of preprocessing commonly constitute problems in the analysis of various types of bioanalytical data, ranging from NMR spectroscopy (Xi and Rocke, 2008), mass spectrometry (Du *et al.*, 2006) or Raman spectroscopy (Zhang *et al.*, 2010). While several different approaches have been proposed for these techniques, it has been commonly observed that preprocessing, in some cases heavily, affects subsequent analysis (Du *et al.*, 2006). On the other hand, the baseline artifacts in other bioanalytical spectra tend to be less complex than resonant Mie scattering in infrared spectra, and it thus appears reasonable to assume that other types of bioanalytical data can greatly benefit from our purely data-driven unsupervised preprocessing approach using stacked autoencoders.

7 Conclusion

We have tackled and successfully solved two related problems in spectral histopathology that have not been studied previously, likely because they could not be solved with conventional classifiers: First, we demonstrated that unsupervised pre-training allows to train classifiers that can classify unprocessed raw pixel spectra of infrared microscopic images, and thus may substitute the physical model based preprocessing of infrared image spectra. At the same time, these classifiers are well-established regularizers and thus hold the promise of generalizing stronger and having less tendency towards overfitting. Second, we have introduced the concept of transferring across domains of different tissue preparation as a method to assess whether a classifiers generalizes well or will rather tend to overfit the given training data.

Acknowledgement

We want to thank Angela Kallenbach-Thieltges for providing the RF classifier for FFPE tissue thin sections. We thank an anonymous referee on helpful remarks regarding resonant Mie theory.

Funding

This research was supported by the Protein Research Unit Ruhr within Europe (PURE) funded by the Ministry of Innovation, Science and Research (MIWF) of North-Rhine Westphalia, Germany (grant number: 233-1.08.03.03-031-68079).

Conflict of Interest: none declared.

References

Bassan,P. *et al.* (2009) Resonant Mie scattering in infrared spectroscopy of biological materials–understanding the 'dispersion artefact'. *Analyst*, **134**, 1586–1593.

Bassan,P. *et al.* (2010) Resonant Mie scattering (RMieS) correction of infrared spectra from highly scattering biological samples. *Analyst*, **135**, 268–277.

Bengio,Y. *et al.* (2013) Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, **35**, 1798–1828.

Bird,B. *et al.* (2012) Infrared spectral histopathology (SHP): a novel diagnostic tool for the accurate classification of lung cancer. *Lab. Investig.*, **92**, 1358.

Chen,X.-W. and Lin,X. (2014) Big data deep learning: challenges and perspectives. *IEEE Access*, **2**, 514–525.

Du,P. *et al.* (2006) Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, **22**, 2059–2065.

Glorot,X. and Bengio,Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256.

Großerueschkamp,F. *et al.* (2015) Marker-free automated histopathological annotation of lung tumour subtypes by FTIR imaging. *Analyst*, **140**, 2114–2120.

Großerueschkamp,F. *et al.* (2017) Spatial and molecular resolution of diffuse malignant mesothelioma heterogeneity by integrating label-free FTIR imaging, laser capture microdissection and proteomics. *Sci. Rep.*, **7**, 44829.

Guo,S. *et al.* (2017) Common mistakes in cross-validating classification models. *Anal. Methods*, **9**, 4410–4417.

Guo,S. *et al*. (2018) Extended multiplicative signal correction based model transfer for Raman spectroscopy in biological applications. *Anal. Chem*., **90**, 9787–9795.

Kallenbach-Thieltges,A. *et al*. (2013) Immunohistochemistry, histopathology and infrared spectral histopathology of colon cancer tissue sections. *J. Biophotonics*, **6**, 88–100.

Kohler,A. *et al*. (2008) Estimating and correcting Mie scattering in synchrotron-based microscopic Fourier transform infrared spectra by extended multiplicative signal correction. *Appl. Spectroscopy*, **62**, 259–266.

Konevskikh,T. *et al*. (2018) An improved algorithm for fast resonant Mie scatter correction of infrared spectra of cells and tissues. *J. Biophotonics*, **11**, e201600307.

Kuepper,C. *et al*. (2016) Label-free classification of colon cancer grading using infrared spectral histopathology. *Faraday Discuss*., **187**, 105–118.

Kuepper,C. *et al*. (2018) Quantum cascade laser-based infrared microscopy for label-free and automated cancer classification in tissue sections. *Sci. Rep*., **8**, 7717.

Marcsisin,E.J. *et al*. (2012) Noise adjusted principal component reconstruction to optimize infrared microspectroscopy of individual live cells. *Analyst*, **137**, 2958–2964.

Miljković,M. *et al*. (2012) Line shape distortion effects in infrared spectroscopy. *Analyst*, **137**, 3954–3964.

Mohlenhoff,B. *et al*. (2005) Mie-type scattering and non-beer-lambert absorption behavior of human cells in infrared microspectroscopy. *Biophys. J*., **88**, 3635–3640.

Pan,S.J. *et al*. (2010) A survey on transfer learning. *IEEE Trans. Knowl. Data Eng*., **22**, 1345–1359.

Rifai,S. *et al*. (2011). Contractive auto-encoders: explicit invariance during feature extraction. In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*. Omnipress, pp. 833–840.

Romeo,M. and Diem,M. (2005) Correction of dispersive line shape artifact observed in diffuse reflection infrared spectroscopy and absorption/reflection (transflection) infrared micro-spectroscopy. *Vibration. Spectroscopy*, **38**, 129–132.

Simonyan,K. *et al*. (2013) Deep inside convolutional networks: visualising image classification models and saliency maps. *arXiv preprint arXiv: 1312.6034*.

Srivastava,N. *et al*. (2014) Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res*., **15**, 1929–1958.

Vincent,P. *et al*. (2008). Extracting and composing robust features with denoising autoencoders. In: *Proceedings of the 25th International Conference on Machine Learning*. ACM, pp. 1096–1103.

Vincent,P. *et al*. (2010) Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res*., **11**, 3371–3408.

Xi,Y. and Rocke,D.M. (2008) Baseline correction for NMR spectroscopic metabolomics data analysis. *BMC Bioinformatics*, **9**, 324.

Yosef,H.K. *et al*. (2017) Noninvasive diagnosis of high-grade urothelial carcinoma in urine by Raman spectral imaging. *Anal. Chem*., **89**, 6893–6899.

Zech,J.R. *et al*. (2018) Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med*., **15**, e1002683.

Zhang,Z.-M. *et al*. (2010) Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst*, **135**, 1138–1146.

Zhou,B. *et al*. (2016) Learning deep features for discriminative localization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929.