Bioinformatics, 36(10), 2020, 3215–3224
doi: 10.1093/bioinformatics/btaa085
Advance Access Publication Date: 12 February 2020
Original Paper

Bioimage informatics

ColocML: machine learning quantifies co-localization between mass spectrometry images

Katja Ovchinnikova[1], **Lachlan Stuart**[1,†], **Alexander Rakhlin**[2,†], **Sergey Nikolenko**[3,4] **and Theodore Alexandrov**[1,5,6,*]

[1]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany, [2]Neuromation OU, Tallinn, Estonia, [3]National Research Institute Higher School of Economics, [4]Steklov Institute of Mathematics at St. Petersburg, St. Petersburg, Russia, [5]Metabolomics Core Facility, European Molecular Biology Laboratory, Heidelberg, Germany and [6]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA

*To whom correspondence should be addressed.

†The authors wish it to be known that these authors contributed equally.

Associate Editor: Alfonso Valencia

Received on September 4, 2019; revised on January 22, 2020; editorial decision on February 2, 2020; accepted on February 4, 2020

Abstract

Motivation: Imaging mass spectrometry (imaging MS) is a prominent technique for capturing distributions of molecules in tissue sections. Various computational methods for imaging MS rely on quantifying spatial correlations between ion images, referred to as co-localization. However, no comprehensive evaluation of co-localization measures has ever been performed; this leads to arbitrary choices and hinders method development.

Results: We present ColocML, a machine learning approach addressing this gap. With the help of 42 imaging MS experts from nine laboratories, we created a gold standard of 2210 pairs of ion images ranked by their co-localization. We evaluated existing co-localization measures and developed novel measures using term frequency–inverse document frequency and deep neural networks. The semi-supervised deep learning Pi model and the cosine score applied after median thresholding performed the best (Spearman 0.797 and 0.794 with expert rankings, respectively). We illustrate these measures by inferring co-localization properties of 10 273 molecules from 3685 public METASPACE datasets.

Availability and implementation: https://github.com/metaspace2020/coloc.

Contact: theodore.alexandrov@embl.de

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

1 Introduction

Metabolites and lipids play key roles in fuelling and making up cells, ultimately determining their types and states. Concentrations of metabolites and lipids are carefully regulated to maintain homeostasis in tissues, organs and organisms, and are profoundly and sometimes irreversibly altered in disease. Capturing spatial distributions of molecules in tissue sections is a prerequisite for any hypothesis-driven or discovery-oriented investigation of biology and medicine on the levels of tissues and the organism. In the past two decades, a window of opportunity has been opened by the development and further maturation of imaging mass spectrometry (imaging MS), a powerful and versatile technology for spatial molecular analysis (Buchberger *et al.*, 2018; Doerr, 2018; Dreisewerd and Yew, 2017) with a particular interest in clinical (Vaysse *et al.*, 2017) and pharmaceutical applications (Schulz *et al.*, 2019). For a tissue section, imaging MS generates a hyperspectral image encompassing thousands to millions of ion images, each image representing the distribution of a particular molecule or several molecules in the section. Rapid development and growing popularity of imaging MS, as well as the high dimensionality and sheer size of generated data, measuring up to hundreds of gigabytes for a tissue section, have stimulated the development of computational methods and software (Alexandrov, 2012). Various methods have been developed for low-dimensional data representation (based on PCA, NMF, t-SNE, biclustering), finding spatial regions of interest with spatial segmentation, search for markers associated with a region of interest, and, recently, for metabolite annotation (Palmer *et al.*, 2017). Many of these methods use some measure of spatial similarity between ion images, often referred to as *spatial co-localization*. Various measures for quantifying co-localization have been proposed, including the Pearson correlation, cosine score, Euclidean L_2-measures (Alexandrov, 2012; Alexandrov *et al.*, 2010; McCombie *et al.*, 2005; McDonnell *et al.*, 2008) sometimes applied to transformed images, e.g. after hotspot removal or log-transformation (Watrous *et al.*, 2011). Recently, new measures adopted from other fields have been proposed, including the structural similarity index (SSIM) and hypergeometric similarity measure (Aaron *et al.*, 2018; Ekelöf *et al.*, 2018; Kaddi *et al.*, 2011). However, despite the ubiquity of using spatial co-localization in imaging MS and a variety of measures proposed, no rigorous and comprehensive evaluation of co-localization measures has ever been performed.

This leads to arbitrary and often *ad hoc* choice of a co-localization measure in every particular study, laboratory or software package. Moreover, it hinders the progress of imaging MS methods since new co-localization measures are faced with scepticism without objective criteria to demonstrate their advantages. This gap has persisted for over a decade due to the lack of ground truth data that would allow one to evaluate a measure objectively. Obtaining ground truth data is challenging because it requires a comprehensive inventory of which molecules are represented in imaging MS data and which of them are co-localized. This is not possible for tissues and hardly possible even for authentic standards due to our limited understanding of ionization of complex mixtures.

Here, we are addressing this apparent gap by presenting ColocML, a machine learning approach to quantify co-localization between ion images. First, we present a gold standard set of pairs of ion images ranked by imaging MS experts by the perceived co-localization. This effort was enabled by METASPACE, the open knowledge base of spatial metabolomes (Alexandrov *et al.*, 2019), through being able to select a large number of public representative datasets, employ modern web-based technologies for user-friendly and facilitated image ranking, engage a large number of experts and consolidate their rankings into a high-quality gold standard set. Second, using the gold standard set of pairs of images manually ranked by their co-localization, we have evaluated a variety of co-localization measures, including the cosine score, Pearson correlation and SSIM. Moreover, we propose several novel measures for co-localization, e.g. using term frequency–inverse document frequency (tf–idf) adopted from natural language processing as well as approaches based on deep learning.

We found the semi-supervised deep learning-based Pi model as well as the cosine score applied after median thresholding to be the most optimal spatial co-localization measures for imaging MS. We propose to use them in data analysis methods relying on co-localization. Our work provides a gold standard set [available at GitHub (https://github.com/metaspace2020/coloc)] which can be used for evaluating future measures, and illustrates how using open-access data, web technologies, community engagement and deep learning open novel avenues to addressing long-standing challenges in imaging MS.

2 Materials and methods

2.1 Experiment design to collect expert knowledge

In artificial intelligence and computer vision, a gold standard set is a collection of images manually tagged or ranked by experts called rankers. Having a gold standard set enables training and evaluation of machine learning models and algorithms. However, creating an unbiased, representative and high-quality gold standard set is a substantial challenge on its own. To the best of our knowledge, there exists no gold standard set of co-localized images for imaging MS. We aimed at creating a gold standard set that would quantify the perceived by experts degree of co-localization for different ions. We designed the gold standard set to consist of *target-comparison sets* where each set includes one *target ion* and 10 *comparison ions* ranked according to their co-localization with the target ion.

To create a gold standard set of co-localized ion images, we selected public datasets from METASPACE with the aim to have a manageable number of diverse yet representative high-quality datasets from different laboratories. First, we selected laboratories with at least three active contributors of public data, nine laboratories in total. For every laboratory, we selected active contributors to METASPACE, 42 rankers in total. We aimed at asking each ranker to rank up to 20 sets and at every set to be ranked by three rankers. For each laboratory, we randomly selected round (20 × N_TL × 2/3) public datasets submitted by this laboratory to METASPACE, where N_TL is the number of rankers from a given laboratory.

From each dataset, we randomly selected 1 target ion and 10 comparison ions constituting a target-comparison set. We then used the RankColoc web app (described later) to go through the target-comparison sets and exclude noisy images or images with only a few pixels. For each laboratory, we aimed at obtaining round (20 × N_TL/3) high-quality sets, although it was not always possible due to the quality of the datasets. This allowed us to have the same target-comparison set ranked by three rankers to later estimate the between-rankers agreement and to obtain average ranks.

2.2 Pilot study

Before creating the gold standard set, we ran a pilot study to investigate the difficulty of ranking ion images in the target-comparison sets according to their co-localization, as well as to learn potential pitfalls and obstacles of the ranking process. The pilot study was basically a full study including dataset selection, web app implementation, rankers recruitment, gold standard set creation and agreement evaluation, but performed in a smaller format with five rankers only.

2.3 Web app for manual ranking of ion images

The RankColoc web app (https://github.com/metaspace2020/coloc/tree/master/RankColoc) was developed with the aim to facilitate image ranking as well as help inspect ranked sets. For a public dataset in METASPACE, the web app downloads ion images from METASPACE using the GraphQL API (https://github.com/metaspace2020/metaspace/tree/master/metaspace/python-client), and shows a target and 10 comparison ion images. The web app helps a ranker rank each comparison image from 0 to 9 by dragging and dropping it into 1 of the 10 rank boxes or leave it unranked. Several images can be assigned the same rank. The web app page includes instructions for rankers. For each ranker, we assigned a collection of target-comparison sets and generated unique URLs containing the sets and the ranker ID. Ranking results were stored in real time, associated with the ranker ID, and could be opened by either the same ranker or a curator. Figure 1 shows screenshots of RankColoc web app with examples of the ranked sets. Supplementary Video S1 illustrates the ranking process.

2.4 Evaluating obtained rankings

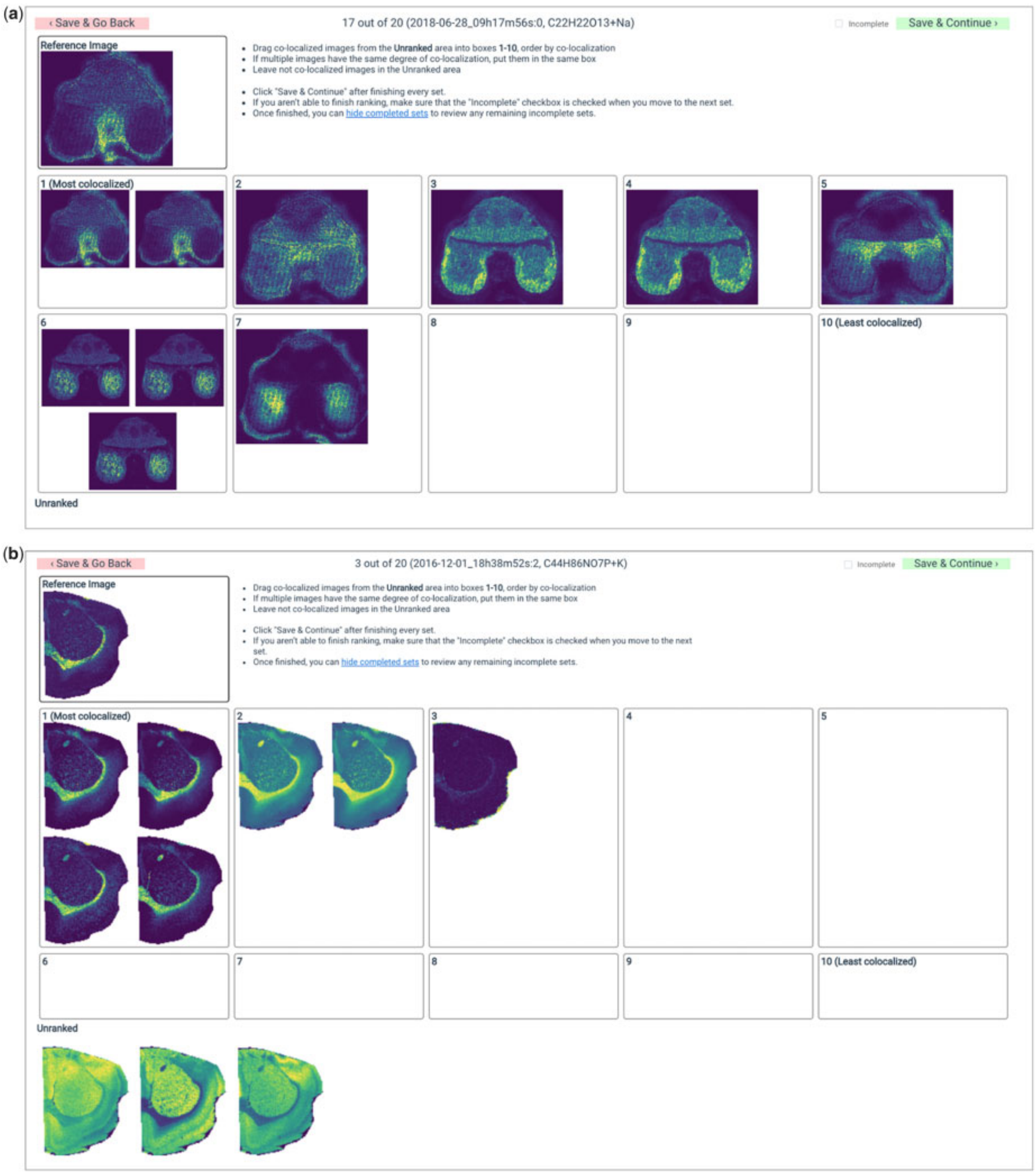
We assessed the complexity of the task and reproducibility of the rankers' judgements by calculating pairwise correlations between the rankers, i.e. correlations between their ranks of comparison ion images in the same sets. The images left unranked (i.e. perceived by rankers as completely not co-localized with the target ion image) were assigned the rank 10. We computed average Spearman and Kendall ranker-pairwise rank correlation for each laboratory, ranker and set.

2.5 Creating the gold standard set

To ensure the high quality of the resulting gold standard set, we have excluded (i) sets for which the average Spearman pairwise correlation between rankers was <0.4, and (ii) rankers whose average Spearman correlation with other rankers was <0.4. After some rankers were excluded, some sets ended up with just one ranking. We excluded those sets as well. The resulting gold standard set contains pairs of ion images (target image and a comparison image) with each pair assigned an average rank across three rankers. The ranks range from 0, representing the highest co-localization, to 10, representing the lowest or no co-localization.

2.6 Co-localization measures that require no learning

Our implementation of the co-localization measures is available at the GitHub repository (https://github.com/metaspace2020/coloc/tree/master/measures).



- Drag co-localized images from the **Unranked** area into boxes **1-10**, order by co-localization
- If multiple images have the same degree of co-localization, put them in the same box
- Leave not co-localized images in the Unranked area

- Click "Save & Continue" after finishing every set.
- If you aren't able to finish ranking, make sure that the "Incomplete" checkbox is checked when you move to the next set.
- Once finished, you can hide completed sets to review any remaining incomplete sets.

- Drag co-localized images from the **Unranked** area into boxes **1-10**, order by co-localization
- If multiple images have the same degree of co-localization, put them in the same box
- Leave not co-localized images in the Unranked area

- Click "Save & Continue" after finishing every set.
- If you aren't able to finish ranking, make sure that the "Incomplete" checkbox is checked when you move to the next set.
- Once finished, you can hide completed sets to review any remaining incomplete sets.

Fig. 1. Screenshots of the RankColoc web app showing two target-comparison sets ranked by experts. (**a**) MALDI-imaging dataset from a wheat seed section, submitted to METASPACE by Dhaka Bhandari, Justus Liebig University Giessen. (**b**) MALDI-imaging dataset from a rat brain tissue section, submitted to METASPACE by Berin Boughton, University of Melbourne (METASPACE URLs, for example datasets: https://metaspace2020.eu/annotations?ds=2018-06-28_09h17m56s, https://metaspace2020.eu/annotations?ds=2016-12-01_18h38m52s)

2.6.1 Correlation and cosine-based measures

First, we considered the commonly used co-localization measures: Pearson correlation, Spearman correlation and cosine similarity applied to flattened ion images, i.e. 1D vectors of pixel intensities.

2.6.2 SSIM measure

Following (Ekelöf *et al.*, 2018), we considered the SSIM (Wang *et al.*, 2004) with the Gaussian weights.

2.6.3 tf–idf-Based measure

We developed a measure based on the tf–idf concept from the field of natural language processing (Leskovec *et al.*, 2014). Using flattened ion images, we calculated the tf–idf value for each pixel–ion pair to quantify how important a pixel p is for the particular ion i with respect to all ions in the dataset D:

$$\mathrm{tf}-\mathrm{idf}(p,i,D)=\mathrm{tf}(p,i)\ *\ \mathrm{idf}(p,D),$$

$$\mathrm{tf}(p,i)=\mathrm{int}(p,i)/\sum_{p'\in P_D}\mathrm{int}(p',i)$$

$$\mathrm{idf}(p,D)=\log(|I_D|/|\{i\in I_D:\mathrm{int}(p,i)>0\}|),$$

where P_D is the set of all pixels in D, I_D is the set of all ions in D, and $\mathrm{int}(p,i)$ is the intensity of i in p. We then created tf–idf vectors of the same dimensionality as the intensity vectors and quantified co-localization of ion images as the cosine similarity between the corresponding tf–idf vectors.

2.6.4 Image transformations

For all considered ion intensity-based measures, we applied the following transformations to the ion image prior to calculating co-localization: (i) hotspot removal, namely reducing intensities of pixels with intensities >0.99 quantile by replacing them with the 0.99 quantile value; (ii) denoising by applying the median filter (Huang *et al.*, 1979) with a square window of size ranging from 1 (no filter applied) to 5 with step 1; (iii) applying quantile thresholding, namely setting to zero those pixel intensities which are below a given quantile value, for quantiles ranging from 0 to 0.9 with step 0.05. Evaluation whether using a transformation is beneficial as well as optimizing the size of the median filter and the quantile value was performed using the 5-fold cross-validation for each measure. Measures with the best performing filters were then applied to the entire gold standard set.

2.7 Co-localization measures based on deep learning

Our implementation of the co-localization measures is available at the GitHub repository (https://github.com/metaspace2020/coloc/tree/master/measures).

With the advent of deep learning, models based on neural networks have become the method of choice for processing unstructured data such as images. Therefore, in our study we have developed several methods exploiting current state-of-the-art deep learning approaches that would learn ion co-localization from the gold standard set.

2.7.1 Xception-based model

This model, illustrated in Figure 2, is based on the well-known Xception convolutional architecture designed to extract informative features from images (Chollet, 2017). We introduced the following modifications. First, the input has two channels corresponding to the target and comparison ion images. The two channels pass through the Xception architecture without the final classification layer, which in our case is replaced with a regression output. The Xception-based model is supervised, and its target variable is the rank as specified in the gold standard set, with the mean squared error (MSE) loss function.

2.7.2 Mu model

The Mu model is a variation of the Xception-based model with the difference that the top layers are replaced with two 2048-dimensional outputs followed by a discriminator. The mu model encodes a pair of ion images into two 2048-dimensional representations, computed image similarity as the Pearson correlation coefficient between the representations, and then regresses the similarity score onto the rank target with the MSE loss.

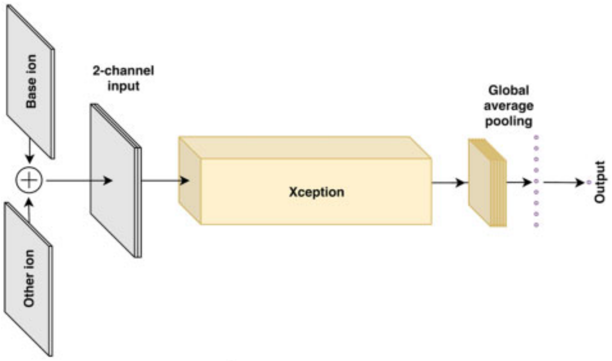


Fig. 2. Architecture of the Xception-based deep learning model

2.7.3 Unsupervised UMAP

We developed a model based on the uniform manifold approximation and projection (UMAP), a recently developed non-linear dimensionality reduction technique with broad applications in biology (McInnes *et al.*, 2018). In this model, we applied UMAP to embed flattened ion images (i.e. 1D vectors of pixel intensities) into 20D space using the 'cosine' distance metric. After the unsupervised embedding model defined the distance between ion images, we calculated the Pearson correlation coefficient between the corresponding embedded vectors to rank comparison images with respect to the target image. This model is unsupervised and does not use the gold standard set.

2.7.4 UMAP+GBT model

Since in our case supervision is actually possible, we extended the UMAP model with a supervised model on top. Namely, we used gradient boosted trees (GBT), a state-of-the-art regression model (Chen and Guestrin, 2016), feeding UMAP 20D features as input and regressing them onto rank targets from the gold standard set with the MSE loss function.

2.7.5 Pi model

The Pi model is based on the recently developed approach of temporal ensembling for semi-supervised learning (Laine and Aila, 2016). This approach uses an ensemble of network outputs from different training epochs as quasi-targets for training on unlabelled samples, which has been shown to significantly improve the final model quality. In our case, the Pi model follows the general architecture of the Xception-based model, but the last layers are replaced with two heads for two loss components: (i) supervised loss as in the Xception-based model, the MSE between the network prediction and the rank and (ii) unsupervised loss intended to stabilize the network prediction; we define it as the squared error between network predictions on a pair of ion images and the same pair of ion images subjected to various image augmentations (intensities and geometric transformations).

The unsupervised loss component has allowed us to use ∼40 000 unlabelled ion images from 3685 public METASPACE datasets, gathering ∼56 000 unlabelled pairs from them for training in addition to the labelled gold standard set.

2.8 Evaluation of the co-localization measures

For each set, we calculate Spearman and Kendall correlation coefficients separately for each target-comparison set and report the mean and median values over all sets.

3 Results

3.1 The co-localization gold standard set

We have initially selected 239 datasets from METASPACE with 304 target-comparison sets of ion images for 42 rankers from 9 laboratories (Table 1).

3.2 Pilot study

The pilot study was crucial to inform us about the complexity and subjectivity of the task and to design the final version of the web app and the study accordingly. In particular, we learned that ranking comparison images was more natural for the rankers than ordering them because this allowed the rankers to assign the same rank to several comparison images. Selecting high-quality datasets and skipping noisy ion images and images with just a few non-zero pixels was crucial for obtaining reproducible rankings. Some rankers preferred to leave non-co-localized images unranked and we have implemented this option for the final study.

3.3 Agreement between experts

Table 1 shows the average pairwise ranker correlation values for each laboratory that represent agreement between rankers. Note that the agreement values cannot and shall not be compared across different laboratories because every laboratory ranked images

Table 1. Information about the co-localization gold standard set created from public METASPACE datasets contributed by nine laboratories with target-comparison sets ranked manually by 42 experts from these laboratories

Laboratory	Number of rankers	Number of datasets	Number of sets	Average pairwise agreement			
				Spearman		Kendall	
				Mean	Med	Mean	Med
University of Copenhagen	8	66	78	0.534	0.730	0.489	0.620
EMBL	8	33	53	0.763	0.804	0.666	0.710
University of Melbourne	6	29	40	0.806	0.836	0.732	0.742
JLU Giessen	5	29	33	0.612	0.688	0.545	0.597
IBMP	3	20	20	0.638	0.681	0.550	0.604
PNNL	3	20	20	0.686	0.687	0.596	0.609
MPI Bremen	3	19	20	0.843	0.925	0.787	0.879
UT Austin	3	16	20	0.808	0.866	0.745	0.770
M4I	3	7	20	0.730	0.680	0.660	0.608
Total	42	239	304	0.700	0.773	0.629	0.666
Final version after filtering out rankers and sets	**38**	**182**	**234**	**0.791**	**0.800**	**0.711**	**0.708**

Note: The bold highlighted row corresponds to the best results.

Table 2. Performance of the co-localization measures requiring no learning

Co-localization measure	Correlation with the expert rankings in the gold standard set						Optimal parameters of image transformations		
	Spearman			Kendall					
	Mean	Med	Mean SD	Mean	Med	Mean SD	Quant value	Hotspot remov.	Med win size
Cosine	**0.794**	**0.849**	**0.012**	**0.682**	**0.720**	**0.014**	**0.5**	**No**	**3**
tf–idf-cosine	0.769	0.825	0.012	0.653	0.689	0.013	0.65	No	0
Spearman	0.737	0.783	0.011	0.620	0.647	0.013	0.7	No	0
Pearson	0.788	0.838	0.014	0.674	0.719	0.013	0	No	0
SSIM	0.559	0.623	0.015	0.581	0.488	0.016	0	No	0

Note: The values of the Spearman and Kendall correlation coefficients to the gold standard are shown, together with the optimal parameters for image transformations. 'Mean' and 'Med' stand for the mean and median of the correlation values across the gold standard. 'Mean SD' is the standard deviation of the mean correlation value over 100 bootstrapped samples from the gold standard. The best performance is achieved by the cosine measure. The bold highlighted row corresponds to the best results.

from different METASPACE datasets. For example, ranking images with a simple and clear spatial structure led to higher agreement values. The mean correlation across all sets was 0.700 (Spearman) and 0.629 (Kendall). After excluding sets and rankers with low agreement, the mean agreements for the final version of the gold standard set is 0.791 (Spearman) and 0.711 (Kendall).

3.4 The gold standard set

The final version of the gold standard set includes 234 sets with 2210 ion image pairs from 182 public imaging datasets from METASPACE ranked by 38 rankers from 9 laboratories, available at https://github.com/metaspace2020/coloc/tree/master/GS. The datasets represented human (37%), mouse (21%), pig (7%), rat (6%) and other organisms; brain (27%), kidney (11%), skin (9%), seed (4%) and other organs; MALDI (84%) and DESI (16%) ionization; DHB (44%), DAN (17%), DHA (6%), BPYN (5%) and other MALDI matrices; Orbitrap (69%) and FTICR (31%) mass analysers; positive (68%) and negative (32%) polarity. For every target-comparison pair of ion images, average rank across three rankers has been assigned. The ranks range from 0, representing the highest co-localization, to 10, representing the lowest or no co-localization.

3.5 Evaluation of co-localization measures that require no learning

Table 2 shows the performance of co-localization measures requiring no learning, measured using the gold standard set as Spearman and Kendall rank correlation with the expert rankings. For each measure, we show its best performance and optimal image transformation parameters. The best performing measure is the cosine similarity with quantile threshold 0.5, without hotspot removal and with median filter with window size 3. The second best measure is the Pearson correlation measure with no image transformation applied. The SSIM measure recently proposed in the context of imaging MS (Ekelöf *et al.*, 2018) was outperformed by other measures.

Table 3 shows the effect of using different types of image transformation on the performance of the cosine measure. Surprisingly, applying hotspot removal did not improve the performance. Denoising images by using the moving median filter improved the performance only marginally (Spearman correlation from 0.792 to 0.794), whereas using quantile thresholding led to a larger improvement (Spearman correlation from 0.779 to 0.794). However, taken into account the SD values of the mean estimates, this improvement cannot be claimed to be significant.

3.6 Evaluation of co-localization measures based on deep learning

Table 4 shows the performance of co-localization measures based on deep learning models. The Pi model achieved the best performance, with a slight improvement over cosine similarity and similar to the human-to-human agreement between the experts in our study (mean Spearman of 0.791; see Table 1). This is no surprise since the Pi model is a state-of-the-art semi-supervised model that makes

Table 3. The effect of using different types of image transformations onto the performance of the cosine-based measure

Image transformation applied before calculating cosine similarity	Correlation with the expert rankings in the gold standard set						Optimal parameters of image transformations		
	Spearman			Kendall					
	Mean	Med	Mean SD	Mean	Med	Mean SD	Quant value	Hotspot remov.	Med win size
Quantile thresholding, hotspot removal, denoising	**0.794**	**0.849**	**0.012**	**0.682**	**0.720**	**0.014**	**0.5**	**No**	**3**
Quantile thresholding, hotspot removal (no denoising)	0.792	0.844	0.012	0.681	0.720	0.012	0.5	No	—
Hotspot removal, denoising (no quantile thresholding)	0.779	0.842	0.012	0.665	0.719	0.013	—	No	3

Note: The optimal parameters for image transformations are shown. ‘Mean’ and ‘Med’ stand for the mean and median of the correlation values across the gold standard. ‘Mean SD’ is the standard deviation of the mean correlation value over 100 bootstrapped samples from the gold standard. The bold highlighted row corresponds to the best results.

Table 4. The performance of deep learning-based models measured as Spearman and Kendall correlations with expert rankings in the gold standard set

Image transformation applied before calculating cosine similarity	Correlation with the expert rankings in the gold standard set					
	Spearman			Kendall		
	Mean	Med	Mean SD	Mean	Med	Mean SD
Xception model	0.777	0.820	0.011	0.682	0.716	0.011
Pi model	**0.797**	**0.847**	**0.011**	**0.712**	**0.752**	**0.011**
Unsupervised UMAP	0.761	0.827	0.012	0.656	0.686	0.015
UMAP+GBT	0.758	0.845	0.016	0.672	0.741	0.016
Mu model	0.725	0.804	0.016	0.638	0.705	0.016

Note: ‘Mean’ and ‘Med’ stand for the mean and median of the correlation values across the gold standard. ‘Mean SD’ is the standard deviation of the mean correlation value over 100 bootstrapped samples from the gold standard. The best performance is achieved by the semi-supervised Pi model that makes use of both labelled and unlabelled data. The bold highlighted row corresponds to the best results.

use of both labelled data from the gold standard set and unlabelled data from METASPACE.

3.7 Inferring molecular relationships by mining public METASPACE data

We have applied the best derived co-localization measures to illustrate how they can be used on a large scale to mine data from the public knowledge base METASPACE. More specifically, we aimed to infer co-localization relationships between all molecules represented in public METASPACE annotations. For this, we downloaded ion images for all annotations from 3685 public datasets in METASPACE using its API (https://github.com/metaspace2020/metaspace/tree/master/metaspace/python-client), calculated co-localizations between all ion images within a dataset using either the cosine score after median (0.5 quantile) thresholding or deep learning Pi model, the best performing methods in their respective classes, and averaged co-localization across all datasets. Finally, we visualized all 10 273 resulting molecular formulas in a 2D space using UMAP with the average co-localization used as the pre-computed distance. The annotations that are more co-localized on average are shown closer to each other (Fig. 3).

To investigate whether the inferred co-localization properties are associated with chemical properties of the molecules, we highlighted glycerolipids, an important class of lipids which are known to be easily detectable by imaging MS (Fig. 3a). Note that the assignment of a molecular annotation (formula) to a molecular class was performed accounting for potential ambiguity, with unambiguously assigned annotations shown in green and ambiguously assigned annotations shown in red (Fig. 3). One can see that glycerolipids indeed form dense clusters that indicates their high average co-localization. A subclass of glycerolipids, triradylcglycerols (with the classes names as in HDMB) represent the majority of the glycerolipids in METASPACE and form the densest clusters (Fig. 3b). Sparser representation of glycerolipids in the negative polarity data (Fig. 3c) illustrates the common knowledge of the positive mode being the preferred way of ionization for this class of lipids. Using another co-localization measure (deep learning-based Pi model instead of the cosine) also confirms the findings but shows a visible difference in data organization. This reflects the robust capacity of both measures to capture chemically associated co-localization and also shows the differences between them, which can be potentially used in the future to further improve the results.

Another class of lipids, glycerophospholipids, represents a large part of molecules in METASPACE, clearly forming a cluster in the UMAP chemical space (Fig. 3e). Performing examination in a way similar to Figure 3a–d, one can see that the molecular subclass of glycerophospholipids, glycerophosphoethanolamines, represents the core of the cluster of co-localized glycerophospholipids (Fig. 3f). Opposite to glycerolipids, glycerophospholipids are known to be ionizable in both positive and negative modes, and this is reflected in their strong presence as well as clustered appearance on Figure 3g. Examining deep learning Pi score-based mapping (Fig. 3h), one can see that despite dense spacing, there is clearly less separation visible to the class of glycerolipids (Fig. 3h versus d) compared to the cosine score-based UMAP visualization (Fig. 3g versus a), which makes cosine score-based results easier for interpretation.

4 Discussion

4.1 Other approaches to obtain gold standard data

In addition to the approach presented in this manuscript, we have also considered creating gold standard data either by simulating it or by authentic standards or mixtures of standards. However, our

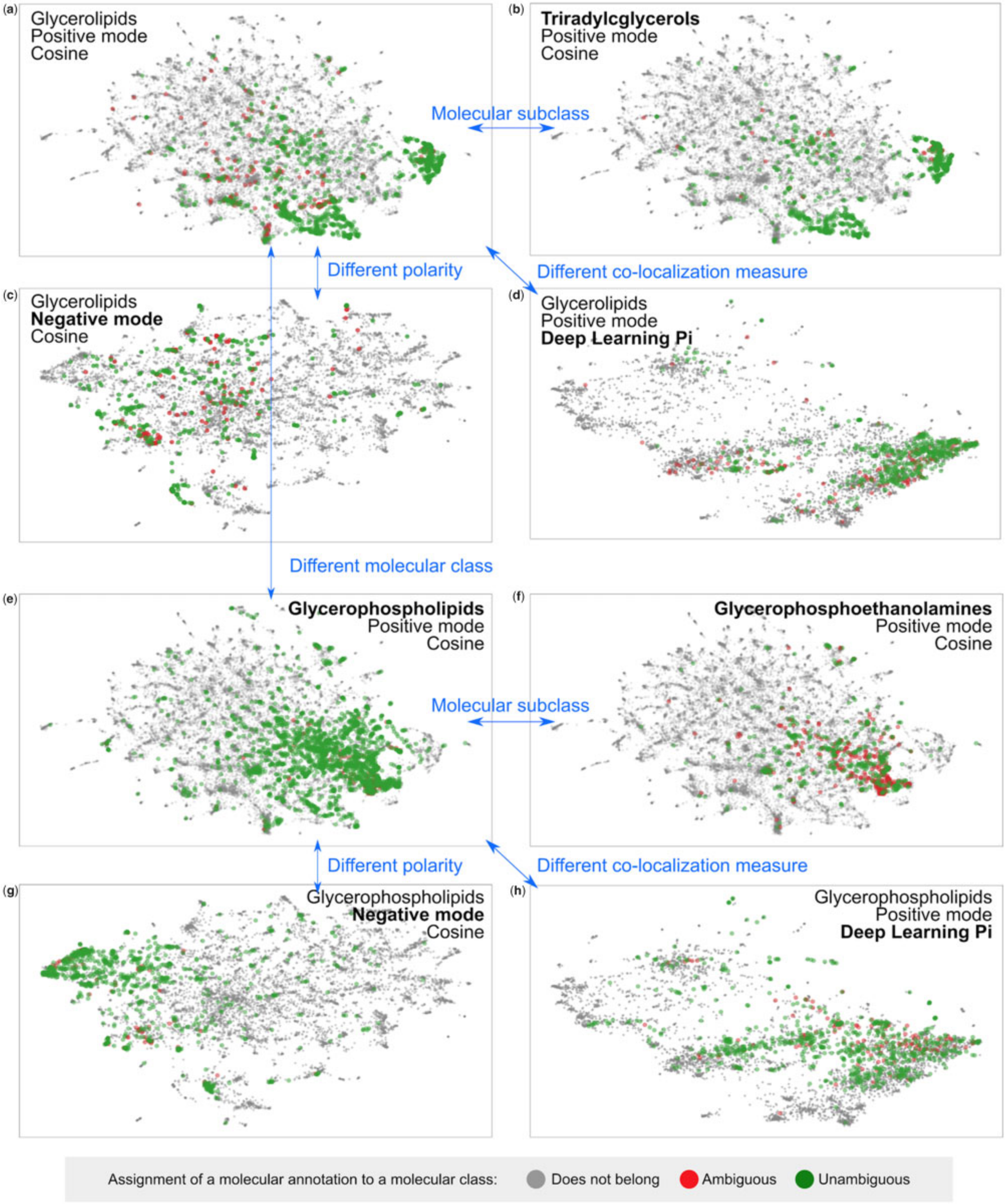


Fig. 3. Visualization of co-localization molecular relationships as learned from METASPACE. Dots representing annotations (each corresponding to 1 of 10 273 unique molecular formulas) are mapped based on their average co-localization across 3685 public METASPACE datasets. For a molecular class, the green colour represents unambiguous assignment when all isomers belong to the class whereas the red colour represents ambiguous assignment when some isotopes belong to another class. (Color version of this figure is available at *Bioinformatics* online.)

pilot experiments indicated that neither of these approaches would be adequate for evaluating co-localization due to being over-simplistic as well as not representing complex spatial patterns, noise and background adequately enough. Simulating imaging MS data is by itself an unsolved challenge and proposed methods are rather limited and did not gain recognition. Spotting authentic standards, first, would hardly represent the variety of technologies, spatial resolutions and types of tissues that we can get access to through

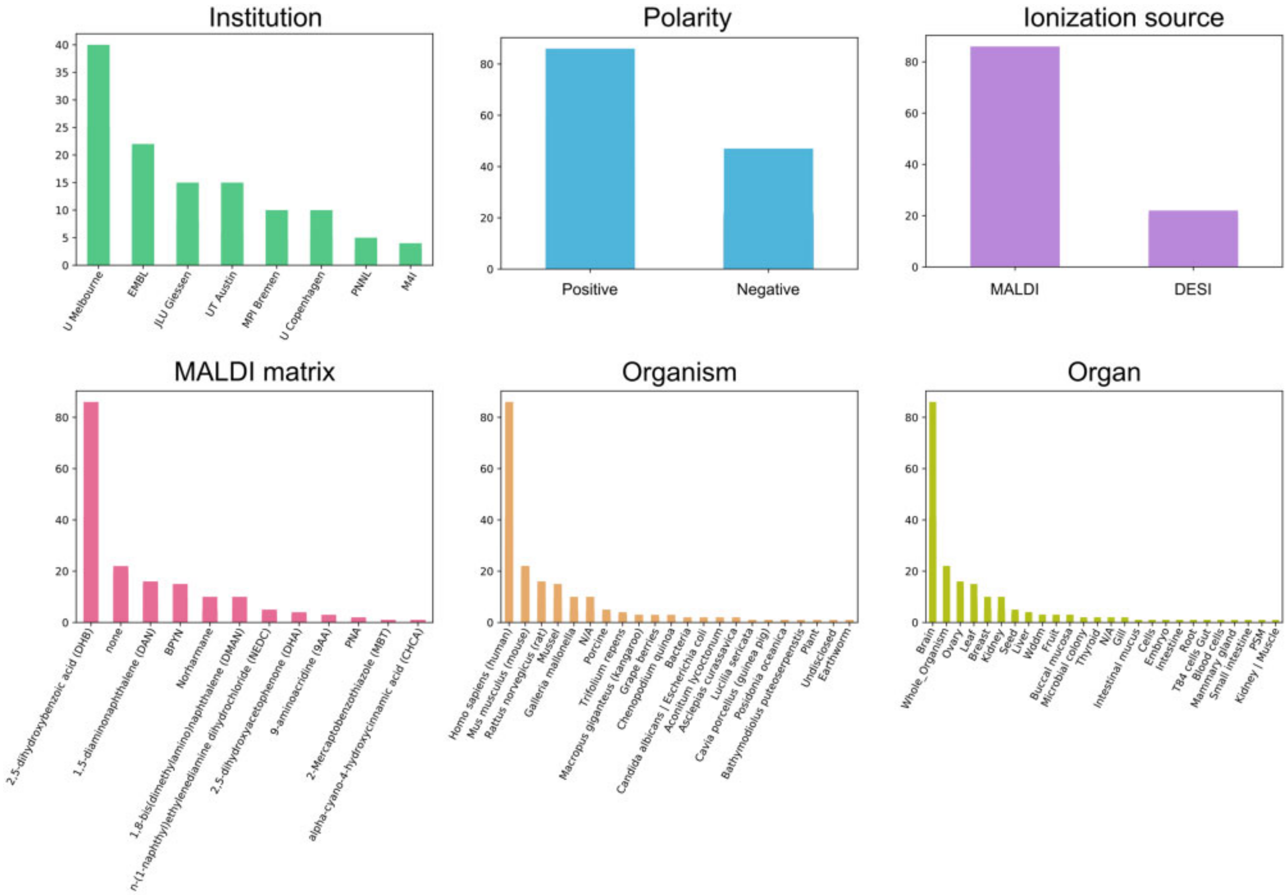


Fig. 4. Histograms of the properties of the public METASPACE datasets selected for the gold standard

METASPACE. Second, experimental data from spotted authentic standards would provide only a limited insight into which ions should be co-localized because of in-source fragmentation and clusters formation which potentially can lead to co-detections of ions even from different standards. For example, adenosine could be detected from both spotted ADP and ATP as an in-source fragment and this represents only one known example out of various possible in-source fragments.

4.2 Gold standard

Creating a high-quality gold standard set of expert-ranked pairs of target-comparison images was possibly the most challenging part of the study. Not only it required scientific formulation of the co-localization problem and development of an experiment design able to capture the perceived extent of co-localization from the experts, but it was also the most time-consuming part of our study to organize the whole ranking experiment by selecting datasets, recruiting almost 50 experts, communicating with them, reminding them to complete the task, and when necessary coming back to them with requests for corrections. Altogether it required 95 emails solely for communicating with the rankers. Despite having expertise in performing crowdsourcing studies in imaging MS (Ovchinnikova *et al.*, 2019; Palmer *et al.*, 2015) and overwhelmingly positive support of METASPACE users in performing the ranking, running this study would not be possible without access to diverse public data in METASPACE and without using modern web technologies employed for the RankColoc web app that both critically facilitated the process. The achieved average pairwise correlation between the rankers (mean Spearman 0.791) confirms a strong inter-ranker agreement. This indicates that there is a consensus between experts with respect to perceived co-localization and, importantly, that this consensus was successfully captured in the gold standard set, thus validating our efforts.

Performing the pilot study was essential to avoid pitfalls and refine the experimental design and the web app for more objective ranking before engaging a large number of experts. Nevertheless, after performing the complete study, we see opportunities for next-level improvement. For example, in the spirit of active learning, we could choose comparison ions not randomly but those ions where our models are most uncertain about their ranking.

Figure 4 shows the statistics for the properties of the METASPACE datasets selected for the gold standard. When selecting them, we mainly were driven by the requirement of having at least three rankers from every laboratory. Thus, big laboratories with high numbers of active METASPACE received more representation in the gold standard set. This also led to some bias towards the types of samples and the mass spectrometry used. However, comparing the properties of the datasets in the gold standard to the overall METASPACE (Alexandrov *et al.*, 2019), we see that in general the gold standard is relatively representative. We have evaluated whether there is a significant bias due to overrepresentation of brain tissue datasets in the gold standard. For this, we 10 times randomly sub-sampled one-third of all brain datasets in the gold standard. The calculated the averaged mean Spearman for the best cosine measure was 0.787 which is lower than for the full gold standard (0.794) but the difference is not significant. This indicates that the effect of this particular bias is present but not significant.

Taking into account the efforts necessary for producing such a gold standard set, we do not expect it to be repeated on a larger scale in the near future. However, we are considering to implement an online approach where a target-comparison set or a reduced version of it will be occasionally shown to METASPACE users. This approach would provide a continuous population of the gold standard set. However, it should be carefully designed to ensure the quality and check for consistency, since the ranking task will be split into small subtasks and performed over a period of time by a larger diverse crowd of rankers.

4.3 Co-localization measures

We hope that our results, comparing a variety of deep learning models, and this discussion will be helpful for future deep learning applications in imaging MS. In Supplementary Figure S1 and Tables S1 and S2, we show values for all developed methods for an example target-comparison set.

Interestingly, the unsupervised model UMAP performed on par with the supervised UMAP+GBT model, namely its own version enhanced with supervision through GBT. This may indicate that the structural properties of co-localization are relatively evident in the data.

At the same time, the achieved performance for both the best deep learning model (mean Spearman 0.797 for the Pi model) and the cosine measure (mean Spearman 0.794) are close to the average pairwise agreement between the rankers (mean Spearman 0.791) that indicates that we achieved close to theoretically best performance. This would also explain only a slight improvement when using an advanced deep learning-based methods compared to the cosine similarity.

The slight positive difference of the best performance compared to the average pairwise agreement between rankers (0.797 or 0.794 versus 0.791) does not necessarily indicate an overfitting but can be due to the averaging of rankings in the produced gold standard set, thus introducing positive effects of averaging compared to the values used for the rankers agreement calculation.

We would like to note other specifics of the considered problem which potentially do not allow to capitalize on the full potential of the deep learning methods: ion images from the same dataset have the same size and structure and can be compared pixel-by-pixel after flattening. Ion images also do not undergo changes in the view angle or brightness or other non-linear deformations that would apply to, e.g. photos used in computer vision where deep learning significantly benefits from its capacity to extract abstract visual features thus allowing comparison of different images showing the same object. Here, future efforts can be focused on developing next-level methods for spatial association between molecules that would consider 'molecular microenvironment' rather than 'tissue section' context.

The results potentially indicate a bottleneck in the size of the gold standard set (2340 pairs of ranked images), since the best performance was achieved by a semi-supervised model which, besides the gold standard set, used all public METASPACE data for deriving a representation of ion images.

It was somewhat surprising to obtain the best performance for the cosine measure applied after the median thresholding which, for an ion image, sets all pixel intensities smaller than the median intensity to zero thus neglecting half of the pixels. Note that the cosine measure applied after the median thresholding considers only the areas where both images have high-intensity pixels (with intensities above the median value). Furthermore, zeroing half of the pixels provides an efficient denoising by substantially reducing the values of the cosine measure between noisy images. Compared to two similarly looking images, cosine distance between two random (noisy) images is reduced more after median thresholding due to the smaller overlap in the sets of non-zero pixels. Taking these considerations into account, we speculate that the cosine measure combined with the median thresholding achieves high performance because it corresponds to the perception of rankers which make their judgement about the similarity of images based on high-intensity areas of images. Continuing this line of discussion, let us bring attention to the second best performing measure which was the Pearson correlation without any image transformation applied. Since Pearson correlation is not recommended for highly skewed distributions whereas the intensities in imaging MS are Poisson distributed (Alexandrov *et al.*, 2010), we evaluated whether the log-transformation of pixel intensities would improve the performance. Interestingly, it was not the case and the mean Spearman with the expert rankings after log-transformation was 0.742 that is lower than without any transformation (0.788). Potentially, this reflects the same assumption we made about rankers perception, as using the Pearson correlation for highly skewed data without log-transformation makes it biased towards high-intensity pixels.

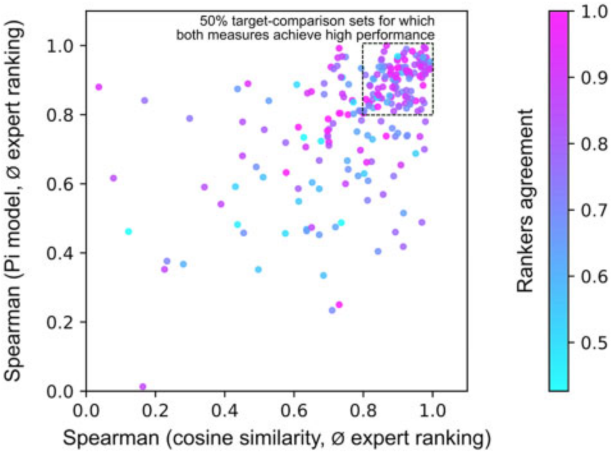


Fig. 5. Scatterplot showing for each of 2340 target-comparison sets from the gold standard how well the best co-localization measures (deep learning Pi model, cosine similarity after median thresholding) reproduce the average (ø) expert ranking, as measured by the Spearman correlation. Each dot is coloured according to the rankers agreement for the respective target-comparison set

Comparing the best co-localization measures (deep learning Pi model and cosine similarity after median thresholding), we investigated how well they correspond to expert ranking for each target-comparison pair from the gold standard. Figure 5 shows that there is no visible difference between these two measures: for 50% of all target-comparison sets both measures achieve high performance (Spearman correlation with the expert ranking is >0.8). Despite the fact that error analysis of low-valued sets has not revealed any factors that would allow us to improve the measures, one can potentially combine the considered measures and thus achieve a better performance with an ensemble ranking. Interestingly, Figure 4 highlights that the target-comparison pairs for which both measures performed well have also visibly high values of the rankers agreement. This provides another confirmation that the developed co-localization measures reproduce the perceived co-localization when experts themselves agree on it.

4.4 Applications

A wide coverage of organisms, organs, ionization types, MALDI matrices and mass analysers represented in the imaging MS datasets used in the gold standard set ensures broad applicability for the findings and measures developed in this study. We expect key applications of the developed and evaluated co-localization methods to be in the search for molecular biomarkers associated with either a particular molecule or a region of interest. They should also improve distance-based methods for data analysis, e.g. representation of the full dataset using clustering of ion images (Alexandrov *et al.*, 2013). Moreover, we expect this work to provide a scientifically rigorous justification for using these measures in systems biology approaches aimed at uncovering molecular relationships between molecules by assuming the tissue representing cells of different phenotypes. Here, cutting-edge methods relying on distance or similarity measures, such as UMAP demonstrated in this paper, can replace more conventional methods such as PCA, NMF or t-SNE.

Acknowledgements

We thank the rankers of public METASPACE data who shared their data publicly and helped us create the gold standard set: Asta Mari-a Joensen, Charlotte Bagger, Julian Schneemann, Meghan Friis, Fernanda Endringer Pinto, Anne Mette Handler, Andreas Danielsen, Katharina Clitherow, Sophie Jacobsen (Janfelt Lab, University of Copenhagen), Mohammed Shahraz, Don Nguyen, Sergio Triana, Veronika Saharuka, Luca Rappez, Cristina Gonzalez Lopez, Aslihan Inal (EMBL), Dinaiz Thinagaran, Sanuli Paralkar, Farheen

Farzana, Edita Ritmejeryte, Nicholas Sing, Berin Boughton (Metabolomics Australia, University of Melbourne), Michael Waletzko, Dhaka Bhandari, Domenic Dreisbach, Patrik Kadesch (Spengler Lab, Justus Liebig University Giessen), Elisa Ruhland, Julien Delecolle, Claire Villette (Heintz Lab, IBMP, University of Strasbourg), Dusan Velickovic, Christopher Anderton, Arunima Bhattacharjee (Anderton Lab, Pacific Northwest National Laboratory), Manuel Liebeke, Benedikt Geier, Emilia Sogen (Liebeke Lab, Max Planck Institute for Marine Microbiology Bremen), Marta Sans, Jialing Zhang, Kyana Garza (Eberlin Lab, University of Texas Austin) and Shane Ellis, Pieter Kooijman, Lennart Huizing (M4I Institute, Maastricht University). We thank the reviewers and the editor for helping us improve the manuscript.

Funding

This work was supported by the European Union's Horizon 2020 programme [N°634402, N°777222 to K.O., L.S., T.A.], the Russian Foundation for Basic Research [18-54-74005 to S.N., T.A.] and the European Research Council Consolidator [METACELL N°773089 to T.A.].

Conflict of Interest: Until 2020, Theodore Alexandrov was on the Scientific Advisory Board of SCiLS, a company developing software for imaging mass spectrometry.

References

Aaron,J.S. *et al.* (2018) Image co-localization—co-occurrence versus correlation. *J. Cell Sci.*, **131**, jcs211847.

Alexandrov,T. (2012) MALDI imaging mass spectrometry: statistical data analysis and current computational challenges. *BMC Bioinformatics*, **13**, S11.

Alexandrov,T. *et al.* (2010) Spatial segmentation of imaging mass spectrometry data with edge-preserving image denoising and clustering. *J. Proteome Res.*, **9**, 6535–6546.

Alexandrov,T. *et al.* (2013) Analysis and interpretation of imaging mass spectrometry data by clustering mass-to-charge images according to their spatial similarity. *Anal. Chem.*, **85**, 11189–11195.

Alexandrov,T. *et al.* (2019) METASPACE: a community-populated knowledge base of spatial metabolomes in health and disease. *bioRxiv*, pp. 1–22. https://doi.org/10.1101/539478.

Buchberger,A.R. *et al.* (2018) Mass spectrometry imaging: a review of emerging advancements and future insights. *Anal. Chem.*, **90**, 240–265.

Chen,T. and Guestrin,C. (2016) XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *San Francisco, California, USA*, pp. 785–94. ACM.

Chollet,F. (2017) Xception: deep learning with depthwise separable convolutions. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, *Honolulu, HI, USA*. pp. 1800–1807.

Doerr,A. (2018) Mass spectrometry imaging takes off. *Nat. Methods*, **15**, 32.

Dreisewerd,K. and Yew,J.Y. (2017) Mass spectrometry imaging goes three dimensional. *Nat. Methods*, **14**, 1139–1140.

Ekelöf,M. *et al.* (2018) Evaluation of digital image recognition methods for mass spectrometry imaging data analysis. *J. Am. Soc. Mass Spectrom.*, **29**, 2467–2470.

Huang,T. *et al.* (1979) A fast two-dimensional median filtering algorithm. *IEEE Trans. Acoust. Speech Signal Process.*, **27**, 13–18.

Kaddi,C. *et al.* (2011) Hypergeometric similarity measure for spatial analysis in tissue imaging mass spectrometry. In: *Proceedings IEEE Int Conf Bioinformatics Biomed*, *Atlanta, GA, USA*. pp. 604–607.

Laine,S. and Aila,T. (2016) Temporal ensembling for semi-supervised learning. *arXiv [cs.NE]*. arXiv. pp. 1–13. http://arxiv.org/abs/1610.02242.

Leskovec,J. *et al.* (2014) *Mining of Massive Datasets*. Cambridge University Press, Cambridge.

McCombie,G. *et al.* (2005) Spatial and spectral correlations in MALDI mass spectrometry images by clustering and multivariate analysis. *Anal. Chem.*, **77**, 6118–6124.

McDonnell,L.A. *et al.* (2008) Mass spectrometry image correlation: quantifying colocalization. *J. Proteome Res.*, **7**, 3619–3627.

McInnes,L. *et al.* (2018) UMAP: uniform manifold approximation and projection for dimension reduction. *arXiv [stat.ML]*. arXiv. pp. 1–51. http://arxiv.org/abs/1802.03426.

Ovchinnikova,K. *et al.* (2019) OffsampleAI: artificial intelligence approach to recognize off-sample mass spectrometry images. BMC Bioinformatics, in press. https://doi.org/10.1186/s12859-020-3425-x.

Palmer,A. *et al.* (2015) Using collective expert judge-ments to evaluate quality measures of mass spectrometry images. *Bioinformatics*, **31**, i375–384.

Palmer,A. *et al.* (2017) FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry. *Nat. Methods*, **14**, 57–60.

Schulz,S. *et al.* (2019) Advanced MALDI mass spectrometry imaging in pharmaceutical research and drug development. *Curr. Opin. Biotechnol.*, **55**, 51–59.

Vaysse,P.-M. *et al.* (2017) Mass spectrometry imaging for clinical research—latest developments, applications, and current limitations. *Analyst*, **142**, 2690–2712.

Wang,Z. *et al.* (2004) Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, **13**, 600–612.

Watrous,J.D. *et al.* (2011) The evolving field of imaging mass spectrometry and its impact on future biological research. *J. Mass Spectrom.*, **46**, 209–222.