analytical chemistry

Active Learning for Convenient Annotation and Classification of Secondary Ion Mass Spectrometry Images

Michael Hanselmann,[†] Jens Röder,[†,‡] Ullrich Köthe,[†] Bernhard Y. Renard,[§] Ron M. A. Heeren,[∥] and Fred A. Hamprecht*[,†]

[†]Heidelberg Collaboratory for Image Processing (HCI), Interdisciplinary Center for Scientific Computing (IWR), University of Heidelberg, Speyerer Straße 6, Heidelberg, Germany

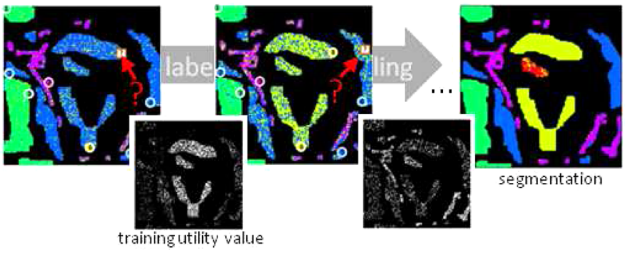
[‡]Robert Bosch GmbH, CR/AEM5, Robert-Bosch-Straße 200, Hildesheim, Germany

[§]Research Group Bioinformatics (NG 4), Robert Koch Institute, Nordufer 20, Berlin, Germany

[∥]FOM-AMOLF, Science Park 104, 1098 XG, Amsterdam, The Netherlands

Supporting Information

ABSTRACT: Digital staining for the automated annotation of mass spectrometry imaging (MSI) data has previously been achieved using state-of-the-art classifiers such as random forests or support vector machines (SVMs). However, the training of such classifiers requires an expert to label exemplary data in advance. This process is time-consuming and hence costly, especially if the tissue is heterogeneous. In theory, it may be sufficient to only label a few highly representative pixels of an MS image, but it is not known a priori which pixels to select. This motivates *active learning* strategies in which the algorithm itself queries the expert by automatically suggesting promising candidate pixels of an MS image for labeling. Given a suitable querying strategy, the number of required training labels can be significantly reduced while maintaining classification accuracy. In this work, we propose active learning for convenient annotation of MSI data. We generalize a recently proposed active learning method to the multiclass case and combine it with the random forest classifier. Its superior performance over random sampling is demonstrated on secondary ion mass spectrometry data, making it an interesting approach for the classification of MS images.



Mass spectrometry imaging (MSI)[1,2] allows a detailed analysis of the spatial distribution of proteins, peptides, lipids, or metabolites.[3,4] With recent efforts to standardize proteomics experiments,[5−8] MSI continuously moves closer to clinical application.[3,9−11] In many of these recent studies, the MS image is spatially partitioned into coherent regions associated with cancer or healthy tissue or into regions corresponding to different cell types. Manual analysis requires the expert to inspect multiple m/z channel images. Moreover, analyzing the channel images independently may not even be sufficient for discriminating tissue types with similar molecular signatures. For these reasons and with data sizes of up to several gigabytes,[12] direct manual analysis becomes tedious or infeasible, emphasizing the need for automated methods.

Previous studies have shown that unsupervised methods such as hierarchical clustering,[13] principal component analysis (PCA),[14] or probabilistic latent semantic analysis (pLSA)[15] are useful for segmenting MS images into spectrally coherent regions based on their molecular signatures only. At the same time, they are intrinsically limited by their inability to learn from expert annotations. One consequence is the lack of clear criteria for model optimization.[16] If the underlying mathematical assumptions are inept for the data at hand, the user has very limited influence on the segmentation outcome.

Many recent studies have thus considered supervised approaches and demonstrated that, given a set of spatially resolved annotations or (immunohistochemical) expert labels, supervised classifiers can be used for automated discrimination of tissue types.[17−21] Even so, technical and biological variability between experiments often remains significant.[22] Depending on the precise application, this limits the classification accuracies that can be achieved, especially in studies where the size of the training set is small. In such scenarios, where training of classifiers that generalize well to new MSI data is difficult, more robust and reliable results might be obtained by training the classifier anew for each separate MSI set. However, labeling of MSI data is time-consuming and consequently very expensive. It is thus desirable to reduce the number of required labels (i.e., labeling time for the expert) without jeopardizing classification accuracy. This motivates the application of semisupervised

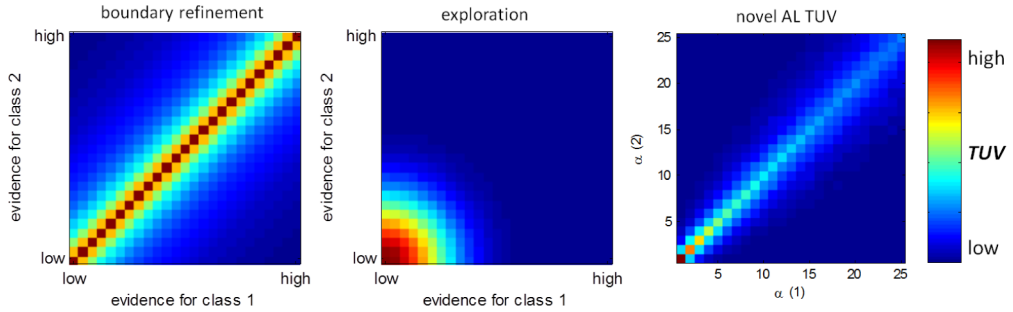


Figure 1. Training utility value (TUV) of a candidate point in a binary classification setting. In (pure) decision boundary refinement, or uncertainty learning, candidate points with equal amounts of evidence for either class are preferred, regardless of how much evidence there is. In (pure) exploration, the candidate points receive a high score if the (local) evidence for both classes is low. Only the absolute "amount" of evidence is considered; its consistency is neglected. On the right, the newly proposed TUV function is shown for different parameter settings, where the evidence for classes 1 and 2 is measured by $\alpha_1 \in \{1, ..., 25\}$ and $\alpha_2 \in \{1, ..., 25\}$. We observe that our TUV function reconciles exploration and decision boundary refinement (also see the Supporting Information, part B).

learning (SSL) techniques,[23−25] active learning (AL) strategies (see Settles[26] for a review), or hybrid approaches.[27]

SSL methods typically base their classification output on two sources of information: the labels given by the user and the underlying structure of the unlabeled data points. An interesting and highly interactive method for matrix-assisted laser desorption ionization (MALDI) MSI analysis was recently published by Bruand et al.[25] While SSL approaches can exploit the information hidden in the unlabeled observations, they lack a concept for guiding the labeling expert. In contrast, in active learning, the algorithm iteratively queries the expert to label that observation for which additional knowledge may be most beneficial for improving the classifier's performance. By labeling the samples (observations) of a data set in a smart order, a high performance level can often be obtained with fewer training samples. Although AL methods have shown excellent performance in many fields such as speech recognition,[28] image classification,[29] remote sensing,[30−32] and biomedical imaging,[33,34] only a few researchers have applied them to MS data.[35−37] None of these publications is on MSI.

In this paper, we generalize a recently proposed AL strategy[38] to the multiclass setting and combine it with the random forest classifier,[39] which has previously been used for efficient classification of MSI data.[21] We show on real world MS images that our approach results in high classification accuracies after only a few learning steps and is thus suitable for efficient annotation of MSI data sets. We further demonstrate that the algorithm has an inbuilt capacity for novelty detection, alerting the expert to previously unlabeled but distinct classes rather than blindly making a prediction. Given the same number of labels, our querying strategy outperforms traditional nonactive learning by up to 10% in sensitivity and 2−4% in positive predictive value. In our experiments, random sampling requires more than twice as many labels to achieve the same performance level. Finally, our strategy does not suffer from the high variability between runs that are characteristic for the random sampling approach.

METHODS

Active Learning. Active learning (AL) aims at achieving steep learning curves, that is, high classification accuracies after seeing as few labeled training examples as possible. It is motivated by the observation that a classifier can benefit more from judiciously chosen and informative training examples than from large numbers of redundant and hence less informative examples.[40] Typically, AL approaches are iterative and "guide" the labeler in the sense that the algorithm chooses observations for which it needs labels.[26] In each round, the algorithm requests a label for that observation (pixel) x of an (MS) image that has the maximum training utility value (TUV) in the set U of all unlabeled observations and is thus expected to contribute most to improving the classifier's performance. After label assignment, the classifier is trained with the augmented label set, all unlabeled observations are reclassified, and the algorithm continues by presenting its next query. These steps are repeated until either the human expert is satisfied with the classification result or a predefined stopping criterion is met.

A meaningful TUV function balances two strategies: *exploration* of the feature space and *refinement* of the current decision boundary. The aim of exploration is to sample from those regions of feature space from which so far only a few training examples are available. The rationale is that a test sample can only be classified well if enough (local) evidence is available. Whereas exploration thus seeks good sample coverage of the whole feature space, the refinement strategy tries to improve the classifier by sampling points that are close to the decision boundary, that is, for which approximately equal probability for two or more classes is present. Figure 1 illustrates these strategies for the binary case.

The proposed active querying strategy can be illustrated with the following thought experiment: consider three different points in a feature space, and do not assume that the true decision boundary comes from a simple parametric class, such as a hyperplane. Points 1 and 2 lie on the currently estimated decision boundary, and point 3 lies far away from it. There are many labels available in the vicinity of point 1, a few in the neighborhood of point 2, and none surrounding point 3. Let $x^{(i)}$ with $i \in \{1, 2, 3\}$ denote the three points.

A pure refinement strategy would favor points $\{1, 2\}$ over 3. An exploratory strategy would take more interest in 3 than in $\{1, 2\}$. We use a strategy that prefers $\{2, 3\}$ over 1, for the following reasons: Point 3 is interesting, because we know nothing about its true class (remember that we do not assume a simple parametric model for the decision boundary). Point 2 is interesting because the location of the decision boundary is based on an estimate of $\hat{p}(Y|x^{(2)})$, which being a random variable of itself is of necessity imprecise when based on only few labeled points. There is thus some potential to be informed, or surprised, by an additional label at point 2. Point 1 is uninteresting because its estimate $\hat{p}(Y|x^{(1)})$ is based on a large

number of nearby training examples and we do not expect the decision boundary to change substantially in response to yet another label at that point. Finally, we factor the marginal density $\hat{p}(x)$ of all labeled and unlabeled points into the proposed training utility value. The reason is that estimating the decision boundary well is only relevant in populated regions of feature space. The vehicle used to capture the above intuition is a *second-order distribution*, that is, the distribution of the probabilistic point estimate $\hat{p}(Y|x)$. This distribution and its use in a training utility value are defined next.

Training Utility Value Function. Above, we have informally discussed favorable properties of the TUV function. This section approaches the problem from a more theoretical perspective and may be skipped by the less mathematically interested reader.

Let $(\mathcal{X}, \mathcal{Y}) = \{(x^{(1)}, y^{(1)}), ..., (x^{(N)}, y^{(N)})\}$ be the set of N training samples, that is, mass spectra $x^{(k)}$ with M channels and corresponding class labels $y^{(k)} \in \{1, ..., d\}$. Let further L be a loss function, that is, a function that quantifies the penalty associated with an incorrect classification. The lowest achievable classification error, for a given loss function L, data distribution $p(x, y)$, and classification rule θ, is given by the overall *expected risk* $\int_{\mathcal{X}} R(\pi(x))p(x)\mathrm{d}x$. The conditional risk for misclassifying a point at position x is given by

$$R(\pi(x)) := \mathbb{E}_{Y|x}(L(Y = y, \theta(\pi(x)))) \quad (1)$$

$$\pi(x) := [p(Y = 1|x), ..., p(Y = d|x)]^T \quad (2)$$

Here, $L(y, z)$ is the loss associated with a prediction z if the true class label is y; $\pi(x) \in \mathbb{S}^d$ is the vector of class conditional probabilities for each of the d classes which, thanks to the normalization constraint $\sum_{y\in\mathcal{Y}} p(y|x) = 1$, lies in the unit simplex $\mathbb{S}^d$ with d vertices (see Figure 2 for an example with $d = 3$ vertices). Finally, θ is a classification rule $\mathbb{S}^d \rightarrow \{1, ..., d\}$ that maps any point in the simplex to one of the d classes. The canonical mapping function θ employs the winner-takes-all strategy, i.e. maps each point from the simplex to its closest vertex (and hence to the class associated with that vertex).

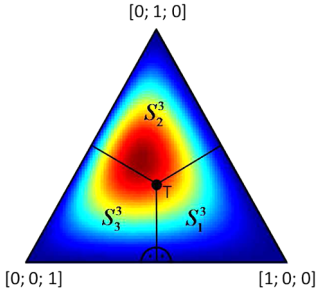


Figure 2. Each vertex of the simplex $\mathbb{S}^3$ corresponds to one of the $d = 3$ classes of interest. The mapping function θ (cf. eq 3) maps each point on the simplex to one of these classes. In the canonical case, each point is assigned to the closest vertex and hence to the class associated with that vertex. Figuratively, threshold point T (which lies in the center of the simplex) is used to partition $\mathbb{S}^3$ into three parts $\mathbb{S}_j^3$, $j = 1, ..., 3$. $\mathbb{S}_j^3$ is the Voronoi region associated with the j-th vertex. The posterior estimate for a test point can now be interpreted as a point on this simplex. In the TUV for Random Forests section, we further describe how a Dirichlet distribution can be employed to describe the second-order distribution of the posterior. Color-coding is used to show an example for such a second-order distribution, where blue indicates low and red indicates high probability. A uniformly colored simplex would correspond to an uninformative prediction. In contrast, in the example the plotted Dirichlet distribution is concentrated in part $\mathbb{S}_2^3$ of the simplex, indicating a preference for class two.

In practice, the true class conditional probabilities $\pi(x)$ are not known but need to be estimated from training data.[41] Classifiers such as logistic regression or polychotomous logistic regression offer point estimates $q_y^0(x) := \hat{p}(Y = y|x)$, $y \in \mathcal{Y}$ which can be compiled in a d-dimensional vector $q^0(x) = [\hat{p}(Y = 1|x), ..., \hat{p}(Y = d|x)] \in \mathbb{S}^d$. Plugging this point estimate into the conditional risk gives

$$R(q^0(x)) := \sum_{y\in\mathcal{Y}} L(y, \theta(q^0(x)))\cdot q_y^0(x) \quad (3)$$

This quantity is the key ingredient of uncertainty sampling, which has been presented in many variants.[42−44] This class of active learning algorithms seeks to reduce the estimated expected risk by querying additional labels near the decision boundary, where the conditional risk is greatest. The implicit hope is that additional labels may drive the updated class conditional probability toward one of the simplex vertices, that is, to obtain unequivocal evidence for the dominance of one class. Uncertainty sampling is very simple to implement and widely used, but it is a pure exploitation/refinement strategy: it will never explore uncharted regions of feature space. Indeed, it will spend all of its queries around the current decision boundary. In addition, uncertainty sampling only relies on a point estimate of the posterior distribution and does not consider the uncertainties of the class conditional probability estimates themselves. This "second-order" uncertainty is implicitly taken into account in schemes such as error reduction sampling.[45,46] However, such look-ahead schemes require a (rank-one) update of the current classification boundary and turn out to be relatively expensive.

The novelty in ref 38 is that it makes explicit, and capitalizes on, the uncertainty of the class-conditional probability itself. The latter, like any estimate that is obtained from finite training data, is subject to uncertainty. The prerequisite for their procedure is that the classifier must provide not merely a point estimate $q^0(x)$ for the class conditional probability but a full *second-order distribution* over $q(x)$ as expressed by a probability density function $g(q(x))$. More specifically, an estimated second-order distribution over the class-conditional probability can be written as

$$g(q(x)) := \frac{\partial G(q(x))}{\partial q(x)} \quad (4)$$

$$G(q(x)) := Pr(\hat{p}(Y = 1|x) \leq q_1(x) \wedge ... \wedge \hat{p}(Y = d|x) \leq q_d(x)) \quad (5)$$

with density g and cumulative distribution function G.

If such a second-order distribution is available, the point estimate $q^0(x)$ can be identified with $q^0(x) \equiv \mathbb{E}_q(q(x))$ and $R(q^0(x))$ from eq 3 can be rewritten as $R(\mathbb{E}_q(q(x)))$.

Now, in ref 38, we argue that this estimate is overly conservative and tends to overrate the utility of samples whose intrinsic (i.e., Bayesian) uncertainty is high. We contrast it with the following distributional estimate, which measures the risk at location x arising from intrinsic uncertainty and insufficient training combined:

$$\mathbb{E}_q(R(q(x))) := \sum_{y\in\mathcal{Y}} \int L(y, \theta(q(x)))\cdot q_y(x)\cdot g(q(x))\mathrm{d}q(x) \tag{6}$$

We further argue that the extent by which these estimates differ, when weighted with the estimated marginal density $\hat{p}(x)$ (to take into account the importance of location x), is a good TUV, or measure of interestingness, for yet unlabeled observations. Specifically, we posit

$$\mathrm{TUV}(x) = \hat{p}(x)(R(\mathbb{E}_q(q(x))) - \mathbb{E}_q(R(q(x)))) \tag{7}$$

and show superior active learning curves when averaging over a large number of data sets.

This TUV function can be seen to naturally balance both exploration and refinement, see Figure 1. In particular, unlike uncertainty sampling strategies, this criterion eventually desists from querying further labels near the decision boundary in areas where multiple labels are already available: these areas exhibit high intrinsic uncertainty that cannot be removed by additional label queries. Also, the proposed criterion does not have additional parameters as required by heuristic strategies that alternate between exploration and exploitation phases.[47]

To summarize this discussion, in areas with few labels and in the absence of a parametric model that is known to govern the true posterior probability of a class, the estimate of the class conditional probability is of necessity imprecise. This uncertainty is reflected in a broad second-order distribution, which leads to lower values of $\mathbb{E}_q(R(q(x)))$ as compared to the more conservative $R(\mathbb{E}_q(q(x)))$. If, on the other hand, the local evidence is high, the second-order distribution is narrow, yielding similar values for both terms. An example is given in Supporting Information part A.

Random Forests. The random forest[39] (cf. Figure 3) is a state-of-the-art ensemble classifier that comprises n_{tree} decision trees. Each individual tree constitutes a crisp classifier and is constructed from a bootstrap sample of size N of all available training samples. Tree construction starts at the root node and proceeds down toward the leaf nodes. In each node, a subset of the M features (i.e., mass channels) is chosen at random (a typical subset size being $\sqrt{M}$), and the feature that allows for the best class separation of the samples in the node is selected. After splitting the node, the algorithm continues on the next level until all nodes are pure, that is, contain samples with consistent class labels. All samples that are not part of the bootstrap sample, the so-called out-of-bag samples, can be used to obtain a performance estimate for the classifier. A query sample is classified by putting it down each of the trees in the ensemble until it reaches the leaf nodes. The distribution over classes obtained for a single query sample cannot strictly be interpreted as a posterior probability but does give an indication of how certain the classifier is in its prediction.

Many studies have shown that the random forest classifier is robust to overfitting and label noise,[39,48] delivers state-of-the-art prediction accuracy,[49,50] can handle a large number of input variables,[51,52] allows for fast training, and is robust with respect to the exact choice of the two hyperparameters: number of trees and size of the random feature subset evaluated at a node.[53]

TUV for Random Forests. We now combine the TUV with the random forest classifier in a multiclass setting. As discussed above, given a test sample, the random forest classifier provides a distribution over tree votes. To obtain both

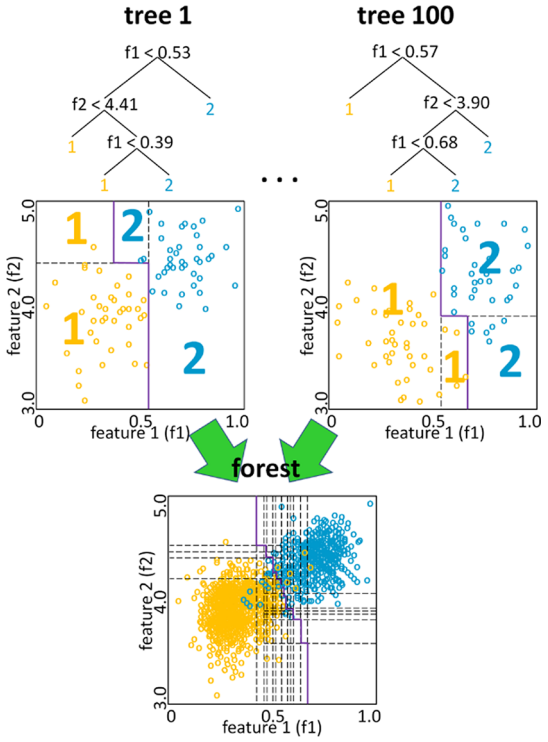


Figure 3. Random forest classifier is an ensemble of decision trees where the single trees are constructed from bootstrap samples. At each node of a tree, the feature that allows for the best class separation is chosen (with respect to the subset of features selected for that node). The corresponding partitioning of the feature space is shown with the decision boundary plotted in purple. The collection of trees forms the random forest whose classification is based on the majority votes of the individual trees.

a density estimate and a meaningful measure for the uncertainty (pure leaves suggest perfect certainty and are hence misleading), we train the random forest with all labeled examples from previous learning rounds plus a predefined fraction of samples from a uniformly distributed auxiliary class "0". After training, all hitherto unlabeled MSI samples are classified. Among these points, the next query candidate is selected.

The number of trees $v_i(x)$ voting for the $d + 1$ classes ($i = 0, 1, ..., d$) can now be interpreted as an indicator for how certain the classifier's assessment for x is. Simply put, the more trees vote for the auxiliary class, the weaker the local evidence for the other classes and thus the higher the uncertainty of the classifier. At the same time, the relative number of votes for the remaining classes is an indicator of how far x lies from the decision boundary. Generalizing the Beta distribution from ref 38 to multiple classes, we model the probability density function $g(q)$ (cf. eq 5) with a Dirichlet distribution, which is parametrized by the number of trees voting for classes 1 to d. This yields $g(q) = Dir(q|\alpha)$ where $\alpha \in \mathbb{N}_+^d$, $\alpha_y = 1 + v_y(x)$ and $\sum_{i=1}^{d}\alpha_i = d + n_{\mathrm{tree}}$ (see Figure 2). The complete mathematical derivation is detailed in Supporting Information part B.

Figure 1 and Supporting Information part C show that this choice yields a TUV function that obeys both exploration and refinement principles. Computation of the TUV requires Monte Carlo integration over parts of the simplex. An efficient implementation is discussed in the Supporting Information, parts D and E; MATLAB code is available from http://hci.iwr.

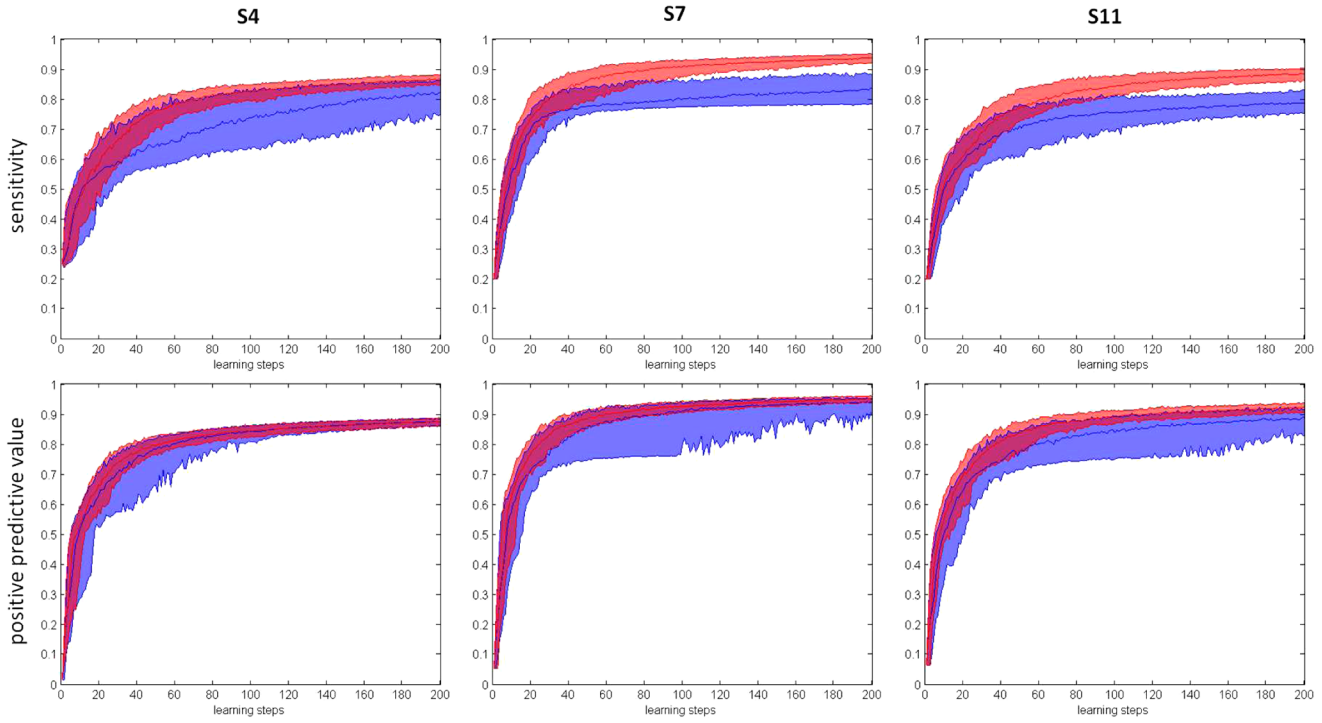


Figure 4. Learning curves obtained for random sampling (RS, blue) and our active learning approach (AL, red). Accuracies are measured by sensitivity (top row) and positive predictive value (bottom row). In each learning step, one additional label is queried. The plots show the median as well as the band between the 95% quantile and the 5% quantile for the 100 repeats. In contrast to RS, our AL approach exhibits significantly lower variance between the different learning runs, and the band around the median gets thinner over the course of iterations. At the same time, it significantly outperforms RS.

uni-heidelberg.de/MIP/Software. An overview of the active learning method is given in algorithm 1:

Query label for observation x with the largest density in feature space
for $k = 1$ to maxIterations **do**
1. Uniformly sample from the bounding box enclosing all observations in feature space and label the obtained auxiliary samples as "0" (frequency controlled by resampling parameter)
2. Combine user-labeled samples and "0"-samples to train a random forest classifier with $d + 1$ classes
3. Classify all unlabeled observations $x \in U$, i.e. all observations to which the user has not yet assigned a label
4. Drop random forest votes for class "0" to obtain d-dimensional vectors α for all unlabeled observations $x \in U$ with $\alpha_i = 1 + v_i(x)$, $i = 1, \ldots, d$ where d is the number of classes and $v_i(x)$ is the number of trees that vote for class i given observation x
5. Query user label for that observation x that has the *highest training utility value* (*TUV*) among all yet unlabeled observations (i.e. $\max_{x \in U}$ TUV (x), cf. eq (7) and Supporting Information D)
end for

EXPERIMENTS

Data. We used secondary ion mass spectrometry (SIMS) data acquired from orthotopic human breast cancer xenografts (MCF-7) grown in mice. For data acquisition, a Physical Electronics TRIFT II TOF SIMS equipped with an Au+ liquid metal ion cluster gun was used. The tumor samples were embedded in gelatin, flash-frozen, cryo-sectioned to $\approx$10 μm, and thaw-mounted on a cold indium tin oxide coated glass slide. The tissues were not washed prior to SIMS analysis, which was confined to a mass range of 0−2000 Da. The spectral resolution was rebinned to 0.1 Da, and the range between 0 and 400 Da was selected, resulting in 4009 mass channels. Due to the large amount of data processed in this study, short acquisition times of 2 s per spot were used. Consequently, the spatial resolution had to be rebinned to 35 × 35 μm^2 per pixel in order to guarantee a reasonable number of ion counts in each mass spectrum.

Three out of the six slices used in a previous study[15,21] were selected for evaluation of our active learning method: one from the bottom (entitled S4), middle (S7), and top (S11) of the stack of available parallel slices of the tumor. The spectra in the three data sets were baseline corrected by channelwise subtraction of the minimum and normalized by their total ion count, and features were extracted with a peak picker based on local maximum detection. The dimensionality of the resulting spectra varied from 64 to 69 for the three sets. Crisp gold standard labels were obtained by Hematoxylin−Eosin (HE) staining of parallel slices, and five classes of interest were identified: necrotic tumor, viable tumor, tumor interface, gelatin, glass/hole (see ref 21 and the Supporting Information part F for a more detailed description). All observations (pixels) for which label information is available were used in the evaluation of the methods. The class distribution among the labels corresponding to these observations determines the (maximum) number of different regions/classes in the segmentation result. Since section S4 only contains labels for four of the five classes, S4 was segmented into four regions. In contrast, S7 and S11 were segmented into five regions.

Evaluation Criteria. We compared our active learning approach (AL-RF) to random sampling (RS) that, in each learning step, randomly queries the label of a hitherto unlabeled

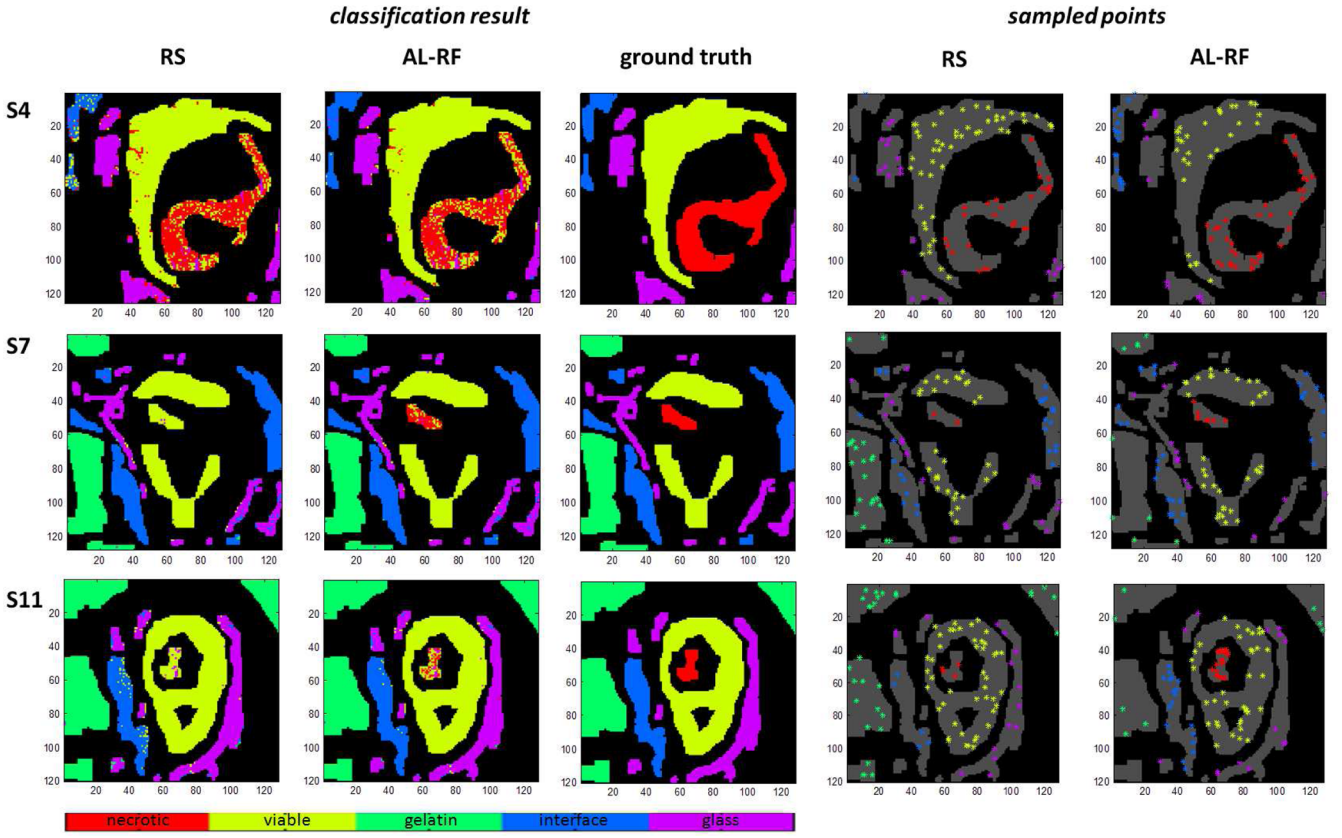


Figure 5. Classification results after 100 learning steps with our active learning method (AL-RF) and random sampling (RS). To obtain the crisp classification, we first averaged the probability maps gathered in the 100 repeats and then took the maximum likelihood estimate in each pixel. On the right, the selected training points for a representative learning run are plotted (we refrain from plotting the training points for all 100 repeats to keep the images uncluttered). Since the area of the necrotic class is comparatively small in slices S7 and S11, RS only selects very few training points for that class, leading to a bad classification result. In contrast, AL-RF requests more training samples for that class, yielding a superior classification. At the same time, it samples less points from the gelatin and glass classes, which have less overlap with the other classes in feature space than, e.g., necrotic and viable tissue, and are thus easier to learn.

observation. Random sampling was used for comparison as it is known to be "surprisingly effective, being competitive with more complex approaches"[54] and performs reasonably well in many studies.[55,56] It has thus been established as the de facto baseline strategy to compare new AL algorithms to. Prediction accuracy was measured by sensitivity (SE) and positive predictive value (PPV). Sensitivity is defined as SE = (TP)/(TP + FN) where TP is the number of true positives and FN is the number of false negatives. The positive predictive value estimates the ratio of samples that are correctly classified as class k among all samples that are classified as k, that is, PPV = (TP)/(TP + FP) where FP is the number of false positives. We averaged the obtained SE and PPV rates over all four (slice S4) or five classes (slices S7 and S11).

Due to the nondeterministic nature of the RS strategy and the Monte Carlo integration, we repeated the AL method and the RS approach 100 times and averaged the obtained results in each learning step. To obtain reliable quality estimates, in addition, we repeated the random forest training and classification in each learning step five times. We drew 300 samples to perform the Monte Carlo integrations and employed stratified sampling to balance the labels in the training set. In both approaches, the learning was started with an empty set of labeled points (in practical applications a number of initial labels might already be given, such as it is, e.g., possible in AMASS[25]), exactly one label was queried in each active learning step where the ground truth label map served as oracle, and a 0−1 loss function was assumed.

RESULTS

Figure 4 and Supporting Information parts G and H report the obtained classification accuracies on the three MSI data sets. Results are given for both querying strategies and an increasing number of learning steps. Ideally, the learning curves are steep, such that high classification accuracies are obtained after only a few learning steps. Since this is typically achieved by first querying the labels that have the highest potential of increasing the classifier's performance, it is also insightful to examine which training points the methods select within a fixed number of learning steps (here, 100). Intuitively, some of the classes are easier to distinguish than others, which is likely to manifest itself in the training point selection of the AL strategy. Results are shown in Figure 5, and a step-by-step example for slice S7 is given in Supporting Information part I. In detail, the following results were obtained for slices S4, S7, and S11.

Slice S4. Figure 4 and Supporting Information part G reveal that our active learning scheme (AL-RF) performed similarly to RS in the first few learning steps and significantly outperformed RS as soon as more than ≈20 learning steps were executed. Due to the steeper learning curve, AL-RF improved on RS by about 10% in sensitivity after 100 iterations. RS needed more than 200 learning steps (i.e., twice as many labels) to achieve

the same performance level. For a large number of learning steps, RS eventually collected a sufficient number of samples from all classes and hence converged toward the sensitivity rates obtained with AL-RF. However, the margin was still more than 5% after 200 iterations (cf. Supporting Information part H). Regarding positive predictive value, AL-RF slightly outperformed RS in the first ≈70 learning steps, that is, in the regime which is most interesting for a learning from sparse annotations.

Slice S7. Over the whole range of the first 200 iterations and especially for low numbers of learning steps, our approach outperformed RS with respect to PPV. At the same time, it significantly outperformed RS regarding sensitivity, leading to a gain of more than 10% after 100 and also after 200 learning steps. Again, RS required more than twice as many labels to reach the performance level of AL-RF after 100 steps. The sensitivity of the RS algorithm increased very slowly such that after 500 iterations the sensitivity was still at a comparably low level of 86%.

Figure 5 reveals that RS resulted in a classifier that mostly confused the necrotic class (indicated in red) with the viable class (light green). In contrast, AL-RF yielded significantly better results. Gelatin and glass spectra did not pose a challenge for either strategy.

Slice S11. Regarding sensitivity as well as positive predictive value, the results obtained for slice S11 proved to be highly similar to the results for slice S7. AL-RF again outperformed RS with respect to both sensitivity and positive predictive value. After 100 and 200 learning steps, it resulted in SE and PPV rates that were approximately 9% respectively 4−6% higher than the results yielded with RS. Figure 5 shows that RS again failed to achieve good classification performance for the necrotic class. AL-RF performed significantly better but still confused several necrotic samples with viable cancer and some with glass. Apparently, additional learning steps are necessary to learn to reliably discriminate necrotic and viable tumor in this data set.

DISCUSSION

Classification Performance. Given a fixed number of learning steps, AL-RF resulted in positive predictive values that were slightly higher or comparable to the ones obtained with RS. At the same time, AL-RF significantly outperformed RS with respect to sensitivity by up to 10%, as soon as more than 15−20 labels were queried. It also exhibited significantly lower variance between runs, as can be seen from Figure 4. The main conclusion is that AL-RF has the potential to reduce labeling times without trading for classification accuracy.

Training Point Selection. Figure 5 shows that RS largely failed to discriminate necrotic from viable tumor tissue. The necrotic area has only small spatial extent, such that RS only selected a few corresponding training points. In comparison, AL-RF selected more than twice as many necrotic samples on slices S7 and S11. This choice seems reasonable, since discriminating viable and necrotic tumor is the most challenging task of our classification problem. In any case, AL-RF yielded a significantly better classification result with respect to these classes (cf. Figure 5). Whereas the necrotic and viable tumor samples are rather close in feature space, the nontissue classes gelatin and glass have little spectral overlap, which simplifies their classification. Indeed, AL-RF queried far fewer samples from these classes than RS, and the corresponding areas in Figure 5 are less densely sampled. We conclude that AL-RF seems to construct training sets that are consistent with our expectations and prior knowledge about the classification task at hand.

Influence of the Number of Trees. There is some freedom in the exact choice of the second-order distribution. The Dirichlet, as a member of the exponential family with the correct support, is a canonical choice. While it allows the combination with the successful random forest classifier, using the tree votes as parameters introduces a certain shortcoming: when increasing the overall number of trees in the ensemble, the parameters specifying the Dirichlet distributions grow larger, which results in a narrower distribution. Thus, ultimately the uncertainty estimate is dependent on the number of trees. However, the number of trees in a random forest is fixed, typically between 100 and 200. Our experiments demonstrate that, for this choice, our criterion works well in practice.

Method's Assumptions. Supervised learning can only be as good as the labels provided, and it is thus important for the expert to ensure that the assigned labels are correct. This requires a certain level of interaction between the AL approach and the microscopy software.

Unsupervised Segmentation Can Assist the Labeling Process. Alternatively, PCA or pLSA scores may be used as overlays when assigning labels. These low-dimensional summaries of the MSI data often reveal structures that are not apparent from individual channel images but are often visible in the stained images (see Supporting Information part J for details).

Computation Time. Training of the random forest and subsequent classification took less than 1 s on a standard desktop PC (2 GHz dual core processor with 2 GBytes of RAM). Computing the risk estimates for all unlabeled observations (cf. Supporting Information part C) required another 1.5−2 s. Performance improvements may be achieved by employing an online version of the random forest classifier[48,57] or by querying multiple labels in each iteration,[58] but this is beyond the scope of this paper. While a speed-up is always desirable, the measured computation times are clearly below the time that an expert typically needs for labeling the query point.

Future Work. Since AL-RF is based on the random forest classifier, which was repeatedly shown to work well on complex MALDI signatures (see e.g., ref 59), and since the results for discriminating similar tissue classes such as viable and necrotic tissue are encouraging, we expect that AL-RF may also become an interesting tool for MALDI MSI analysis. Confirming or refuting this belief is an interesting avenue of future research. Also, the analyzed xenograft tumors are rather homogeneous in nature. Thus, it will be interesting to analyze tissue types that are characterized by spectrally more overlapping signatures. Due to the reasons given above, we believe that AL-RF is suitable for this task.

CONCLUSIONS

Due to the enormous amount of data produced by modern-day instruments, routine clinical application of MSI will not be possible without computational analysis.[60] Robust training of supervised classifiers requires a set of expert labels that reflects the variability between patients and instrument settings. The high variability encountered in practice jeopardizes reproducibility and motivates the collection of expert labels for each newly acquired MSI data set. However, labeling is time-consuming and thus expensive. Consequently, novel algorithms

are needed that yield the highest possible classification accuracies and at the same time require as little user interaction as possible. We have demonstrated how AL can be used for the efficient annotation and classification of SIMS data. We have further demonstrated that it outperforms RS by a large margin if only a small number of labels are made available for training. Harvesting this potential is worthwhile as MSI is moving closer to clinical application.

■ ASSOCIATED CONTENT

Supporting Information

Additional information as noted in the text. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: fred.hamprecht@iwr.uni-heidelberg.de; fax: +49 6221 54 5276.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Kristine Glunde (Johns Hopkins University School of Medicine, Baltimore, MD, USA) and Erika R. Amstalden (FOM-AMOLF, Amsterdam, The Netherlands) for providing the tissue sections and MSI data, as well as Boaz Nadler (Weizman Institute of Science, Rehovot, Israel), Anna Kreshuk (HCI, University of Heidelberg, Heidelberg, Germany), Xinghua Lou (Memorial Sloan-Kettering Cancer Center, New York, NY, USA), and Marc Kirchner (Children's Hospital Boston, MA, USA) for fruitful discussions. We furthermore gratefully acknowledge financial support by the DFG under grant no. HA4364/6-1 (M.H., B.Y.R., F.A.H.) and the Robert Bosch GmbH (J.R., F.A.H.). R.M.A.H. gratefully acknowledges financial support from the programme P24 of the Dutch national program COMMIT. Finally, we thank our reviewers for helpful comments and suggestions.

■ REFERENCES

(1) Caprioli, R.; Farmer, T.; Gile, J. *Anal. Chem.* **1997**, *69*, 4751−4760.
(2) McDonnell, L.; Heeren, R. *Mass Spectrom. Rev.* **2007**, *26*, 606−643.
(3) Seeley, E.; Caprioli, R. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 18126−18131.
(4) Chaurand, P.; Schwartz, S.; Caprioli, R. *Curr. Opin. Chem. Biol.* **2002**, *6*, 676−681.
(5) Taylor, C. F.; Paton, N. W.; Lilley, K. S.; Binz, P.-A.; Julian, R. K., Jr.; Jones, A. R.; Zhu, W.; Apweiler, R.; Aebersold, R.; Deutsch, E. W.; et al. *Nat. Biotechnol.* **2007**, *25*, 887−893.
(6) Slany, A.; Haudek, V.; Gundacker, N.; Griss, J.; Mohr, T.; Wimmer, H.; Eisenbauer, M.; Elbling, L.; Gerner, C. *Electrophoresis* **2009**, *30*, 1306−28.
(7) Franck, J.; Arafah, K.; Elayed, M.; Bonnel, D.; Vergara, D.; Jacquet, A.; Vinatier, D.; Wisztorski, M.; Day, R.; Fournier, I.; Salzet, M. *Mol. Cell. Proteomics* **2009**, *8*, 2023−2033.
(8) Green, F.; Gilmore, I.; Lee, J.; Spencer, S.; Seah, M. *Surf. Interface Anal.* **2010**, *42*, 129−138.
(9) Fournier, I.; Wisztorski, M.; Salzet, M. *Exp. Rev. Proteomics* **2008**, *5*, 413−424.
(10) Seeley, E.; Caprioli, R. *Proteomics: Clin. Appl.* **2008**, *2*, 1435−1443.
(11) Walch, A.; Rauser, S.; Deininger, S.-O.; Höfler, H. *Histochem. Cell Biol.* **2008**, *130*, 421−434.
(12) Eijkel, G.; Kükrer-Kaletas, B.; van der Wiel, I.; Kros, J.; Luider, T.; Heeren, R. *Surf. Interface Anal.* **2009**, *41*, 675−685.
(13) Deininger, S.-O.; Ebert, M.; Fütterer, A.; Gerhard, M.; Röcken, C. *J. Proteome Res.* **2008**, *7*, 5230−5236.
(14) van de Plas, R.; Ojeda, F.; Dewil, M.; van den Bosch, L.; de Moor, B.; Waelkens, E. *Proc. Pac. Symp. Biocomput.* **2007**, *12*, 458−469.
(15) Hanselmann, M.; Kirchner, M.; Renard, B.; Amstalden, E.; Glunde, K.; Heeren, R.; Hamprecht, F. *Anal. Chem.* **2008**, *80*, 9649−9658.
(16) Cord, M.; Cunningham, P. *Machine Learning Techniques for Multimedia*, 1st ed.; Springer: Berlin, Germany, 2008.
(17) Yanagisawa, K.; Shyr, Y.; Xu, B.; Massion, P.; Larsen, P.; White, B.; Roberts, J.; Edgerton, M.; Gonzalez, A.; Nadaf, S.; Moore, J.; Caprioli, R.; Carbone, D. *Lancet* **2003**, *362*, 433−439.
(18) Schwartz, S.; Weil, R.; Thompson, R.; Shyr, Y.; Moore, J.; Toms, S.; Johnson, M.; Caprioli, R. *Cancer Res.* **2005**, *65*, 7674.
(19) Schwamborn, K.; Krieg, R.; Reska, M.; Jakse, G.; Knuechel, R.; Wellmann, A. *Int. J. Mol. Med.* **2007**, *20*, 155−159.
(20) Gerhard, M.; Deininger, S.-O.; Schleif, F.-M. *Symp. Comput.-Based Med. Syst.* **2007**, *20*−*22*, 403−405.
(21) Hanselmann, M.; Köthe, U.; Kirchner, M.; Renard, B.; Amstalden, E.; Glunde, K.; Heeren, R.; Hamprecht, F. *J. Proteome Res.* **2009**, *8*, 3558−3567.
(22) Meyer, H.; Stühler, K. *Proteomics* **2007**, *7* (Suppl 1), 18−26.
(23) Zhu, X. *Semi-Supervised Learning Literature Survey*; Computer Sciences Technical Report 1530; University of Wisconsin: Madison, WI, 2005.
(24) Chapelle, O.; Zien, A.; Schölkopf, B. *Semi-Supervised Learning*; The MIT Press: Cambridge, MA, 2006.
(25) Bruand, J.; Alexandrov, T.; Sistla, S.; Wisztorski, M.; Meriaux, C.; Becker, M.; Salzet, M.; Fournier, I.; Macagno, E.; Bafna, V. *J. Proteome Res.* **2011**, *10*, 4734−4743.
(26) Settles, B. *Active Learning Literature Survey*; Computer Sciences Technical Report 1648; University of Wisconsin: Madison, WI, 2009.
(27) Rajan, S.; Ghosh, J.; Crawford, M. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 1231−1242.
(28) Riccardi, G.; Hakkani-Tür, D. *IEEE Trans. Speech Audio Process.* **2006**, *13*, 1−8.
(29) Joshi, A.; Porikli, F.; Papanikolopoulos, N. *Proc. IEEE Conf. Comput. Vision Pattern Recognit.* **2009**, 2372−2379.
(30) Li, J.; Bioucas-Dias, J.; Plaza, A. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 4085−4098.
(31) Mitra, P.; Shankar, B.; Pal, S. *Pattern Recognit. Lett.* **2004**, *25*, 1067−1074.
(32) Tuia, D.; Ratle, F.; Pacifici, F.; Kanevski, M.; Emery, W. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 2218−2232.
(33) Doyle, S.; Madabhush, A. Consensus of ambiguity: theory and application of active learning for biomedical image analysis. *5th IAPR International Conference on Pattern Recognition in Bioinformatics*, Radboud University Nijmegen, Nijmegen, The Netherlands, September 22−24; Dijkstra, T., Tsivtsivadze, E., Marchiori, E., Heskes, T., Eds.; Springer: Berlin, Germany, 2010; pp 313−324.
(34) Oh, S.; Lee, M. S.; Zhang, B.-T. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2011**, *2*, 316−325.
(35) Zomer, S.; del Nogal Sánchez, M.; Brereton, R.; Pérez Pavón, J. *J. Chemom.* **2004**, *18*, 294−305.
(36) Iyuke, F. M.Sc. Thesis, Ottawa-Carleton Institute for Biomedical Engineering, Ottawa, Canada, 2011.
(37) Shi, J.; Lin, W.; Wu, F.-X. Statistical analysis of mascot peptide identification with active logistic regression. *Proceedings of the 4th International Conference on Bioinformatics and Biomedical Engineering*, Chengdu, China, June 18−20; 2010; p 1−4.
(38) Röder, J.; Kunzmann, K.; Nadler, B.; Hamprecht, F. Active learning with distributional estimates. *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, Catalina Island, USA, August 15−17; Murphy, K., de Freitas, N., Eds.; AUAI Press: Corvallis, OR, 2012; p 715.
(39) Breiman, L. *Mach. Learn.* **2001**, *45*, 5−32.

(40) Schohn, G.; Cohn, D. Less is more: active learning with support vector machines. *Proceedings of the 17th International Conference on Machine Learning*, Stanford University, Stanford, CA, USA, June 29 to July 2; Langley, P., Ed.; Morgan Kaufmann: San Francisco, CA, 2000; pp 839−846.
(41) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: Berlin, Germany, 2009.
(42) Baum, E. *IEEE Trans. Neural Networks* **1991**, *2*, 5−19.
(43) Tong, S.; Koller, D. Support vector machine active learning with applications to text classification. *Proceedings of the 17th International Conference on Machine Learning*, Stanford University, Stanford, CA, USA, June 29 to July 2; Langley, P., Ed.; Morgan Kaufmann: San Francisco, CA, 2000; pp 999−1006.
(44) Scheffer, T.; Decomain, C.; Wrobel, S. Active hidden Markov models for information extraction. *Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis*, Cascais, Portugal, September 13−15; Hoffmann, F., Hand, D. J., Adams, N. M., Fisher, D. H., Guimarães, G., Eds.; Springer: Berlin, Germany, 2001; pp 309−318.
(45) Roy, N.; McCallum, A. Toward optimal active learning through sampling estimation of error reduction. *Proceedings of the 18th International Conference on Machine Learning*, Williams College, Williamstown, MA, USA, June 28 to July 1; Brodley, C. E., Pohoreckyj, A., Eds.; Morgan Kaufmann: San Francisco, CA, 2001; pp 441−448.
(46) Zhu, X.; Lafferty, J.; Ghahramani, Z. In Workshop on the Continuum from Labeled to Unlabeled Data. *Proceedings of the 20th International Conference on Machine Learning*, Washington DC, USA, August 21−24; Fawcett, T., Mishra, N., Eds.; Morgan Kaufmann: San Francisco, CA, 2003; pp 58−65.
(47) Brinker, K. Incorporating diversity in active learning with support vector machines. *Proceedings of the 20th International Conference on Machine Learning*, Washington DC, USA, August 21−24; Fawcett, T., Mishra, N., Eds.; Morgan Kaufmann: San Francisco, CA, 2003; pp 59−66.
(48) Saffari, A.; Leistner, C.; Santner, J.; Godec, M.; Bischof, H. On-line random forests. *3rd IEEE ICCV Workshop on On-line Computer Vision*, Kyoto, Japan, September 27 to October 4; IEEE: New York, 2009; pp 1393−1400.
(49) Caruana, R.; Karampatziakis, N.; Yessenalina, A. An empirical evaluation of supervised learning in high dimensions. *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, June 5−9; Cohen, W. W., McCallum, A., Roweis, S. T., Eds.; Morgan Kaufmann: San Francisco, CA, 2008; pp 96−103.
(50) Ulintz, P.; Zhu, J.; Qin, Z.; Andrews, P. *Mol. Cell. Proteomics* **2006**, *5*, 497−509.
(51) Lin, Y.; Jeon, Y. *J. Am. Stat. Soc.* **2006**, *101*, 578−590.
(52) Breiman, L. *Consistency of a Simple Model of Random Forests*; Technical Report 670 for Statistics Department; University of California: Berkeley, CA, 2004; pp 1−10.
(53) Pardo, M.; Sberveglieri, G. *Sens. Actuators* **2008**, *131*, 93−99.
(54) Cawley, G. C. Baseline methods for active learning. *JMLR Workshop and Conference Proceedings*, Sardinia, Italy, May 16, 2010; Guyon, I., Cawley, G., Dror, G., Lemaire, V., Statnikov, A., Eds.; Journal of Machine Learning Research, 2011; Vol. *16*, pp 47−57.
(55) Guo, Y.; Schuurmans, D. In *Advances in Neural Information Processing Systems (NIPS)*; Neural Information Processing Systems Foundation, 2008; pp 593−600.
(56) Settles, B.; Craven, M. An analysis of active learning strategies for sequence labeling tasks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii, USA, October 25−27; Association for Computational Linguistics: Stroudsburg, PA, 2008; pp 1070−1079.
(57) Fuchs, T.; Buhmann, J. Inter-active learning of randomized tree ensembles for object detection. *3rd IEEE ICCV Workshop on On-line Computer Vision*, Kyoto, Japan, September 27 to October 4; IEEE: New York, 2009; pp 1370−1377.
(58) Cebron, N.; Berthold, M. *Data Min. Knowl. Discovery* **2009**, *18*, 283−299.
(59) Wu, B.; Abbott, T.; Fishman, D.; McMurray, G.; Mor, G.; Stone, K.; Ward, D.; Williams, K.; Zhao, H. *Bioinformatics* **2003**, *19*, 1636−1643.
(60) Eidhammer, I.; Flikka, K.; Martens, L.; Mikalsen, S.-O. *Computational Methods for Mass Spectrometry Proteomics*; John Wiley and Sons: Chichester, England, 2007.