

## Automated biomarker candidate discovery in imaging mass spectrometry data through spatially localized Shapley additive explanations

Tideman, Leonoor E.M.; Migas, Lukasz G.; Djambazova, Katerina V.; Patterson, Nathan Heath; Caprioli, Richard M.; Spraggins, Jeffrey M.; Van de Plas, Raf

**DOI**

[10.1016/j.aca.2021.338522](https://doi.org/10.1016/j.aca.2021.338522)

**Publication date**

2021

**Document Version**

Final published version

**Published in**

Analytica Chimica Acta

**Citation (APA)**

Tideman, L. E. M., Migas, L. G., Djambazova, K. V., Patterson, N. H., Caprioli, R. M., Spraggins, J. M., & Van de Plas, R. (2021). Automated biomarker candidate discovery in imaging mass spectrometry data through spatially localized Shapley additive explanations. *Analytica Chimica Acta*, 1177, Article 338522. <https://doi.org/10.1016/j.aca.2021.338522>

**Important note**

To cite this publication, please use the final published version (if applicable). Please check the document version above.

**Copyright**

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

**Takedown policy**

Please contact us and provide details if you believe this document breaches copyrights. We will remove access to the work immediately and investigate your claim.



# Automated biomarker candidate discovery in imaging mass spectrometry data through spatially localized Shapley additive explanations



Leonoor E.M. Tideman<sup>a</sup>, Lukasz G. Migas<sup>a</sup>, Katerina V. Djambazova<sup>b, d</sup>,  
Nathan Heath Patterson<sup>b, c</sup>, Richard M. Caprioli<sup>b, c, d, e, f</sup>, Jeffrey M. Spraggins<sup>b, c, d</sup>,  
Raf Van de Plas<sup>a, b, c, \*</sup>

<sup>a</sup> Delft Center for Systems and Control, Delft University of Technology, Delft, Netherlands

<sup>b</sup> Mass Spectrometry Research Center, Vanderbilt University, Nashville, TN, USA

<sup>c</sup> Department of Biochemistry, Vanderbilt University, Nashville, TN, USA

<sup>d</sup> Department of Chemistry, Vanderbilt University, Nashville, TN, USA

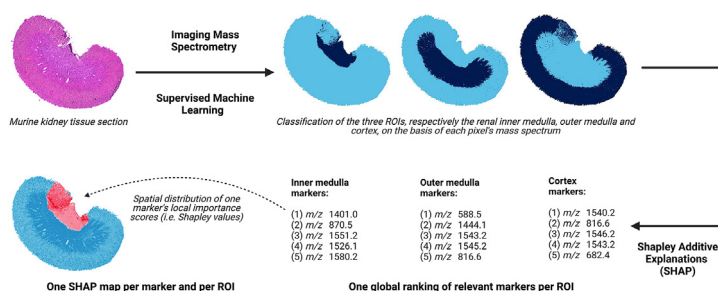
<sup>e</sup> Department of Pharmacology, Vanderbilt University, Nashville, TN, USA

<sup>f</sup> Department of Medicine, Vanderbilt University, Nashville, TN, USA

## HIGHLIGHTS

- Biomarker discovery in imaging mass spectrometry data can be considered a feature ranking problem.
- Interpretability methods enable automated estimation of biomarker candidate potential.
- Our biomarker candidate discovery workflow uses Shapley additive explanations (SHAP).
- SHAP measures the direction and magnitude of a feature's influence on a classifier.
- SHAP maps provide insight into the spatial specificity of biomarker candidates.

## GRAPHICAL ABSTRACT



## ARTICLE INFO

### Article history:

Received 12 January 2021

Received in revised form

4 April 2021

Accepted 11 April 2021

Available online 26 April 2021

### Keywords:

Supervised machine learning

Model interpretability

Biomarker discovery

## ABSTRACT

The search for molecular species that are differentially expressed between biological states is an important step towards discovering promising biomarker candidates. In imaging mass spectrometry (IMS), performing this search manually is often impractical due to the large size and high-dimensionality of IMS datasets. Instead, we propose an interpretable machine learning workflow that automatically identifies biomarker candidates by their mass-to-charge ratios, and that quantitatively estimates their relevance to recognizing a given biological class using Shapley additive explanations (SHAP). The task of biomarker candidate discovery is translated into a feature ranking problem: given a classification model that assigns pixels to different biological classes on the basis of their mass spectra, the molecular species that the model uses as features are ranked in descending order of relative predictive importance such that the top-ranking features have a higher likelihood of being useful biomarkers. Besides providing the

**Abbreviations:** IMS, imaging mass spectrometry; *m/z*, mass-to-charge ratio; MALDI, matrix assisted laser desorption/ionization; ML, machine learning; PI, permutation importance; Q-TOF, quadrupole time-of-flight; SHAP, Shapley additive explanations; XGBoost, extreme gradient boosting.

\* Corresponding author. Delft Center for Systems and Control (Delft University of Technology, 3ME faculty, Mekelweg 2 - Gebouw 34, 2628 CD, Delft, the Netherlands).

E-mail address: [raf.vandepas@tudelft.nl](mailto:raf.vandepas@tudelft.nl) (R. Van de Plas).

<https://doi.org/10.1016/j.aca.2021.338522>

0003-2670/© 2021 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Imaging mass spectrometry  
 Shapley additive explanations  
 Explainable artificial intelligence

user with an experiment-wide measure of a molecular species' biomarker potential, our workflow delivers spatially localized explanations of the classification model's decision-making process in the form of a novel representation called SHAP maps. SHAP maps deliver insight into the spatial specificity of biomarker candidates by highlighting in which regions of the tissue sample each feature provides discriminative information and in which regions it does not. SHAP maps also enable one to determine whether the relationship between a biomarker candidate and a biological state of interest is correlative or anticorrelative. Our automated approach to estimating a molecular species' potential for characterizing a user-provided biological class, combined with the untargeted and multiplexed nature of IMS, allows for the rapid screening of thousands of molecular species and the obtention of a broader biomarker candidate shortlist than would be possible through targeted manual assessment. Our biomarker candidate discovery workflow is demonstrated on mouse-pup and rat kidney case studies.

© 2021 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

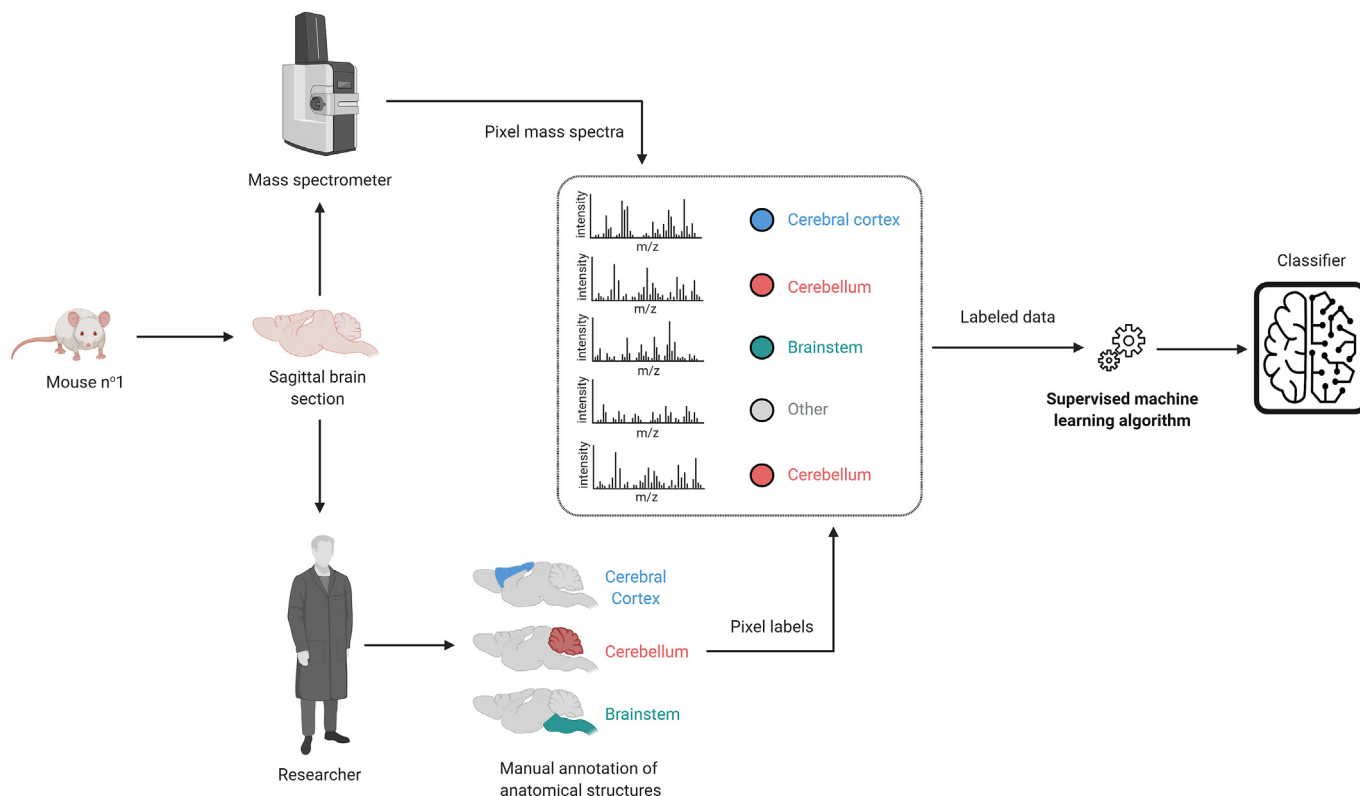
A biomarker can generally be considered an objectively measurable indicator of a specific biological state or disease condition [1,2]. Biomarkers can be used to differentiate between anatomical structures, cell types, and disease states, and lend themselves to the screening, diagnosis, and monitoring of disease, the identification of new drug targets, and the assessment of therapeutic response [1,3–5]. In our work, the term “biomarker candidate” refers to a putative molecular biomarker (i.e. a chemical species) that is differentially expressed between biological states [2]. One technology for discovering such molecular markers at scale is mass spectrometry, which characterizes molecular species in terms of their mass-to-charge ratio ( $m/z$ ). Imaging mass spectrometry (IMS) is a multiplexed, label-free imaging technology that uses mass spectrometry for the molecular mapping of tissues down to cellular resolution [6–8]. An IMS experiment involves collecting spatially localized mass spectra for each pixel in a grid of measurement locations across a sample surface [9,10]. Each pixel has an associated mass spectrum and each mass spectrum plots the measured signal intensity, which is indicative of relative abundance, versus the analytes'  $m/z$  values. The spatial distribution and relative abundance of an analyte can be visualized as an ion image, which plots the signal intensity measured for that analyte across all pixels of the sample's surface [11,12]. IMS is an excellent tool for biomarker discovery for the following three reasons: it is able to concurrently detect hundreds to thousands of analytes within a single experiment in an untargeted manner, it can probe analytes from a wide range of molecular classes (e.g. peptides, proteins, lipids, glycans, metabolites), and it enables the mapping of analytes' spatial distributions in relation to the (patho)histology of tissue samples [5,13,14].

One way novel biomarker candidates can be discovered is by observing the differential expression of molecules between distinct sample classes (e.g. different cell types, different organs, different stages of a disease) [2,15]. However, the large size and high-dimensionality of IMS datasets, which commonly yield several hundreds of thousands of pixels and several hundreds to thousands of molecular ions tracked per pixel, pose a challenge. Manually examining the spatial mapping of thousands of molecular species across the surface of a sample is laborious and risks introducing human subjectivity into the process, leading to results whose reproducibility cannot necessarily be guaranteed [11,16]. The amount of data generated by IMS experiments is so large that it has become more efficient (and in many cases necessary) to computationally search for biomarker candidates among a multitude of ion intensity signals [5,17]. In this work, we suggest a machine learning (ML) workflow for performing biomarker candidate

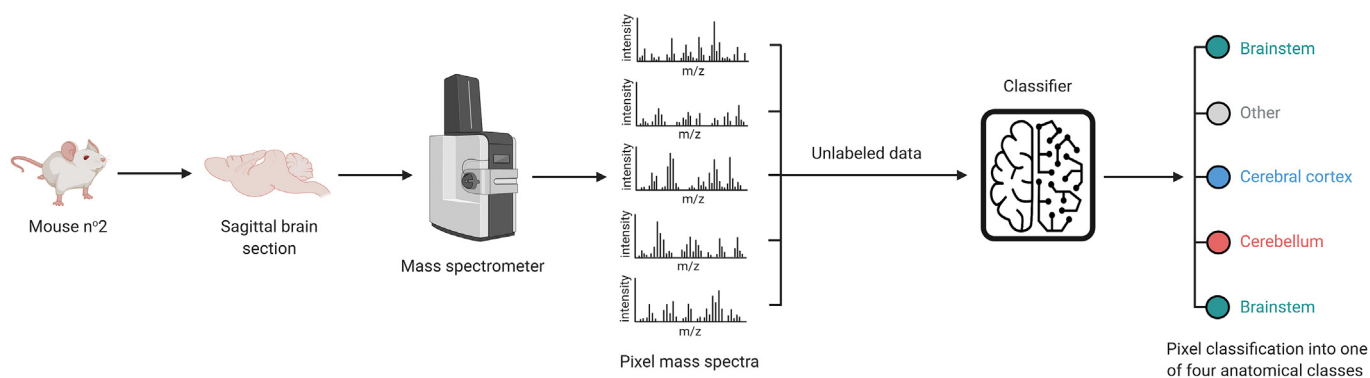
discovery that provides one with a shortlist of molecular species that are characteristic of the class for which biomarkers are sought. Our approach uses supervised ML models to classify mass spectra into different biological classes of interest and then uses state-of-the-art methods from the field of interpretable ML [18–20] to determine the discriminative relevance, and biomarker potential, of each molecular species.

In our work, an IMS dataset is represented by a data matrix  $X \in \mathbb{R}^{m \times n}$  whose rows  $x_i = X_{(i,:)}$ , for  $i = 1, 2, 3 \dots m$ , correspond to the mass spectra of the pixels making up the sample's surface and whose columns  $x^j = X_{(:,j)}$ , for  $j = 1, 2, 3 \dots n$ , correspond to the  $m/z$  bins per spectrum. The  $m$  rows and  $n$  columns of  $X$  can be respectively referred to as observations and features. Classification is a form of supervised ML in which the observations  $x_i$  are annotated with discrete class labels  $y_i$  that represent user-provided knowledge related to these observations. Binary classification problems involve a positive class (e.g. diseased tissue), labeled as  $y_i = +1$ , and a negative class (e.g. healthy tissue), labeled as  $y_i = -1$ . The positive class is usually the class of interest: in our case studies, it is the class for which we want to discover biomarker candidates. Problems with multiple target classes (e.g. multiple cell types or functional tissue units) can be decomposed into multiple binary classification problems, each of which involve differentiating one class from the remaining classes. In the context of our work, classification is therefore the task of learning a multivariate function  $f^* : \mathbb{R}^n \mapsto \{-1, +1\}$ , called a classifier or classification model, that assigns each pixel to a class according to the molecular information provided by its mass spectrum  $x_i$ . Note a difference between the classifier's discrete class prediction  $\hat{y}_i = f^*(x_i) \in \{-1, +1\}$  and its continuous raw output  $f(x_i) \in \mathbb{R}$ , where  $f : \mathbb{R}^n \mapsto \mathbb{R}$ . The classifier's prediction is the class label assigned to a particular observation  $x_i$ , whereas the classifier's raw output can be interpreted as the score (e.g. probability, log-odds ratio) of  $x_i$  being assigned to the positive class. Fig. 1a illustrates the process of building a classifier in IMS: a supervised ML algorithm fits a classification model to a labeled IMS dataset called the training dataset (i.e. mass spectra  $x_i$  whose class membership  $y_i$  is known). The resulting model can then be used to classify new unlabeled data (i.e. mass spectra  $x_i$  whose class membership  $y_i$  is unknown) as illustrated by Fig. 1b. The performance of a classifier is measured by its ability to generalize, that is to correctly predict the labels for new data instances such that  $\hat{y}_i = y_i$ .

Traditionally, applications of supervised ML in IMS focus on maximizing the predictive performance of classifiers designed to automate user-defined recognition tasks, without necessarily examining their decision-making processes. However, we suggest that examining the relationship between a classifier's features and



**a:** Diagram of the classifier building process. Imaging mass spectrometry data is obtained from a sagittal mouse brain tissue section. The pixels making up the sample surface, and their corresponding mass spectra, are assigned to one of four anatomical classes (cerebral cortex, cerebellum, brainstem, or other). Given the labeled data, a supervised machine learning algorithm learns a classifier that can correctly assign the pixels' mass spectra to the correct anatomical classes.



**b:** Diagram of the prediction process. The classifier is provided with a new imaging mass spectrometry dataset obtained from another mouse brain tissue section. The unlabeled pixels making up the new dataset are assigned to one of four anatomical classes (cerebral cortex, cerebellum, brainstem, or other) on the basis of their mass spectra.

**Fig. 1.** Diagrams of the classifier building and prediction processes in imaging mass spectrometry. Icons from Refs. [21,22].

its prediction is important because it can reveal which features, and thus which molecular species, enable the differentiation of classes.

<sup>1</sup> There is no formal definition of supervised machine learning model interpretability that is agreed-upon within the computer science and data science communities [23–25]. The definition we propose, namely explaining a model's decision-making process by quantifying the influence of its input features on its output, is specific to the analysis of imaging mass spectrometry data using supervised machine learning methods other than deep learning. We use the terms interpretability and explainability interchangeably.

Model interpretability is the ability to explain the predictions of a supervised ML model by reporting the relative predictive importance of its features<sup>1</sup>. The importance, or relevance, of a feature is a measure of how it influences the model's prediction, considering both its direct effect (i.e. statistical association with the prediction) and its indirect effect (i.e. statistical association between features) [19,26]. The local predictive importance of a feature measures its influence on the predictive model's output for a specific observation (e.g. the mass spectrum of one pixel), whereas the global predictive importance of a feature measures its influence on the

predictive model's output for all observations (e.g. all pixels of a sample) [19,20,26]. In addition to reporting which features drive the decision-making processes of supervised ML models, interpretability methods also facilitate model troubleshooting (e.g. debugging, monitoring, checking for bias). For example, in the context of IMS data analysis, interpretability methods make it possible to trace whether the decision-making process of a classifier is based on genuine biological patterns rather than on instrumental patterns or chemical noise that are spuriously associated to the class labels. ML interpretability methods effectively address the issue of supervised ML algorithms producing "black-box" models with unintelligible predictive mechanisms [18–20]. The importance of ML interpretability for knowledge discovery has recently been discussed in genomics [27] and single-cell mass spectrometry [28]. To our knowledge, our work is the first application of ML interpretability methods to IMS data for the purpose of biomarker candidate discovery. Our aim is to formulate and demonstrate how ML interpretability methods can be used to understand how the spatial distribution and relative abundance of certain molecular species relate to the classification of different regions of a tissue sample, effectively automating biomarker candidate discovery in IMS data.

Our approach to aiding biomarker discovery is to automate and accelerate the identification of discriminative features by empirically learning which molecular species' overexpression or underexpression enable the recognition of a user-defined class [15]. We translate the problem of biomarker discovery into a feature ranking problem: ML interpretability methods computationally estimate the importance of each feature, or  $m/z$  value, with regards to a specific classification task and produce a ranking of the features in descending order of predictive importance. Ranking the features in terms of predictive importance facilitates the identification of a shortlist of molecular species that are characteristic of a class of interest, and thus have a higher likelihood of being useful biomarkers. In addition to providing one with a global understanding of which molecular species hold potential for recognizing a user-provided class, our approach uses SHAP maps to give the user spatially localized insight into each biomarker candidate's relationship with the class of interest. SHAP maps are a novel graphical representation of a model's decision-making process that can yield a nuanced local assessment of a biomarker candidate's potential and spatial specificity. Our biomarker candidate discovery workflow is therefore a scalable computational tool that enables one to rapidly, efficiently, and automatically filter the multitude of molecular species recorded by IMS down to a panel of promising biomarker candidates that deserve further study and validation.

## 2. Machine learning methodology

### 2.1. Extreme gradient boosting for imaging mass spectrometry data classification

There are many applications of supervised ML in IMS: random forests [17,29,30], support vector machines [29,31], convolutional neural networks [32,33], and gradient boosting machines [34,35] are frequently used classification model types. Decision trees are particularly suitable for IMS data analysis because they are non-linear and non-parametric predictive models that can account for complex dependencies between features, do not make assumptions about the underlying data distribution, and do not require feature scaling. A decision tree is a directed graph that partitions the feature space by recursive binary splitting: its nodes correspond to subsets of the data, and its branches correspond to the partitioning of a feature above or below a splitting threshold [36–38]. Given

that a single decision tree is neither flexible nor stable enough to achieve high predictive performance on IMS data classification tasks, combining multiple decision trees into an ensemble model is usually a preferable strategy [37,39]. We therefore choose to use XGBoost models for classification. XGBoost is a fast and scalable implementation of (stochastic) regularized gradient boosting that was developed by Chen and Guestrin in 2016 [40] based on the work of Friedman [41,42], Freund and Schapire [43]. An XGBoost model is an ensemble of regression trees (i.e. decision trees that output real values in their terminal nodes) that can perform classification by additive logistic modeling [44,45].

$$\min_f \sum_{i=1}^{m_{train}} \mathcal{L}(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(\tau_k) \quad (1)$$

Regularized gradient boosting is a forward stagewise additive modeling algorithm for solving numerical optimization problems of the form of Equation (1).  $\mathcal{L}$  is a differentiable loss function (e.g. negative log-likelihood) that measures the difference between the observations' labels  $y_i$  and the predictive model's predictions  $\hat{y}_i = f^*(x_i)$ , and  $\Omega$  is a regularization term that penalizes the complexity of the regression trees making up the ensemble in order to avoid overfitting<sup>2</sup> [40,46]. In Equation (1), the regression trees are written  $\tau_k$ , for  $k = 1, 2, 3 \dots K$ , and  $m_{train}$  refers to the number of observations making up the training dataset. The XGBoost algorithm builds a classification model from sequentially added regression trees, each of which is focused on the observations that the previously added trees classified incorrectly [46–48]. Given an initial prediction  $\tau_0$  (e.g. the logarithm of the odds), the accuracy of the ensemble model is iteratively improved by functional gradient descent: each newly added regression tree is parameterized to approximate the negative gradient of the loss function  $\mathcal{L}$  [46]. In order to avoid overfitting, the contribution of each newly added regression tree is weighted using a shrinkage parameter  $\nu$ , with  $0 < \nu < 1$  ( $\nu = 0.3$  in our case studies), which determines the learning rate of the boosting procedure [40,41,49]. In our automated biomarker candidate discovery workflow, the XGBoost learning process is stochastic because the regression trees making up the ensemble are learned on randomized subsamples of the training set, and because the features used for node splitting are chosen among a random subset of features [40,49]. The idea is to randomly subsample the rows and columns of the data matrix during training to make each regression tree slightly different from the other regression trees, and hence prevent overfitting.

$$f(x_i) = \log\left(\frac{p(x_i)}{1 - p(x_i)}\right) = \tau_0(x_i) + \sum_{k=1}^K \nu \tau_k(x_i) \quad (2)$$

$$\text{with } p(x_i) = \text{Pr}(y_i = +1|x_i)$$

Equation (2) defines the raw output  $f(x_i)$ , or raw untransformed margin value, of an XGBoost classifier as the (natural) logarithm of the odds, called the log-odds [49,50]. The odds are defined as the ratio of the probability  $p(x_i)$  of observation  $x_i$  being assigned to the positive class over the probability of observation  $x_i$  being assigned to the negative class. The XGBoost classifier is an additive logistic regression model because it represents the log-odds as a linear combination of regression trees [45,47]. The probability of the

<sup>2</sup> Overfitting is a common issue that occurs when a regression or classification model adapts too closely to the training data and memorizes not only the relationship between inputs and outputs but also the noise. The structure of an overfit model is very sensitive to changes in its training data. Such a model usually performs poorly on new data. The risk of overfitting tends to increase when handling high-dimensional data like imaging mass spectrometry data.



model predicting a positive outcome (i.e. assigning an observation to the class of interest) can be obtained from the log-odds thanks to a logistic transformation [45]:  $p(x_i) = S(f(x_i))$  where  $S: \mathbb{R} \mapsto [0, 1]$  is the sigmoid function. Since the sigmoid function is a non-decreasing saturation function, an increase in the log-odds implies an increase in the probability of predicting a positive outcome, and, conversely, a decrease in the log-odds implies a decrease in the probability of predicting a positive outcome. The classification model's prediction  $\hat{y}_i = f^*(x_i)$  is either +1 and -1 depending on whether  $p(x_i)$  is above or below a given threshold  $\eta$ , with  $0 < \eta < 1$  ( $\eta = 0.5$  in our case studies).

## 2.2. Shapley additive explanations for measuring biomarker candidate relevance

Our workflow for biomarker candidate discovery in IMS data uses Shapley additive explanations (SHAP) to quantify the local and global predictive importance of features (e.g.  $m/z$  values in IMS) with respect to a given classification task. SHAP is a state-of-the-art interpretability method based on Shapley values from cooperative game theory. It regards the features as players that form coalitions (i.e. ordered subsets) to achieve the classification or regression model's output, which is the game's payout. SHAP is called a model-agnostic interpretability method because it can derive post-hoc explanations for the predictions of any type of classification model by relating its input to its outputs [19,20,26]. SHAP was developed by Lundberg and Lee [51,52] based on the work of Strumbelj and Kononenko [53,54], and on Ribeiro et al.'s idea of locally approximating the decision-making process of a "black-box" supervised ML model using inherently interpretable local surrogate models [55].

In order to explain the prediction made by a model on a specific observation, SHAP computes the contribution of each feature to the model's output using Shapley values. The Shapley value of a feature is its contribution to the model's output for a specific observation, averaged over all possible feature orderings [52,56]. In the words of Lundberg et al., "Shapley values are computed by introducing each feature, one at a time, into a conditional expectation function of the model's output, and attributing the change produced at each step to the feature that was introduced" [56]. Equation (3) defines the Shapley value  $\phi_i^j(f, x_i)$  of feature  $x^j$ , with  $j \in \{1, 2, 3, \dots, n\}$ , when explaining the predictive model's decision-making process for one specific observation  $x_i = X_{(i, \cdot)}$ , with  $i \in \{1, 2, 3, \dots, m\}$ . Since a feature's contribution to the model's output depends on the order in which other features were introduced, the feature's Shapley value is obtained by averaging its contribution over all possible feature orderings. In Equation (3), the set of all possible feature orderings is written  $\Pi$ . The set of features that we are conditioning on, written  $S_j^\pi$ , is the set of all features that precede feature  $x^j$  in ordering  $\pi$ .

$$\phi_i^j(f, x_i) = \sum_{\pi \in \Pi} \frac{1}{n!} [f_{x_i}(S_j^\pi \cup j) - f_{x_i}(S_j^\pi)] \quad (3)$$

The set function  $f_{x_i}(S_j^\pi)$  is defined by Equation (4) as the conditional expectation of the model's output. The  $n$ -dimensional vector  $x_i$  is considered to be a random variable where only the features belonging to subset  $S_j^\pi$  (i.e. the features before  $x^j$  in the feature ordering  $\pi$ ) are known [57]. The unknown features (i.e. the features after  $x^j$  in the feature ordering  $\pi$ ) are obtained by sampling from the training dataset [57]. In Equation (4),  $f(x_i)$  is the model's raw output for observation  $x_i$ , rather than the predicted class label  $f^*(x_i) \in \{-1, +1\}$ . SHAP does not require knowledge of an observation's true class label ( $y_i$  for  $i \in \{1, 2, 3, \dots, m\}$ ) to evaluate the degree to which a

model depends on a specific feature. SHAP can therefore be used to explain the decision-making process of a model on new unlabeled data, which is useful for measuring the influence of each feature on the model's generalization performance.

$$f_{x_i}(S_j^\pi) = E[f(x_i) | S_j^\pi] \quad (4)$$

The sign of a feature's Shapley value provides information about the direction of its effect on a predictive model's output. A positive Shapley value indicates that feature  $x^j$  increases the raw output  $f(x_i)$  of the predictive model for observation  $x_i$ . Conversely, a negative Shapley value indicates that feature  $x^j$  decreases the raw output. The Shapley value's magnitude indicates how strongly the corresponding feature influences the model's local decision-making process. In our work, we refer to the Shapley value of a feature for a given observation as its local SHAP importance score. In the context of IMS data classification, the Shapley value  $\phi_i^j(f, x_i)$  reports the contribution of the  $j^{\text{th}}$   $m/z$  bin or peak when assigning the  $i^{\text{th}}$  pixel's mass spectrum to a class. Computing the local SHAP importance scores of all features (i.e.  $m/z$  values) for all observations (i.e. mass spectra) yields an  $m \times n$  matrix whose  $(i, j)^{\text{th}}$  entry is the Shapley value of feature  $x^j$  for observation  $x_i$ .

SHAP owes its reliability to the fact that it satisfies the local accuracy and consistency properties [56]. The local accuracy property, also known as the efficiency property in cooperative game theory, guarantees that the Shapley values of all features add up to the difference between the predictive model's raw output  $f(x_i)$  for a given observation  $x_i$  and the model's expected output  $E[f(x_i)]$  over the entire dataset [56]. The local accuracy property is given by Equation (5). SHAP offers contrastive explanations that compare the model's local output to its average global output. In IMS terminology, the local accuracy property states that, given a mass spectrum of interest, the sum of the Shapley values of its molecular features (i.e.  $m/z$  values) is equal to the classifier's raw output for that mass spectrum minus the classifier's average raw output over all mass spectra. SHAP distributes the difference between the classifier's output for a mass spectrum of interest and the classifier's average output, among the different  $m/z$  values that the classifier uses as inputs.

$$f(x_i) = \varphi_0(f) + \sum_{j=1}^n \phi_i^j(f, x_i) \text{ with } \varphi_0(f) = E[f(x_i)] \quad (5)$$

The consistency property, also known as the monotonicity property in cooperative game theory, states that if a model changes so that some feature's influence on the output increases, the importance score assigned to that feature does not decrease [58]. Consistency is necessary for the ranking of a model's features according to their importance scores because it guarantees that a feature with a higher importance score than another feature is actually more important to the model than the other feature. Note that impurity-based measures of global feature importance, which are popular for measuring feature importance in decision tree ensembles<sup>3</sup> and have been used in IMS [34], are actually inconsistent and can therefore produce unreliable feature rankings [58].

<sup>3</sup> Gini importance, which is the default measure of feature importance in Scikit-Learn's implementation of random forest [59] and gain, which is the default measure of feature importance in the Scikit-Learn wrapper interface for XGBoost's implementation of extreme gradient boosting [60], are two examples of popular yet inconsistent impurity-based approaches for estimating features' predictive importances. The Gini importance of a feature is computed by averaging the weighted decrease in node impurity achieved by splitting a node using that feature over all decision trees making up the ensemble [61].

$$\Phi_j(f, X) = \frac{1}{m} \sum_{i=1}^m |\varphi_i^j(f, x_i)| \quad (6)$$

A global measure of feature importance can be obtained by averaging the magnitude of each feature's local SHAP scores, or Shapley values, over all observations in the dataset [56]. Equation (6) defines what we refer to as the global SHAP score  $\Phi_j$  of feature  $x^j = X_{(:,j)}$  for  $j \in \{1, 2, 3, \dots, n\}$ . The global SHAP score of a feature quantifies its influence on the model's decision-making process, averaged over all possible feature orderings and all observations. Computing the global SHAP importance scores of all features yields an  $n$ -dimensional vector where  $n$  is the total number of features. In the context of IMS data analysis, the global SHAP score of a feature is an experiment-wide measure of the feature's predictive importance with respect to a given classification task. Promising biomarker candidates can be easily identified by ranking IMS features (i.e.  $m/z$  values) in descending order of global SHAP importance. Retaining the top-ranking features yields a shortlist of biomarker candidates that are worthy of further study.

In our workflow for biomarker candidate discovery in IMS data, we use a fast implementation of SHAP called TreeSHAP, or TreeExplainer [62], that is specific to decision tree based predictive models like XGBoost. Unlike other SHAP implementations (e.g. KernelSHAP) that calculate sampling-based approximations of Shapley values (often in exponential time), TreeSHAP is able to compute the exact Shapley values of features within low-order polynomial time by exploiting the structure of decision trees [56,58]. When using TreeSHAP to measure the local and global SHAP importance scores of features, one has to choose between two feature perturbation approaches [62]. In this paper, we opt for the tree-path dependent approach because it involves computing the observational, rather than interventional, Shapley values [57]. Observational Shapley values are defined by Equations (3) and (4), whereas interventional Shapley values define the set function differently. The difference between observational and interventional Shapley values relates to how SHAP handles statistical dependencies between the features that the model uses as inputs [57]. Accounting for high-dimensional feature dependencies is what makes measuring the predictive importance of IMS features, many of which are involved in common biochemical processes, particularly challenging. Another measure of global feature predictive importance, called permutation importance (PI), has been used for ranking IMS features with regards to tissue classification tasks [17] despite it only partially accounting for feature inter-dependencies. PI defines the importance of a feature as the average decrease in model accuracy when its values are randomly permuted across all observations<sup>4</sup>. The feature importance scores delivered by PI can be misleading when the classifier's features exhibit statistical dependencies [65]. Unlike PI, SHAP accounts for linear and non-linear feature inter-dependencies [56]. Furthermore, PI relies upon out-of-distribution data instances that are not necessarily realistic [66], whereas computing the global SHAP score of a

<sup>4</sup> Permutation importance was originally developed by Breiman, under the name of mean decrease accuracy, as a model-specific method for measuring feature importance in random forests [63]. The idea is to randomly permute a feature across all observations to break its association with the model prediction (and the other features) and effectively cancel its predictive power [64]. Therefore, if the feature under study is strongly associated to the prediction, permuting its values should result in a large drop in predictive performance. Conversely, if the feature is weakly associated to the prediction, permuting it should have little to no impact on performance.

feature using observational Shapley values constrains the sampling of unknown features to a range of values (i.e. partitions of the feature space) allowed by the decision trees making up the ensemble [64]. A detailed discussion of how TreeSHAP computes observational Shapley values, and how observational Shapley values handle feature dependencies, is beyond the scope of this paper, and we therefore refer the reader to Refs. [56–58]. Observational Shapley values are recommended for knowledge discovery in biology and chemistry because they spread credit among correlated features that are jointly informative of the outcome of interest [57].

### 2.3. SHAP maps for a spatial understanding of a classifier's decision-making process

In addition to automatically establishing an experiment-wide biomarker candidate shortlist by means of global SHAP score ranking, we furthermore introduce a novel spatially-aware representation of local SHAP-based explanations, called a SHAP map. The SHAP map of a molecular feature is obtained by plotting that feature's local SHAP importance scores, or Shapley values, across all pixels. SHAP maps facilitate a spatially localized understanding of a classifier's decision-making process. In the context of biomarker candidate discovery, SHAP maps provide one with a nuanced and location-specific (e.g. cell type specific, tissue region specific) view into a molecular species' biomarker potential. Fig. 2 illustrates how SHAP maps can complement feature rankings for the purpose of biomarker candidate discovery. Unlike global SHAP importance scores, local SHAP importance scores avoid conflating the magnitude of the feature's effect with the prevalence of its effect across the sample surface area.

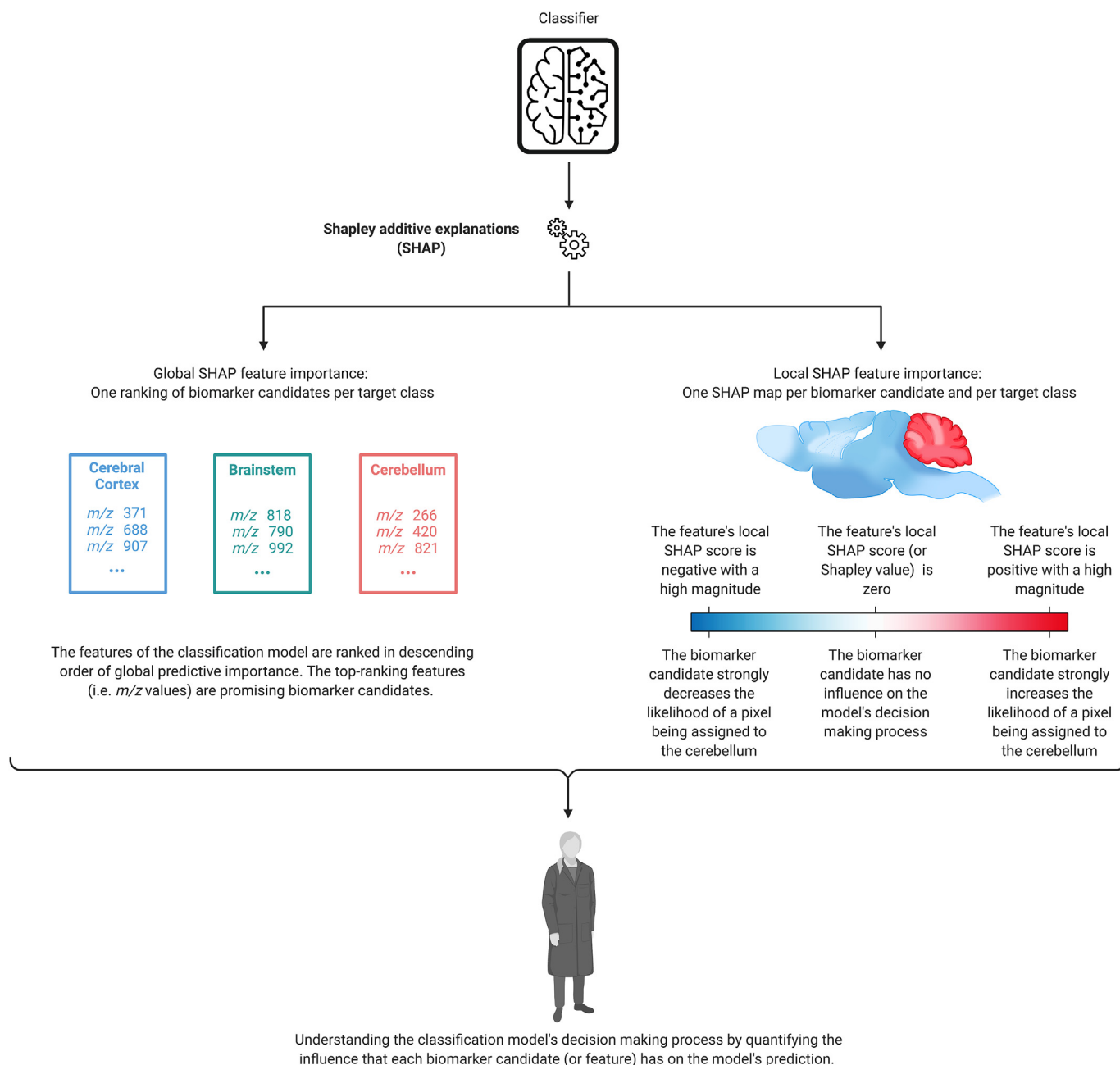
The SHAP map of a feature answers the following two questions:

- Where does the feature increase or decrease the classifier's output?

The feature increases the probability of the classifier assigning a pixel to the class of interest (i.e. the positive class) where its local SHAP scores are positive (red pixels). The feature decreases the probability of the classifier assigning a pixel to the positive class where its local SHAP scores are negative (blue pixels). In our application, studying the sign of a feature's Shapley values together with the feature's spatial distribution (e.g. the feature's ion image) enables the user to determine whether it is the presence or the absence of a feature that is indicative of the biological state or disease condition of interest. If the regions where the feature's measured intensity is high coincide with the regions where the feature's Shapley values are positive, the feature's presence is indicative of the class of interest. The relationship between the feature's abundance and the class prediction is correlative. Conversely, if the regions where the feature's measured intensity is low coincide with the regions where the feature's Shapley values are positive, the feature's absence is indicative of the class of interest. The relationship between the feature's abundance and the class prediction is anticorrelative.

- Where does the feature strongly or weakly influence the classifier's output?

The feature has a relatively large influence on the classifier where its Shapley values have a high magnitude (pixels with high saturation). Conversely, the feature has a relatively small influence on the model where its Shapley values have a low magnitude (pixels with low saturation). Studying the magnitude of a feature's Shapley values provides insight into how large or small the



**Fig. 2.** Diagram of the classifier interpretation process. SHAP is used to measure the local and global predictive importance of the features that the classifier from Fig. 1 uses to assign the pixels making up the sample surface (and their corresponding mass spectra) to one of four different anatomical classes (cerebral cortex, cerebellum, brainstem, or other). The global SHAP scores provide an experiment-wide measure of each biomarker candidate's relevance, whereas the local SHAP scores measure the direction and magnitude of each biomarker candidate's influence on the model output for one single pixel. SHAP maps deliver spatially localized explanations of the classifier's decision-making process. Icons from Refs. [21,22].

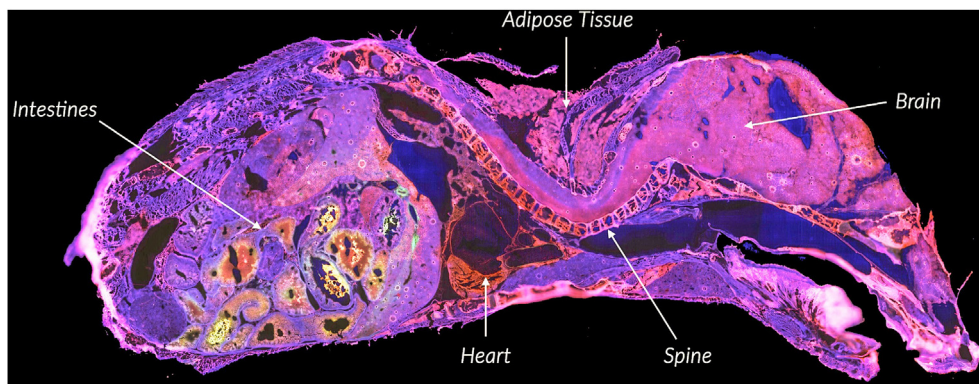
feature's local influence on a classifier is. In our application, we consider a feature (i.e. *m/z* value) to be relevant to recognizing a given class in the regions of the sample where its Shapley values have a high magnitude.

### 3. Results & discussion

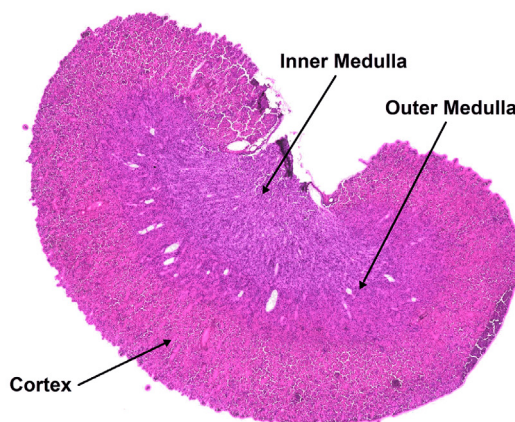
Our biomarker candidate discovery workflow is demonstrated on two IMS datasets that were acquired by matrix assisted laser desorption/ionization (MALDI) quadrupole time-of-flight (Q-TOF) IMS using a prototype MALDI timsToF Pro (Bruker Daltonics,

Germany) in positive ion mode [67]. Please refer to the supplementary material for information regarding the materials, sample preparation, experiments, histology, and IMS data preprocessing. Since the following five case studies do not involve the study of diseased tissue, the ranked features are not indicative of any pathological processes but rather of anatomical structures. Therefore, the term "molecular marker" is preferred over the term "biomarker" in Section 3. It should be noted that, methodologically speaking, there is no difference: in both cases our workflow looks for differentiating markers (i.e. *m/z* values) that correspond to user-provided classes of interest.





**a:** Autofluorescence microscopy image of the mouse-pup whole-body section from which imaging mass spectrometry dataset n°1 is obtained.



**b:** Hematoxylin & eosin stained microscopy image of the rat kidney tissue sample from which imaging mass spectrometry dataset n°2 is obtained.

**Fig. 3.** Microscopy images of the tissue sections imaged in IMS datasets n°1 and n°2.

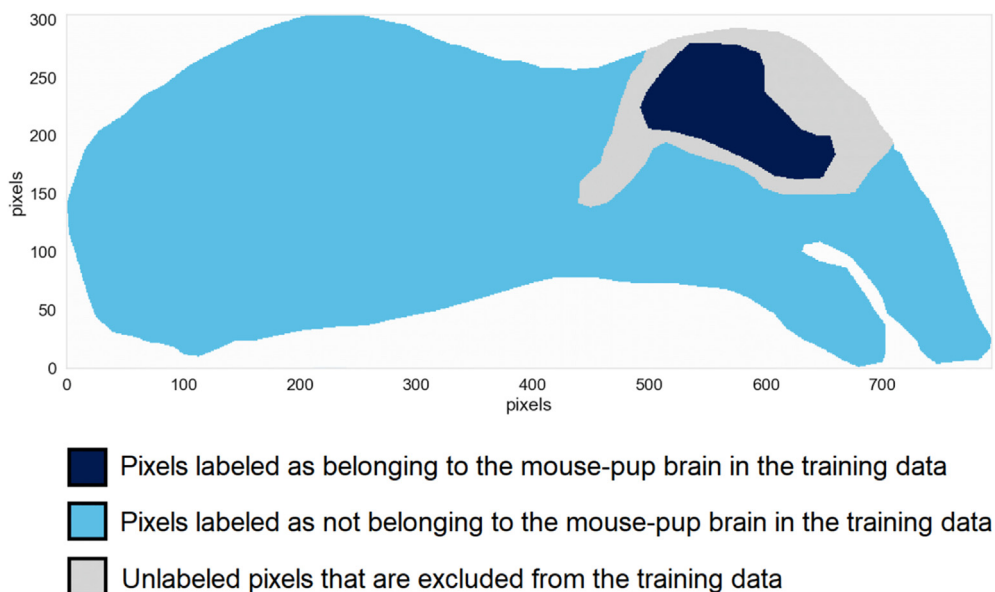
- Dataset n°1 was acquired from the sagittal whole-body section of a mouse-pup. The autofluorescence microscopy image of the tissue section is presented in Fig. 3a and was used to guide annotation of the regions of interest [68]. The sample was cryosectioned at 20  $\mu\text{m}$  thickness and a 1,5-diaminonaphthalene matrix was applied by sublimation. The mean mass spectrum of the dataset was retrieved and peak-picked to produce a feature list of 879 distinct ion species. The  $m/z$  acquisition range is 300-1200 and the pixel size is 50  $\mu\text{m} \times 50 \mu\text{m}$ . The dataset consists of a total of 164,808 pixels. Our workflow is therefore applied to a dataset of 164,808 observations and 879 features. The challenge of molecular marker discovery in the two case studies tied to this dataset therefore amounts to automatically determining which molecular species, among the 879 measured  $m/z$  values, are most relevant to recognizing two anatomical regions: the mouse-pup's brain and its liver.
- Dataset n°2 was acquired from the sagittal section of a rat kidney. The hematoxylin & eosin stained microscopy image of the tissue section is presented in Fig. 3b. The sample was cryosectioned at 12  $\mu\text{m}$  thickness and a 1,5-diaminonaphthalene matrix was applied by sublimation. The mean mass spectrum of the dataset was retrieved and peak-picked to produce a feature list of 1428 distinct ion species. The  $m/z$  acquisition range is 300-2000 and the pixel size is 15  $\mu\text{m} \times 15 \mu\text{m}$ . Our workflow is applied to a data table of 591,534 observations and 1428

features. The challenge of molecular marker discovery amounts to automatically determining which molecular species, among the 1428 measured  $m/z$  values, are most relevant for recognizing three different regions of the kidney: the cortex, the inner and outer medulla.

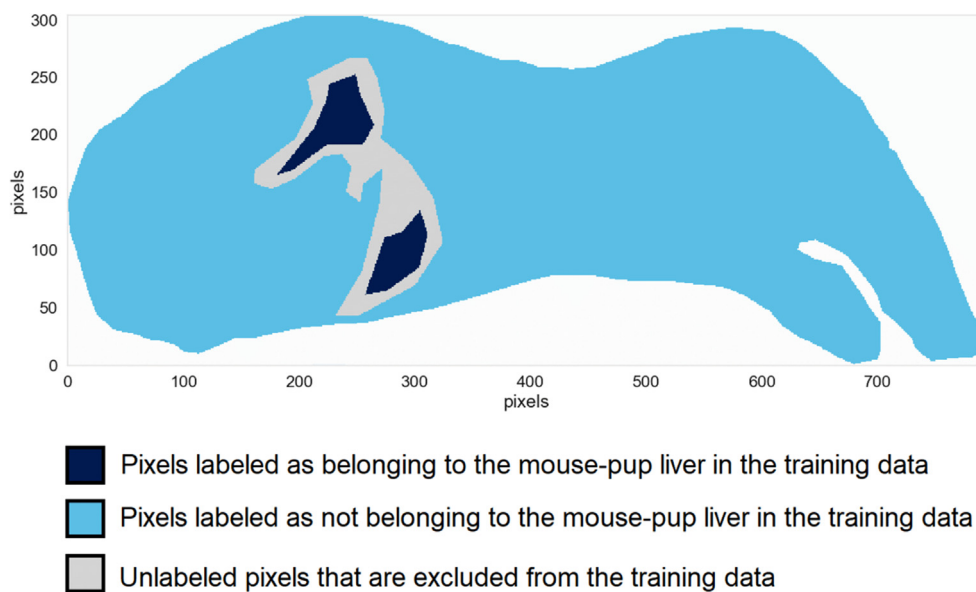
Five anatomical regions were delineated within the two tissue samples on the basis of the microscopy images in Fig. 3. Each tissue region was given a class label: a brain and a liver region in dataset n°1 and a cortex, inner medulla, and outer medulla region in dataset n°2. Our aim is to discover molecular markers for each of these user-provided class labels. The molecular marker discovery is treated separately for each class, using the one-versus-all procedure, yielding five binary classification problems whose target (i.e. positively labeled) classes are the following: the mouse-pup's brain and liver in dataset n°1; the rat kidney's inner medulla, outer medulla, and cortex in dataset n°2. Note that, although user-defined masks are employed to label the data in our case studies, our approach would work equally well if provided with automatically generated class annotations (e.g. tissue segmentation masks generated by clustering algorithms, masks provided by algorithms performing automated recognition in microscopy images). Two case studies (i.e. discovering molecular markers for the mouse-pup's brain in dataset n°1, discovering molecular markers for the rat kidney's inner medulla in dataset n°2) are covered in Section 3,

whereas the remaining three case studies (i.e. discovering molecular markers for the mouse-pup's liver in dataset n°1, discovering molecular markers for the rat kidney's outer medulla and cortex in dataset n°2) are provided in the supplementary material.

As discussed in [subsection 2.1](#), XGBoost models are used to classify the pixels on the basis of their mass spectra. These five classification problems are imbalanced because their corresponding datasets have unequal class cardinality (i.e. the negatively



**a:** Masks used for building the XGBoost classification model designed to recognize the mouse-pup brain.



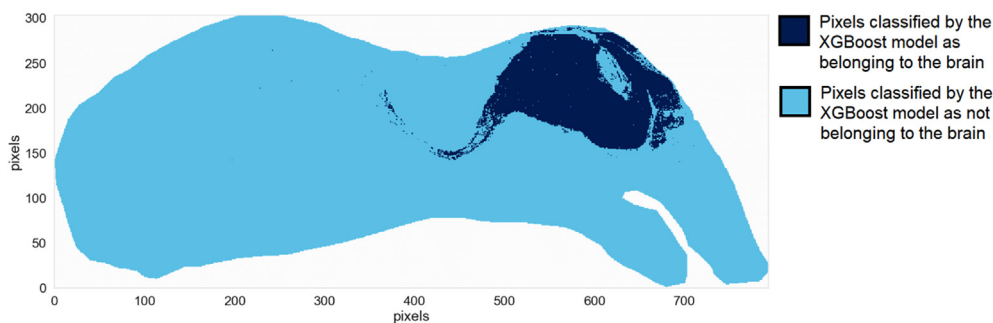
**b:** Masks used for building the XGBoost classification model designed to recognize the mouse-pup liver.

**Fig. 4.** Class-defining masks used as inputs for training the two XGBoost classifiers designed to recognize the mouse-pup brain and liver. For each task, regions of the tissue sample were manually annotated as belonging to one of three categories: dark blue pixels are labeled as belonging to the target organ and make up the positive class, light blue pixels are labeled as not belonging to the target organ and make up the negative class, and gray pixels are close to borders between the target organ and other anatomical structures, making it difficult to annotate them definitively. The latter are therefore excluded from the training data to avoid feeding the supervised machine learning algorithm unreliable annotations during classifier training. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

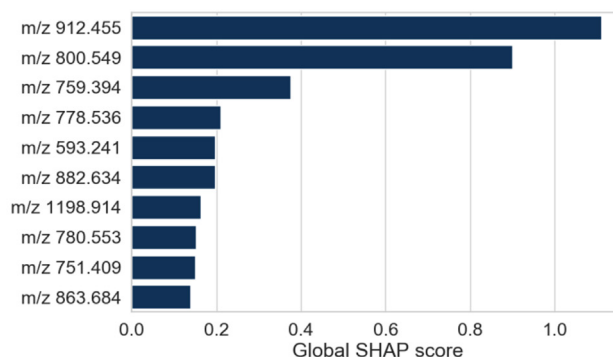
labeled pixels outnumber the positively labeled pixels). We avoid using accuracy (i.e. the proportion of predictions that are correct) to measure the classifier's predictive performance since accuracy tells us little about whether false negatives or false positives are more common [69]. Instead, we choose to measure our classifier's predictive performance using balanced accuracy, precision, and recall. Recall (also called sensitivity or the true positive rate) is the proportion of positive observations that are correctly identified. Precision is the proportion of all positive predictions that are correct. Specificity (also called the true negative rate) is the proportion of negative observations that are correctly identified. Balanced accuracy is the arithmetic mean of sensitivity and specificity [69]. As discussed in subsections 2.2 and 2.3, the TreeSHAP implementation of the SHAP interpretability method (with observational Shapley values) is used to rank the features (i.e.  $m/z$  values) in descending order of global predictive importance. The top-ranking features are highly discriminative with regards to a labeled tissue class and are therefore considered to be promising molecular markers for that class of interest. In addition to automatically establishing a shortlist of molecular species that are statistically related to user-provided tissue class labels, our workflow delivers spatially localized insight into the relationship (e.g. correlative, anticorrelative) between each measured ion species and the class of interest by means of a novel visualization approach called SHAP maps.

### 3.1. Dataset n<sup>o</sup>1: Recognition of the brain and liver of a mouse-pup

Classification-oriented supervised ML algorithms require labeled training data (in our case, labeled pixels) to build a classifier. In the two mouse-pup case studies, anatomical class labels are obtained as user-provided spatial delineations of the mouse-pup's brain and liver in the tissue sample. Exploratory analysis of the IMS data was performed using non-negative matrix factorization to aid in that delineation task [16,70,71]. Please refer to the supplementary material to see how the low-dimensional latent patterns extracted by non-negative matrix factorization from dataset n<sup>o</sup>1 facilitated the manual localization and annotation of the target organs. The target organs (or tissue regions, cell types, or cells) that are provided as masks to the supervised ML algorithm are also the organs (or tissue regions, cell types, or cells) for which we want to discover molecular markers. Fig. 4 shows a spatial representation of the masks used to build the XGBoost classifiers for the mouse-pup cases. Pixels are either annotated as belonging to the target organ (i.e. positive class) or not belonging to the target organ (i.e. negative class). Some pixels (e.g. at the borders of target organs) were difficult to annotate definitively and were excluded from the training set to avoid providing the supervised ML algorithm erroneous or unreliable training examples. Furthermore, the negative class was downsampled to avoid the one-versus-all classification of the brain and liver being severely imbalanced. After downsampling,



**a:** Brain class prediction in the mouse-pup whole-body section. The class labels are predicted by the classifier designed to differentiate the pixels belonging to the mouse-pup's brain from the pixels not belonging to the mouse-pup's brain.



**b:** Global feature importance of the ten  $m/z$  values most relevant to recognizing brain tissue in the mouse-pup, as per the SHAP interpretability method. The features of the XGBoost classifier trained to recognize the mouse-pup's brain tissue are ranked in descending order of global feature importance. The global SHAP score of a feature is obtained by averaging the magnitude of the features' Shapley values, or local SHAP scores, over all labeled and unlabeled pixels in dataset n<sup>o</sup>1. The ten  $m/z$  values shown here score the highest in terms of differentiating relevance, and automatically provide a shortlist of molecular markers for mouse-pup brain tissue.

**Fig. 5.** Mouse-pup brain recognition and global feature ranking.



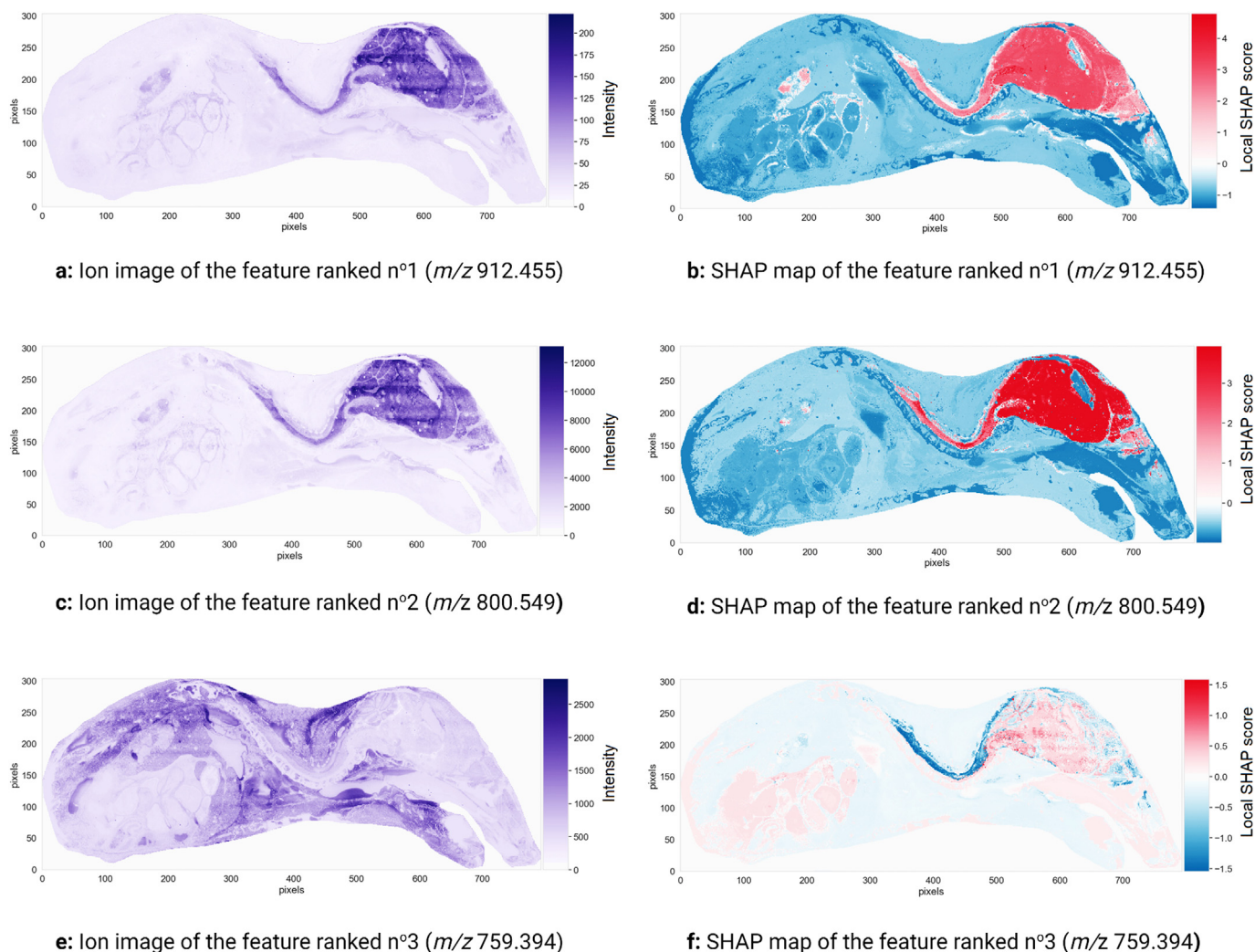
approximately 25% of the pixels used to build the classifier belong to the positive class, and approximately 75% belong to the negative class. Please refer to the supplementary material for the mouse-pup liver case study.

### 3.1.1. Molecular marker discovery for the mouse-pup brain

Our brain molecular marker discovery workflow starts with building an XGBoost classifier from IMS dataset n°1 and the user-provided brain mask shown in Fig. 4a. Fig. 5a is the classification result obtained by supplying the classifier with all IMS measurements (both labeled and unlabeled pixels), and having it automatically recognize brain tissue pixels on the basis of their mass spectra. Fig. 5a shows which mouse-pup tissue regions are predicted to belong to the brain according to the XGBoost classifier. As is apparent from Fig. 5a, the mouse-pup's brain (as well as parts of its spinal cord) is successfully differentiated from the other organs. Since an explanation can only be as good as its underlying model, it is necessary to verify the generalization performance of the XGBoost classifier designed to recognize pixels belonging to the mouse-pup brain. The classifier achieves a balanced accuracy of 0.9925, a precision of 0.9967, and a recall of 0.9974 on the testing

dataset, which is a labeled selection of pixel measurements that is distinct from the training dataset.

Fig. 5b shows the top ten molecular markers of the global ranking of 879 features (i.e.  $m/z$  values) obtained by TreeSHAP. The features are ranked in descending order of global SHAP score, and thus in descending order of relevance to brain tissue recognition, yielding a shortlist of molecular markers for mouse-pup brain tissue. Fig. 5b provides insight into a feature's global (i.e. tissue-wide) relevance to the recognition task of Fig. 5a. The ten top-ranking features of Fig. 5b are annotated further in Table S1 of the supplementary material, including tentative identifications. The spatially localized nature of IMS measurements together with the SHAP map representation developed above allows us to obtain tissue location specific insights into an  $m/z$  value's relevance. Fig. 6 shows the ion images and SHAP maps of the three top-ranking features of Fig. 5b. The left column of Fig. 6 displays the spatial distribution and relative abundance of the three top-ranking molecular features for recognizing the mouse brain. Fig. 6a, 6c, and 6e are ion images of the features ranked n°1, n°2, and n°3 respectively, and they are displayed using a pseudo-color scale whose brightness is indicative of the signal intensity measured at a given pixel. These ion images



**Fig. 6.** Three promising molecular markers for the mouse-pup's brain. The ion images (left) and SHAP maps (right) of three features (i.e.  $m/z$  values) with the most influence on the decision-making process of the classifier trained to recognize the mouse-pup's brain are shown. The ion images plot the spatial distribution and measured intensity of each feature across the sample, and are not specifically tied to the task of recognizing the brain. The SHAP maps plot the spatial distribution of Shapley values, or local SHAP predictive importance scores, of each feature across the sample, and provide information on where and how the feature is relevant to the task of recognizing brain.



provide a classical view on molecular distribution by reporting the ion intensity signal corresponding to the molecular species at hand. However, ion images do not provide any information about how that ion intensity relates to the recognition of brain tissue. The right column of Fig. 6 provides information on the signs and magnitudes of the local SHAP scores across the sample for each top-ranking feature. Fig. 6b, 6d, and 6f are the SHAP maps of the features ranked  $n^{\circ}1$ ,  $n^{\circ}2$ , and  $n^{\circ}3$  respectively. These SHAP maps provide information on where and how a given ion intensity signal relates to the task of brain tissue recognition.

Fig. 6a is the ion image of the feature ranked  $n^{\circ}1$ , whose  $m/z$  value is 912.455 and who has been tentatively identified as [SHexCer(38:6; O5)+K]<sup>+</sup>. The measured intensity of this feature is high in the brain and spinal cord. According to Fig. 6b, feature  $n^{\circ}1$  increases the log-odds (raw) output of the XGBoost classifier in the brain region: the Shapley values in the brain and spinal cord are positive, and negative elsewhere. The presence of feature  $n^{\circ}1$  increases the log-odds (and probability) of the classifier predicting that a given pixel belongs to the brain. The ion image and SHAP map of the feature ranked  $n^{\circ}2$  ( $m/z$  800.549) are very similar to those of the feature in first position. Both top-ranking features are positively correlated with the classifier assigning a pixel to the brain. Measuring high intensity signals for features  $n^{\circ}1$  and  $n^{\circ}2$  in a given pixel increases the log-odds (and probability) of the classifier assigning that pixel to the brain. The high predictive performance of the XGBoost classifier suggests that it is probably a good approximation of the data generating mechanism (i.e. biochemical processes taking place in the tissue). It can therefore be assumed and inferred that measuring a high intensity signal for features  $n^{\circ}1$  and  $n^{\circ}2$  in a given pixel also increases the probability of that pixel actually belonging to the brain. In other words, the presence of these features ( $m/z$  values) is characteristic of the mouse-pup's brain and spinal cord and differentiates the brain and spinal cord from other regions in the tissue.

Fig. 6e indicates that the feature ranked  $n^{\circ}3$ , whose  $m/z$  value is 759.394, has a low intensity both in the brain and the spinal cord. Its measured intensity in the spinal cord is slightly higher than its intensity in the brain. Fig. 6f shows that the Shapley values of that feature are negative in the spinal cord (with a magnitude between  $-1.0$  and  $-1.5$ ). The area highlighted (negatively, hence in dark blue) in Fig. 6f, namely the spinal cord, is where feature  $n^{\circ}3$

plays a role in helping to obtain a biomolecular signature unique to the brain. The way in which this feature helps the classifier correctly identify the brain pixels can be read from the sign of its Shapley values, or local SHAP scores. The local SHAP values in the spinal cord are negative, meaning that whatever the signal is that is measured for this feature in the spinal cord, it lowers the log-odds (and probability) of assigning a pixel to the brain. Studying the ion image of feature  $n^{\circ}3$  furthermore reveals that the ion intensity for  $m/z$  759.394 is low in the spinal cord, but still higher than in the brain. This means that a relative increase in signal intensity of  $m/z$  759.394 strongly decreases the log-odds (and probability) of predicting a pixel belonging to the brain. Unlike the features ranked  $n^{\circ}1$  and  $n^{\circ}2$  that are good molecular markers for both the brain and spinal cord, the feature ranked  $n^{\circ}3$  enables the XGBoost classifier to tell the brain apart from the spinal cord. We would not be able to differentiate the mouse's brain from its spinal cord if we were to use only the two top-ranking features ( $m/z$  912.455 and  $m/z$  800.549). This example illustrates the subtle understanding of molecular marker spatial specificity that can be obtained from SHAP maps. If one needs a molecular marker for both the brain and spinal cord, both  $m/z$  912.455 and  $m/z$  800.549 are good candidates. If one requires the ability to tell brain tissue apart from spinal cord tissue, a more elaborate panel of molecular markers is proposed: if  $m/z$  912.455 and  $m/z$  800.549 are present in high abundance in a tissue area, and if  $m/z$  759.394 is present in very low abundance, the probability of those pixels describing brain tissue (exclusively) is very high.

### 3.2. Dataset $n^{\circ}2$ : Recognition of renal inner medulla, outer medulla, and cortex

Annotating the three different functional tissue regions of the rat kidney - namely the inner medulla, outer medulla, and cortex - is required to generate the class labels needed to train the three corresponding XGBoost classifiers. Similar to the previous case studies (subsection 3.1), exploratory analysis by means of non-negative matrix factorization was used to aid in delineating masks. Fig. 7 shows the pixels annotated as belonging to one of the three target regions. The pixels that were difficult to annotate manually were excluded from the training and testing datasets. Similar to the previous case studies (subsection 3.1), downsampling

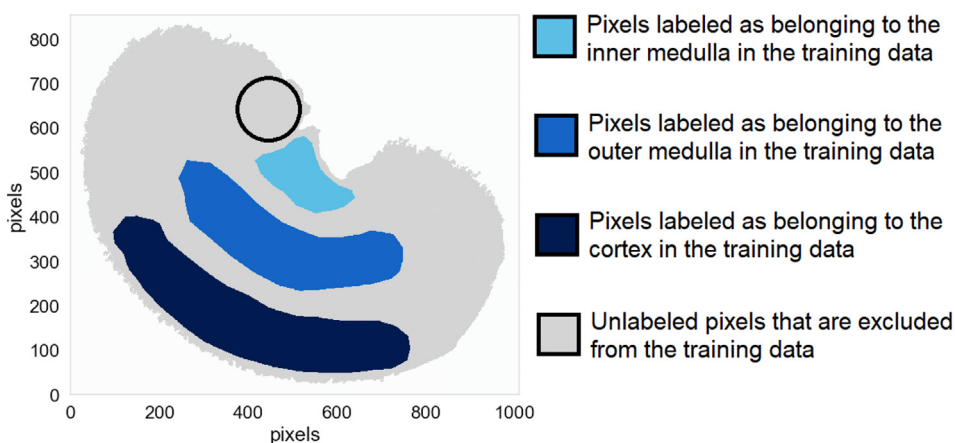
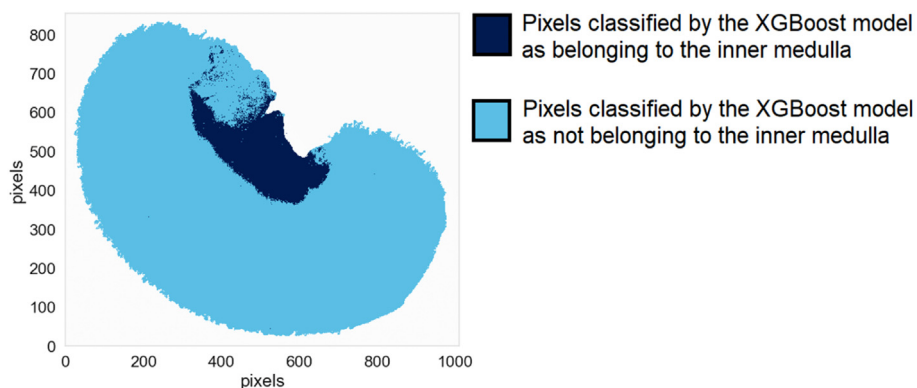
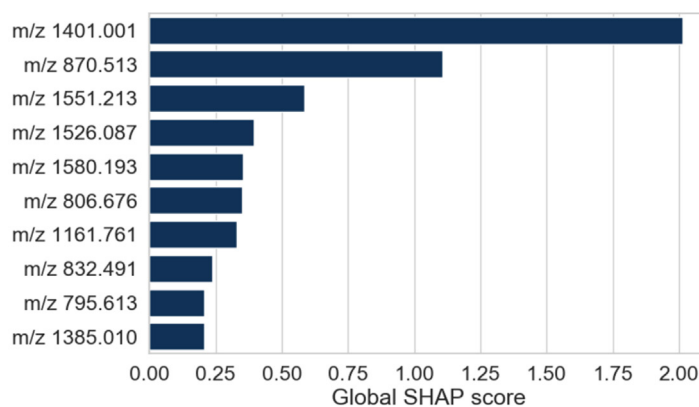


Fig. 7. Masks used as inputs for training the three XGBoost classifiers directed at recognizing the kidney's inner medulla, outer medulla, and cortex. Different regions of the tissue sample were manually annotated as belonging to one of four categories: light blue pixels belong to the inner medulla, medium blue pixels belong to the outer medulla, dark blue pixels belong to the cortex, and gray pixels are close to borders between these anatomical structures, making it difficult to annotate them definitively. The latter are excluded from the training data to avoid feeding the supervised machine learning algorithm unreliable annotations during classifier training. The black circle outlines a region of the renal cortex that was affected by a sample preparation artefact. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)



**a:** Inner medulla class prediction in the rat kidney section. The class labels are predicted by the classifier designed to differentiate the pixels belonging to the inner medulla from the pixels not belonging to the inner medulla.



**b:** Global feature importance of the ten  $m/z$  values most relevant to recognizing the renal inner medulla, as per the SHAP interpretability method. The features of the classifier trained to recognize the inner medulla tissue are ranked in descending order of global feature importance. The global SHAP score of a feature is obtained by averaging the magnitude of the features' Shapley values, or local SHAP scores, over all labeled and unlabeled pixels in dataset n°2. The ten  $m/z$  values shown here score the highest in terms of differentiating relevance, and automatically provide a shortlist of molecular markers for rat renal inner medulla tissue.

**Fig. 8.** Renal inner medulla recognition and global feature ranking.

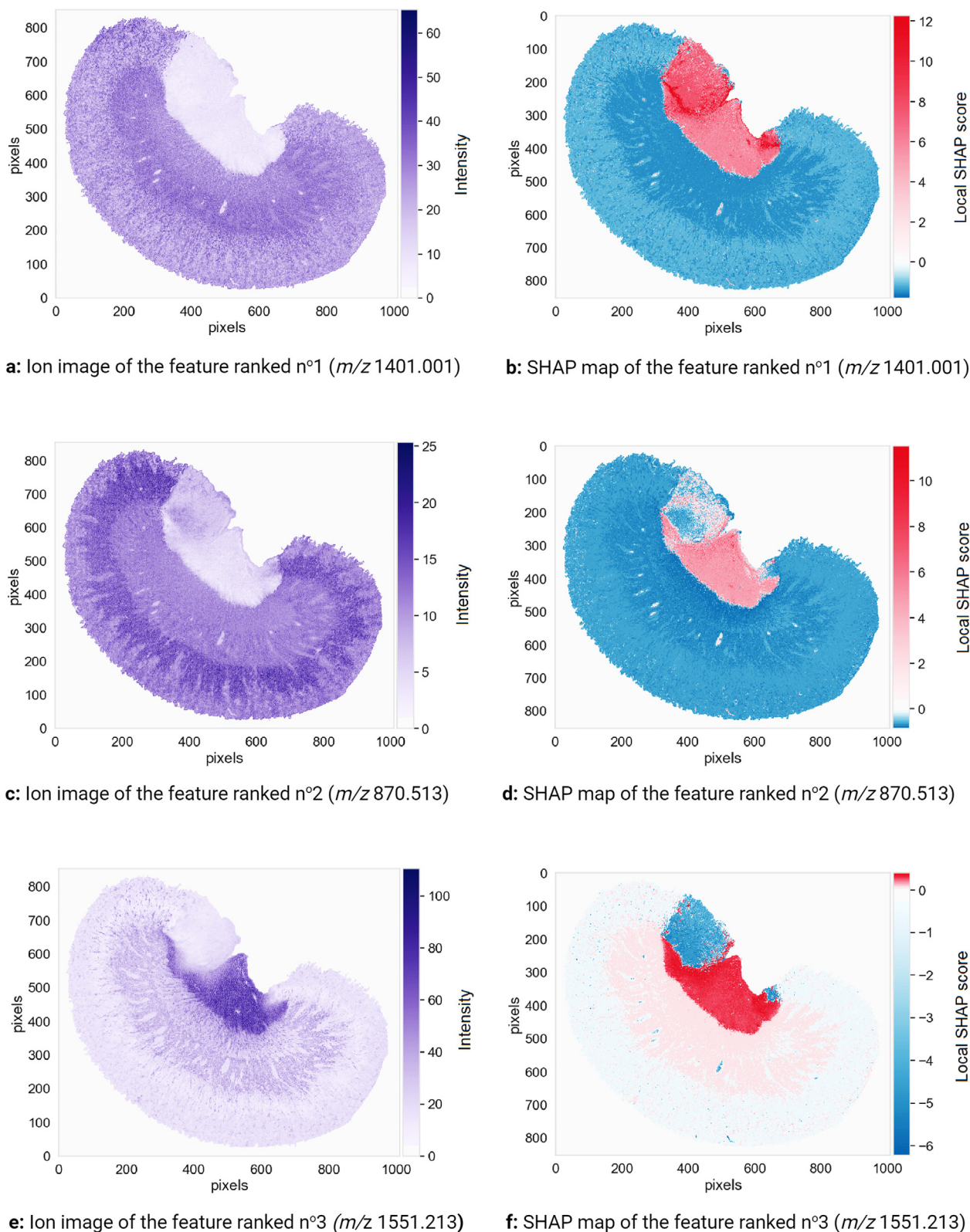
of the negative class was performed. The inner medulla, outer medulla, and cortex are differentiated from the other two regions using one-versus-all classification where the target region is the positive class, and the two other regions make up the negative class. Please refer to the supplementary material for the outer medulla and cortex case studies.

### 3.2.1. Molecular marker discovery for the renal inner medulla

Fig. 8a presents the class prediction result for the renal inner medulla classifier, showing that the inner medulla is successfully differentiated from the outer medulla and cortex. Regarding generalization performance, the XGBoost classifier trained to recognize pixels belonging to the inner medulla achieves a balanced accuracy of 0.9992, a precision of 0.9989, and a recall of 0.9985. Note the slightly noisy region to the top-left of the medulla in Fig. 8a. The difficulties encountered by the classifier in this region, which is outlined by a black circle in Fig. 7, are probably due to a sample preparation artefact known as visceral fat delocalization

[72]. Fig. 8b shows the top ten molecular markers out of the global ranking of 1428 features (i.e.  $m/z$  values) as obtained by TreeSHAP. The features are ranked in descending order of global SHAP score, and thus in descending order of relevance to inner medulla tissue recognition. Fig. 8b therefore provides insight into a feature's global relevance to the recognition task of Fig. 8a. The most important feature to the XGBoost classifier used to assign pixels to the inner medulla (or not) has a  $m/z$  value of 1401.001 and a global SHAP score of 2.036.

The left column of Fig. 9 provides information about the spatial distribution and relative abundance of the three top-ranking molecular features for recognizing the inner medulla: Fig. 9a, 9c, and 9e are the ion images of the features ranked n°1, n°2, and n°3 respectively. The right column of Fig. 9 provides information on the signs and magnitudes of the local SHAP scores, or Shapley values, of each top-ranking feature across the sample: Fig. 9b, 9d, and 9f are the SHAP maps of the three top-ranking features. These SHAP maps provide information on where and how a given ion intensity signal



**Fig. 9.** Three promising molecular markers for the renal inner medulla. The ion images (left) and SHAP maps (right) of three features (i.e.  $m/z$  values) with the most influence on the decision-making process of the classifier trained to recognize the rat's renal inner medulla are shown. The ion images plot the spatial distribution and measured intensity of each feature across the sample, and are not specifically tied to the task of recognizing the inner medulla. The SHAP maps plot the spatial distribution of Shapley values, or local SHAP predictive importance scores, of each feature across the sample, and provide information on where and how the feature is relevant to the task of recognizing the inner medulla.



relates to the task of inner medulla tissue recognition. Combining the left and right columns of Fig. 9 provides insight into the predictive model's decision-making process. The signal intensity measured in the inner medulla for features ranked  $n^{\circ}1$  and  $n^{\circ}2$  is low (Fig. 9a and 9c), and yet their Shapley values are high in the inner medulla (Fig. 9b and 9d). These features,  $m/z$  1401.001 and  $m/z$  870.513 respectively, are negatively correlated to the XGBoost classifier assigning a pixel to the inner medulla. In other words, measuring a low intensity for  $m/z$  1401.001 and  $m/z$  870.513 in a given pixel increases the log-odds (and probability) of the classifier assigning that pixel to the inner medulla. Given the high predictive performance of the classifier, we can assume that measuring a low intensity for  $m/z$  1401.001 and  $m/z$  870.513 in a given pixel also increases the probability of that pixel actually belonging to the inner medulla. Conversely, the feature ranked  $n^{\circ}3$  ( $m/z$  1551.213) is positively correlated with the classifier predicting a pixel as belonging to the inner medulla: its intensities (Fig. 9e) and its Shapley values (Fig. 9f) are both high in the inner medulla. Measuring a high intensity for  $m/z$  1551.213 in a given pixel increases the log-odds (and probability) of the classifier assigning that pixel to the inner medulla. Given the high predictive performance of the classifier, we can assume that measuring a high intensity for  $m/z$  1551.213 in a given pixel increases the probability of that pixel actually belonging to the inner medulla. The absence of features ranked  $n^{\circ}1$  and  $n^{\circ}2$ , and the presence of the feature ranked  $n^{\circ}3$  seem to be characteristic of renal inner medulla tissue.

We now focus on the tissue region to the top-left of the medulla that actually belongs to the cortex, and that was difficult for the XGBoost classifier to correctly differentiate from the inner medulla (see Fig. 8a). The SHAP map of the feature ranked  $n^{\circ}1$  shows (by coloring the difficult-to-classify area red) that this feature strongly increases the log-odds (and probability) of the cortex pixels to the top-left of the inner medulla being erroneously assigned to the inner medulla: the Shapley values of the feature ranked  $n^{\circ}1$  are positive with a high magnitude in this region of the cortex. The SHAP map of the feature ranked  $n^{\circ}3$  shows (by coloring the difficult-to-classify area blue) that the classifier uses this feature to correct for the labeling suggested by the feature ranked  $n^{\circ}1$ : the Shapley values of the feature ranked  $n^{\circ}3$  in the region to the top-left of the inner medulla are negative with a high magnitude. This case study demonstrates an interesting level of nuance in molecular marker discovery, uniquely provided by the SHAP map representation. If only the global SHAP scores of the features are taken into account (i.e. only the global information provided in Fig. 8b, without the localized information provided in Fig. 9b, 9d, and 9f), one might be tempted to consider  $m/z$  1401.001 (corresponding to the feature ranked  $n^{\circ}1$ ) as the most promising marker candidate for inner medulla tissue in this dataset. Although  $m/z$  1401.001 has the most influence on the XGBoost classifier designed to recognize the inner medulla, its global SHAP score is based on a sample-wide assessment of discriminative relevance and disregards subtle spatially localized patterns. In fact, Fig. 9b shows that  $m/z$  1401.001 has a positive influence on the classifier's prediction in the inner medulla but also in a region of the cortex where visceral fat delocalization probably occurred. Unlike  $m/z$  1401.001 and  $m/z$  870.513 (corresponding to the features ranked  $n^{\circ}1$  and  $n^{\circ}2$  respectively),  $m/z$  1551.213 (corresponding to feature ranked  $n^{\circ}3$ ) is exclusive to the inner medulla. Although the global SHAP scores of  $m/z$  1401.001 and  $m/z$  870.513 are higher than that of  $m/z$  1551.213 (respectively 2.036 and 1.007 versus 0.548), a localized study using SHAP maps shows that  $m/z$  1551.213 is the more reliable inner medulla molecular marker of the three because of its high spatial specificity. Unlike the signal of  $m/z$  1551.213, the signals corresponding to  $m/z$  1401.001 and  $m/z$  870.513 were affected by the sample preparation artefact that took place in the renal cortex. This

example also illustrates the importance of not basing one's estimate of a molecular marker candidate's relevance exclusively on its global SHAP predictive importance score. When visualized in the form of SHAP maps, the local SHAP scores (or Shapley values) provide useful spatially localized information as to how and where the molecular marker influences the predictive model's output and (assuming the classifier has good predictive performance) how it ties to the underlying tissue.

#### 4. Conclusion

In this work, we propose an innovative computational approach for automating the discovery of biomarker candidates in molecular imaging data. Our approach enables one to efficiently filter a multitude of molecular species down to a panel of promising biomarker candidates. Applying the automated biomarker candidate discovery workflow to imaging mass spectrometry (IMS) data is especially interesting because of the massively multiplexed nature of IMS. By enabling the untargeted concurrent mapping of hundreds to thousands of molecular species across a tissue sample, IMS enables one to cast a wide net for molecular species with biomarker potential. However, the wide range of candidates can pose difficulties since manual examination of IMS data is impractical. Automating biomarker candidate discovery in IMS using machine learning (ML) methodologies, rather than resorting to manual examination, can help re-establish the practical feasibility of IMS-based biomarker discovery, and can help maintain objectivity, scalability, and reproducibility. Our biomarker candidate discovery workflow produces a ranking of molecular species according to the discriminative relevance they hold for a given tissue structure or disease condition, such that the top-ranking molecular species are highly promising biomarker candidates that merit further study.

Our approach to biomarker candidate discovery is to identify highly discriminative molecular species whose overexpression or underexpression characterize a user-defined biological class of interest. A supervised ML algorithm, called extreme gradient boosting (XGBoost), is used to learn a classifier from labeled imaging mass spectrometry data, and a state-of-the-art ML model interpretability method, called Shapley additive explanations (SHAP), is used to measure the local and global predictive importance of the  $m/z$  values that the classification model uses as features. We translate the task of biomarker candidate discovery into a feature ranking problem: the features are ranked in descending order of global SHAP importance and the top-ranking features are retained for further investigation. The TreeSHAP implementation of Shapley additive explanations, with observational Shapley values, is used for quantifying the local and global predictive importance of features. In order to add nuance to our analysis, we furthermore introduce SHAP maps, a novel representation and visualization that brings a spatial dimension to our understanding of the decision-making processes of a classifier. The SHAP map of a feature is obtained by plotting that feature's local SHAP importance scores, or Shapley values, across all pixels making up the sample surface. A feature's local SHAP importance score is informative of the direction (e.g. positive or negative) and magnitude (e.g. large or small) of the feature's influence on the classifier's output for a given pixel. SHAP maps provide insight into the spatial specificity of biomarker candidates by showing how and where a feature influences the classifier's probability of assigning a pixel, and its corresponding mass spectrum, to the class of interest.

Although our two case studies concern imaging mass spectrometry data, our biomarker candidate discovery workflow is also applicable to other forms of multiplexed imaging data such as multiplexed fluorescence microscopy (e.g. CODEX), imaging mass



cytometry, near-infrared imaging, and Raman spectroscopic imaging, and therefore holds the potential to substantially advance biomarker development across a wide range of spectral imaging modalities. One area where our approach can be employed is in the discovery of clinically relevant molecular signatures for functional tissue units in the context of large-scale molecular mapping projects such as the NIH-sponsored Human BioMolecular Atlas Program [73], which aims to build a complete molecular map of the human body at single-cell resolution, and the Kidney Precision Medicine Project [74], which aims to build a comprehensive molecular, cellular, and anatomical map of the kidney. Our work on ML interpretability for multiplexed imaging may also help advance research in biomedical imaging, for example in the field of data-driven multi-modal image fusion [75], where a cross-modal regression model ties the observations in one imaging modality to the observations in another modality. Obtaining spatially-localized insight into how cross-modal connections are made holds potential for advancing all fusion applications, including prediction to a higher spatial resolution, out-of-sample predictions, as well as cross-modal denoising and relationship discovery [76].

## Funding

Research reported in this publication was supported by the National Institutes of Health (NIH)'s Common Fund, National Institute Of Diabetes And Digestive And Kidney Diseases (NIDDK), and the Office Of The Director (OD) under Award Number U54DK120058 (J.M.S., R.M.C., and R.V.), by NIH's Common Fund, National Eye Institute, and the Office Of The Director (OD) under Award Number U54EY032442 (J.M.S., R.M.C., and R.V.), by NIH's National Institute Of Allergy And Infectious Diseases (NIAID) under Award Numbers R01AI138581 and R01AI145992 (J.M.S. and R.V.), by NIH's National Institute of General Medical Sciences (NIGMS) under Award Number P41GM103391 (R.M.C.), and by the National Science Foundation Major Research Instrument Program CBET - 1828299 (J.M.S. and R.M.C.). The research was furthermore supported by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO; Dutch Research Council), ZonMw, FLAG-ERA, and the European Commission through the FLAG-ERA III JTC project SMART BRAIN (NWO \#680-91-319) under the NWO-domain Exacte en Natuurwetenschappen, together with the NWO-domain Sociale en Geesteswetenschappen, and in association with the European Commission's Human Brain Project (R.V.). SMART-BRAIN is a Partnering Project to the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 785907 (Human Brain Project SGA2). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, the Nederlandse Organisatie voor Wetenschappelijk Onderzoek, ZonMw, FLAG-ERA, or the European Commission.

## CRediT authorship contribution statement

**Leonoor E.M. Tideman:** Conceptualization, Software, Formal analysis, Writing – original draft, Writing – review & editing, Visualization. **Lukasz G. Migas:** Software, Data curation, Writing – review & editing. **Katerina V. Djambazova:** Investigation, Writing – review & editing. **Nathan Heath Patterson:** Software, Writing – review & editing. **Richard M. Caprioli:** Funding acquisition, Project administration, Resources. **Jeffrey M. Spraggins:** Supervision, Investigation, Writing – review & editing, Funding acquisition, Project administration, Resources. **Raf Van de Plas:** Supervision, Conceptualization, Writing – original draft, Writing – review & editing, Funding acquisition, Project administration, Resources.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.aca.2021.338522>.

## References

- [1] B.D.W. Group, Biomarkers and surrogate endpoints: preferred definitions and conceptual framework, *Clin. Pharmacol. Therapeut.* ISSN: 1532-6535 69 (3) (2001) 89–95, <https://doi.org/10.1067/mcp.2001.113989>.
- [2] N. Rifai, M.A. Gillette, S.A. Carr, Protein biomarker discovery and validation: the long and uncertain path to clinical utility., *Nat. Biotechnol.* ISSN: 1546-1696 24 (8) (Aug. 2006) 971–983, <https://doi.org/10.1038/nbt1235>.
- [3] C.A. Crutchfield, et al., Advances in mass spectrometry-based clinical biomarker discovery, *Clin. Proteomics.* ISSN: 1542-6416 13 (Jan. 2016), <https://doi.org/10.1186/s12014-015-9102-9>.
- [4] Z.-Z. Hu, et al., Omics-based molecular target and biomarker identification, Clifton, N.J.), *Methods Mol. Biol.* ISSN: 1064-3745 719 (2011) 547–571, [https://doi.org/10.1007/978-1-61779-027-0\\_26](https://doi.org/10.1007/978-1-61779-027-0_26).
- [5] M. Holzlechner, E. Eugenin, B. Prideaux, Mass spectrometry imaging to detect lipid biomarkers and disease signatures in cancer, *Canc. Rep.* ISSN: 2573-8348 2 (6) (2019) e1229, <https://doi.org/10.1002/cnr.2.1229>.
- [6] R.M. Caprioli, T.B. Farmer, J. Gile, Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS, *Anal. Chem.* ISSN: 0003-2700 69 (23) (Dec. 1997) 4751–4760, <https://doi.org/10.1021/ac970888i>.
- [7] K. Schwamborn, R.M. Caprioli, "MALDI imaging mass spectrometry – painting molecular pictures.", *Mol. Oncol. Thematic Issue: Oncoproteomics.* ISSN: 1574-7891 4 (6) (Dec. 2010) 529–538, <https://doi.org/10.1016/j.molonc.2010.09.002>.
- [8] S.S. Rubakhin, et al., Imaging mass spectrometry: fundamentals and applications to drug discovery, *Drug Discov. Today* 10 (12) (2005) 823–837, [https://doi.org/10.1016/s1359-6446\(05\)03458-6](https://doi.org/10.1016/s1359-6446(05)03458-6).
- [9] K. Schwamborn, M. Kriegsman, W. Weichert, MALDI imaging mass spectrometry - from bench to bedside, *Biochim. Et Biophysica Acta. Proteins Proteomics* 1865. ISSN: 1570-9639 7 (July 2017) 776–783, <https://doi.org/10.1016/j.bbapap.2016.10.014>.
- [10] M. Stoeckli, et al., Imaging mass spectrometry: a new technology for the analysis of protein expression in mammalian tissues, *Apr. Nat. Med.* ISSN: 1546-170X 7 (4) (2001) 493–496, <https://doi.org/10.1038/86573>.
- [11] T. Alexandrov, MALDI imaging mass spectrometry: statistical data analysis and current computational challenges, *S11, BMC Bioinf.* 13 (16) (Nov. 2012) 1471–2105, <https://doi.org/10.1186/1471-2105-13-S16-S11>.
- [12] T.A. Zimmermann, et al., Imaging of cells and tissues with mass spectrometry: adding chemical information to imaging, *Methods Cell Biol.* ISSN: 0091-679X 89 (2008) 361–390, [https://doi.org/10.1016/S0091-679X\(08\)00613-4](https://doi.org/10.1016/S0091-679X(08)00613-4).
- [13] M. Aichler, A. Walch, MALDI imaging mass spectrometry: current frontiers and perspectives in pathology research and practice, *Lab. Invest. J. Tech. Methods Pathol.* ISSN: 1530-0307 95 (4) (Apr. 2015) 422–431, <https://doi.org/10.1038/labinvest.2014.156>.
- [14] K. Schwamborn, Imaging mass spectrometry in biomarker discovery and validation, *J. Proteom.* ISSN: 1876-7737 75 (16) (Aug. 2012) 4990–4998, <https://doi.org/10.1016/j.jpro.2012.06.015>.
- [15] E. Robotti, M. Manfredi, E. Marengo, Biomarkers discovery through multivariate statistical methods: a review of recently developed methods and applications in proteomics, *J. Proteomics Bioinf.* ISSN: 0974-276X (2015) 1–20, <https://doi.org/10.4172/jpb.53-003>.
- [16] N. Verbeek, R.M. Caprioli, R. Van de Plas, Unsupervised machine learning for exploratory data analysis in imaging mass spectrometry, *Mass Spectrom. Rev.* (Oct. 2019), <https://doi.org/10.1002/mas.21602>.
- [17] M. Hanselmann, et al., Toward digital staining using imaging mass spectrometry and random forests, *J. Proteome Res.* ISSN: 1535-3893 8 (7) (July 2009) 3558–3567, <https://doi.org/10.1021/pr900253y>.
- [18] C. Molnar, G. Casalicchio, B. Bischl, Interpretable Machine Learning – A Brief History, State-Of-The-Art and Challenges, *arXiv:2010.09337 [cs, stat]*, Oct. 2020. [arXiv: 2010.09337 \[cs, stat\]](https://arxiv.org/abs/2010.09337).
- [19] V. Belle, I. Papantonis, Principles and practice of explainable machine learning, *arXiv:2009.11698 [cs, stat]*, Sept. 2020. [arXiv: 2009.11698 \[cs, stat\]](https://arxiv.org/abs/2009.11698).
- [20] M. Du, N. Liu, X. Hu, Techniques for interpretable machine learning., *arXiv: 1808.00033 [cs, stat]*, May 2019. [arXiv: 1808.00033 \[cs, stat\]](https://arxiv.org/abs/1808.00033).
- [21] BioRender. <https://biorender.com/>.
- [22] M. Tharapornsakulwong, Angela Noun, Project: machine learning & mechanism icons. <https://thenounproject.com/>.
- [23] Z.C. Lipton, The mythos of model interpretability., *arXiv:1606.03490 [cs, stat]*, Mar. 2017. [arXiv: 1606.03490 \[cs, stat\]](https://arxiv.org/abs/1606.03490).

- [24] F. Doshi-Velez, B. Kim, Towards A Rigorous Science of Interpretable Machine Learning, arXiv:1702.08608 [cs, stat], Feb. 2017. arXiv: 1702.08608 [cs, stat].
- [25] W.J. Murdoch, et al., Interpretable machine learning: definitions, methods, and applications, 44 (Oct. Proc. Natl. Acad. Sci. Unit. States Am. 116 (2019) 22071–22080, <https://doi.org/10.1073/pnas.1900654116>, 0027-8424, 1091-6490, arXiv: 1901.04592
- [26] C. Molnar, Chapter 2 interpretability, in: Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2020. Online book, <https://christophm.github.io/interpretable-ml-book/>.
- [27] C.B. Azodi, J. Tang, S.-H. Shiu, Opening the black box: interpretable machine learning for geneticists., Trends Genet. ISSN: 0168-9525 36 (6) (June 2020) 442–455, <https://doi.org/10.1016/j.tig.2020.03.005>.
- [28] Y.R. Xie, et al., Single-cell classification using mass spectrometry through interpretable machine learning, Anal. Chem. ISSN: 0003-2700 (June 2020), <https://doi.org/10.1021/acs.analchem.0c01660>.
- [29] S. Meding, et al., Tumor classification of six common cancer types based on proteomic profiling by MALDI imaging, J. Proteome Res. ISSN: 1535-3907 11 (3) (Mar. 2012) 1996–2003, <https://doi.org/10.1021/pr200784p>.
- [30] Y. Zhang, X. Liu, Machine learning techniques for mass spectrometry imaging data analysis and applications., Bioanalysis. ISSN: 1757-6180 10 (8) (Mar. 2018) 519–522, <https://doi.org/10.4155/bio-2017-0281>.
- [31] M. Galli, et al., A support vector machine classification of thyroid bioptic specimens using MALDI-MSI data, Adv. Bioinform. 2016. ISSN: 1687-8027 (2016), <https://doi.org/10.1155/2016/3791214>.
- [32] J. Behrmann, et al., Deep learning for tumor classification in imaging mass spectrometry, Bioinformatics 34 (7) (Jan. 2018) 1215–1223. ISSN: 1367-4811.
- [33] J. van Kersbergen, et al., Cancer detection in mass spectrometry imaging data by dilated convolutional neural networks, in: Medical Imaging 2019: Digital Pathology, vol. 10956, International Society for Optics and Photonics, Mar. 2019, p. 109560I, <https://doi.org/10.1117/12.2512360>.
- [34] Z. Zhou, R.N. Zare, Personal information from latent fingerprints using desorption electrospray ionization mass spectrometry and machine learning, Anal. Chem. ISSN: 1520-6882 89 (2) (Jan. 2017) 1369–1372, <https://doi.org/10.1021/acs.analchem.6b04498>.
- [35] K. Margulis, et al., Combining desorption electrospray ionization mass spectrometry imaging and machine learning for molecular recognition of myocardial infarction, 20, Anal. Chem. ISSN: 1520-6882 90 (Oct. 2018) 12198–12206, <https://doi.org/10.1021/acs.analchem.8b03410>.
- [36] M. Krzywinski, N. Altman, Classification and regression trees, Nat. Methods. ISSN: 1548-7105 14 (8) (Aug. 2017) 757–758, <https://doi.org/10.1038/nmeth.4370>.
- [37] R.O. Duda, D.G. Stork, P.E. Hart, Chapter 8: tree-based methods, in: Pattern Classification, 2nd., John Wiley & Sons, 2012.
- [38] L. Breiman, et al., Classification and Regression Trees, Taylor & Francis, 1984, 978-0-412-04841-8.
- [39] S. Russell and P. Norvig. "Chapter 18: learning from examples." In: Artificial Intelligence, a Modern Approach. 3rd. Prentice Hall. ISBN: 978-0-13-604259-4.
- [40] T. Chen, C. Guestrin, " XGBoost, A scalable tree boosting system.", in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, 2016, pp. 785–794, <https://doi.org/10.1145/2939672.2939785>, arXiv: 1603.02754.
- [41] J. Friedman, Stochastic gradient boosting, Comput. Stat. Data Anal. 38 (Mar. 1999) 367–378, [https://doi.org/10.1016/S0167-9473\(01\)00065-2](https://doi.org/10.1016/S0167-9473(01)00065-2).
- [42] J.H. Friedman, Greedy function approximation: a gradient boosting machine, In: Ann. Stat. 29 (5) (Oct. 2001) 1189–1232, <https://doi.org/10.1214/aos/1013203451>, 0090-5364, 2168-8966.
- [43] Y. Freund, R.E. Schapire, A short introduction to boosting., in: In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence. Morgan Kaufmann, 1999, pp. 1401–1406.
- [44] K.P. Murphy, Chapter 16: adaptive basis function models (section 16.4: boosting & section 16.3: generalized additive models), in: Machine Learning: A Probabilistic Perspective, The MIT Press, 2012, 978-0-262-01802-9.
- [45] J. Friedman, T. Hastie, R. Tibshirani, Additive logistic regression: a statistical view of boosting, Ann. Stat. 28 (Apr. 2000) 337–407, <https://doi.org/10.1214/aos/1016218223>.
- [46] T. Hastie, R. Tibshirani, J. Friedman, Chapter 10: boosting and additive trees., in: Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd., Springer-Verlag, 2017, 978-0-387-84857-0.
- [47] T. Hastie, R. Tibshirani, J. Friedman, Chapter 9: additive models, trees, and related methods., in: Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd., Springer-Verlag, 2017, 978-0-387-84857-0.
- [48] M. Kuhn, K. Johnson, Chapter 8: regression trees and rule-based models (section 8.6: boosting.), in: Applied Predictive Modeling, 1st, Springer, 2013, ISBN 978-1-4614-6849-3.
- [49] XGBoost parameters — xgboost 1.2.0-SNAPSHOT documentation. <https://xgboost.readthedocs.io/en/latest/parameter.html>.
- [50] Python API reference — xgboost 1.4.0-SNAPSHOT documentation. [https://xgboost.readthedocs.io/en/latest/python/python\\_api.html#module-xgboost.sklearn](https://xgboost.readthedocs.io/en/latest/python/python_api.html#module-xgboost.sklearn).
- [51] S. Lundberg, S.-I. Lee, An unexpected unity among methods for interpreting model predictions., arXiv:1611.07478 [cs], Nov. 2016. arXiv:1611.07478[cs].
- [52] S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions., arXiv:1705.07874 [cs, stat], May 2017. arXiv:1705.07874[cs, stat].
- [53] E. Strumbelj, I. Kononenko, An efficient explanation of individual classifications using game theory., J. Mach. Learn. Res. ISSN: 1532-4435 11 (Mar. 2010) 1–18.
- [54] E. Strumbelj, I. Kononenko, Explaining prediction models and individual predictions with feature contributions., Knowl. Inf. Syst. ISSN: 0219-3116 41 (3) (Dec. 2014) 647–665, <https://doi.org/10.1007/s10115-013-0679-x>.
- [55] M.T. Ribeiro, S. Singh, C. Guestrin, Why should I trust you?: explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, ACM Press, San Francisco, California, USA, 2016, ISBN 978-1-4503-4232-2, pp. 1135–1144, <https://doi.org/10.1145/2939672.2939778>.
- [56] S.M. Lundberg, et al., From local explanations to global understanding with explainable AI for trees, Nat. Machine Intell. ISSN: 2522-5839 2 (1) (Jan. 2020) 56–67, <https://doi.org/10.1038/s42256-019-0138-9>.
- [57] H. Chen, et al., True to the model or true to the data?, arXiv:2006.16234 [cs, stat], June 2020. arXiv:2006.16234[cs, stat].
- [58] S.M. Lundberg, G.G. Erion, S.-I. Lee, Consistent individualized feature attribution for tree ensembles., arXiv:1802.03888 [cs, stat], Mar. 2019. arXiv: 1802.03888[cs, stat].
- [59] 3.2.4.3.1. Sklearn. Ensemble, RandomForestClassifier — scikit-learn 0.23.1 documentation. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>.
- [60] Python API Reference — XGBoost, XGBoostClassifier 1.2.0 documentation. [https://xgboost.readthedocs.io/en/latest/python/python\\_api.html#xgboost.XGBoostClassifier.feature\\_importances\\_](https://xgboost.readthedocs.io/en/latest/python/python_api.html#xgboost.XGBoostClassifier.feature_importances_).
- [61] G. Louppe, Understanding Random Forests: from Theory to Practice, PhD thesis, University of Liège, Department of Electrical Engineering & Computer Science, July 2014. arXiv:1407.7502.
- [62] Shap.TreeExplainer — SHAP latest documentation. [https://shap.lrbjball.readthedocs.io/en/docs\\_update/generated/shap.TreeExplainer.html](https://shap.lrbjball.readthedocs.io/en/docs_update/generated/shap.TreeExplainer.html).
- [63] L. Breiman, Random forests., Mach. Learn. ISSN: 1573-0565 45 (1) (Oct. 2001) 5–32, <https://doi.org/10.1023/A:1010933404324>.
- [64] C. Molnar, Chapter 5.5: permutation feature importance., in: Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2020. <https://christophm.github.io/interpretable-ml-book/>.
- [65] G. Hooker, L. Mentch, Please stop permuting features: an explanation and alternatives., arXiv:1905.03151 [cs, stat], May 2019. arXiv:1905.03151[cs, stat].
- [66] C. Molnar, et al., Pitfalls to avoid when interpreting machine learning models, arXiv:2007.04131 [cs, stat], July 2020. arXiv:2007.04131[cs, stat].
- [67] J.M. Spraggins, et al., High-performance molecular imaging with MALDI trapped ion-mobility time-of-flight(timsTOF) mass spectrometry, Anal. Chem. ISSN: 0003-2700 91 (22) (Nov. 2019) 14552–14560, <https://doi.org/10.1021/acs.analchem.9b03612>.
- [68] N.H. Patterson, et al., Advanced registration and analysis of MALDI imaging mass spectrometry measurements through autofluorescence microscopy, Anal. Chem. ISSN: 0003-2700 90 (21) (Nov. 2018) 12395–12403, <https://doi.org/10.1021/acs.analchem.8b02884>.
- [69] J. Lever, M. Krzywinski, N. Altman, Classification evaluation, Nat. Methods. ISSN: 1548-7105 13 (8) (Aug. 2016) 603–604, <https://doi.org/10.1038/nmeth.3945>.
- [70] D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature. ISSN: 1476-4687 401 (6755) (Oct. 1999) 788–791, <https://doi.org/10.1038/44565>.
- [71] X.-C. Xiong, et al., Feature extraction approach for mass spectrometry imaging data using non-negative matrix factorization, Chin. J. Anal. Chem. ISSN: 1872-2040 40 (5) (May 2012) 663–669, [https://doi.org/10.1016/S1872-2040\(11\)60544-6](https://doi.org/10.1016/S1872-2040(11)60544-6).
- [72] F. Fournelle, et al., Minimizing visceral fat delocalization on tissue sections with porous aluminum oxide slides for imaging mass spectrometry, Anal. Chem. ISSN: 0003-2700 92 (7) (Apr. 2020) 5158–5167, <https://doi.org/10.1021/acs.analchem.9b05665>.
- [73] M.P. Snyder, et al., The human body at cellular resolution: the NIH human biomolecular Atlas Program, Nature. ISSN: 1476-4687 574 (7777) (Oct. 2019) 187–192, <https://doi.org/10.1038/s41586-019-1629-x>.
- [74] E. Ong, et al., Modelling kidney disease using ontology: insights from the kidney precision medicine project, Nat. Rev. Nephrol. ISSN: 1759-507X 16 (11) (Nov. 2020) 686–696, <https://doi.org/10.1038/s41581-020-00335-w>.
- [75] R. Van de Plas, et al., Image fusion of mass spectrometry and microscopy: a multimodality paradigm for molecular tissue mapping, Nat. Methods. ISSN: 1548-7105 12 (4) (Apr. 2015) 366–372, <https://doi.org/10.1038/nmeth.3296>.
- [76] M.A. Jones, et al., Discovering new lipidomic features using cell type specific fluorophore expression to provide spatial and biological specificity in a multimodal workflow with MALDI imaging mass spectrometry, Anal. Chem. ISSN: 0003-2700 92 (10) (May 2020) 7079–7086, <https://doi.org/10.1021/acs.analchem.0c00446>.