

Eberhard Karls Universität Tübingen
Mathematisch-Naturwissenschaftliche Fakultät
Wilhelm-Schickard-Institut für Informatik

Bachelor Thesis Bioinformatics

Semi-supervised learning for nucleic acid cross-linking mass spectrometry

Emil Paulitz

14.08.2020

Reviewer

Prof. Oliver Kohlbacher
Department of Computer Science
University of Tübingen

Supervisor

Timo Sachsenberg
Address
University of Tübingen

Paulitz, Emilian Nicolaus Simons:

Semi-supervised learning for nucleic acid cross-linking mass spectrometry

Bachelor Thesis Bioinformatics

Eberhard Karls Universität Tübingen

Period: 14.04.2020-14.08.2020

Abstract

Write here your abstract.

Acknowledgements

Leo.org ¹, always finding the right words for me.

¹<https://dict.leo.org/englisch-deutsch/>

Contents

List of Figures	v
List of Tables	vii
List of Abbreviations	ix
1 Introduction	1
1.1 Background	2
1.1.1 Cross-linking	2
1.1.2 Mass Spectrometry	2
1.1.3 Bioinformatics Tool	3
1.1.4 Percolator	4
1.1.5 ROC curves	5
2 Material and Methods	9
2.1 Implementation of the percolator algorithm	9
2.2 Adapting Percolator to Cross-link Identification	11
2.2.1 Different Ranks	11
2.2.2 Characteristics of Cross-linked PSMs	12
2.2.3 Small datasets	14
3 Results	15
3.1 Implementation of the percolator algorithm	15
3.2 Adapting Percolator to Cross-link Identification	16

3.2.1	Different Ranks	17
3.2.2	Characteristics of Cross-linked PSMs	17
3.2.3	Small datasets	20
4	Discussion	23
4.1	Outlook	25
	Bibliography	27

List of Figures

1.1	Example for a mass spectrum	2
1.2	Example for an ROC curve	6
1.3	Examples for pseudo ROC curves	7
2.1	Procedure of Percolator	10
3.1	Results of implementing feature normalization	15
3.2	Examples for pseudo ROC curves as produced by Pycolorator . . .	16
3.3	Results of strategies regarding different ranks	18
3.4	Maintaining proportions works as intended	19
3.5	Results of imputation	19
3.6	Results of splitting the dataset	20
3.7	Correlation between AUC and dataset size	21
3.9	Portion of the dataset with high-confidence PSMs	21
3.10	Found vs expected number of high confidence PSMs	22
3.11	Performance of Pycolorator on smaller datasets	22

List of Tables

3.1	Area under the pseudo ROC curves with and without balancing of classes	17
-----	---	----

List of Abbreviations

MS	Mass Spectrometry
LC	Liquid Chromatography
MS/MS	Tandem Mass Spectrometry
PSM	Peptide Spectrum Match
FDR	False Discovery Rate
ROC	Receiver Operating Characteristics
TPR	True Positive Rate
FPR	False Positive Rate
SVM	Support Vector Machine
AUC	Area Under the Curve

Chapter 1

Introduction

While understanding the human body has provided many i.a. medical advancements, research mostly also reveals unknown facts and connections. As experimental methods advanced, many observed phenomenons could be linked to molecular-scale mechanisms. With new generation DNA sequencing, for example genetic diseases and their exact origin could be identified. While for illnesses like BSE a linkage to misfolded proteins was found, other illnesses like Alzheimer disease, which is also possibly connected to misfolded proteins, lack general treatment and a definite cause as of now [4]. This shows the need for further research and understanding of proteins in biological systems.

Proteomics is the interdisciplinary research field analyzing the composition, interaction and impacts of the proteome (the entirety of proteins) of single cells or up to a whole organism [9, 14]. It has grown in importance to the scientific community over the last decade due to new and refined methods enabling large-scale data acquisition, resulting in the publication of the first human proteome drafts in 2014 [15, 11]. Meanwhile, methods of equally high quality are often missing for the study of many interactions with other (bio-)molecules. In this thesis, research was done to fit one piece of the data acquisition pipeline, the well established Percolator algorithm [8], onto the characteristics of protein-RNA interaction. We will focus on the research to this field based on the study of proteins cross-linked onto RNA.

The scientific background is covered in section 1.1. Chapters 2 and 3 describe the changes done to the Percolator algorithm and their impact. Alongside an outlook for further possible improvements and experiments, these will be discussed in chapter 4, concluding this thesis.

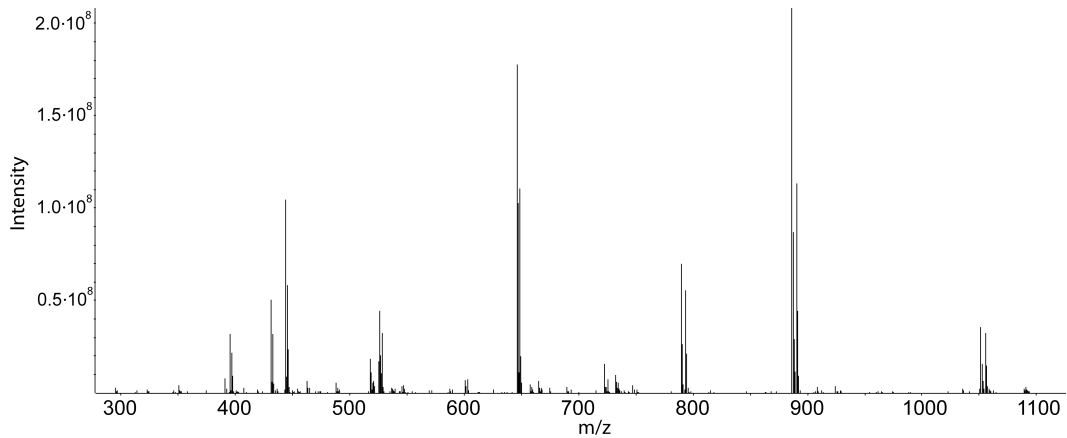


Figure 1.1: Example for a mass spectrum as recorded by a mass spectrometer. The ion intensity correlates with the amount of a molecule in the sample, $\frac{m}{z}$ is the mass-to-charge-ratio. From: Sachsenberg [14]

1.1 Background

1.1.1 Cross-linking

Cross-linking is the process of artificially inducing a chemical bond between different molecules, using for example UV light [14]. The hope of applying this to peptides and RNA, is to receive insight into their *in vivo* interactions, and possibly also conclusions about protein-DNA interaction.

1.1.2 Mass Spectrometry

For quantitatively characterizing the proteome of a sample, large scale measuring techniques are needed. Mostly, mass spectrometry (MS) is used, or more specifically, as for the data in this thesis, tandem mass spectrometry (MS/MS) combined with liquid chromatography (LC). In order to analyze the protein sample with MS, its complexity has to be reduced as much as possible, for example using LC [14]. As Han et al. [9] explains, the mass spectrometer then produces mass spectra, which have to be analyzed further. It does so by first ionizing the substrate, because it can only detect charged particles. Then, the sample is separated in the mass analyzer by the ratio $\frac{m}{z}$, mass of the particles to their charge. The detector then quantifies the amount of a particle in the sample. The result is a mass spectrum, as shown in figure 1.1.

Because mass alone does not give enough information about a peptide to determine its sequence, tandem mass spectrometry is often used to gather more detailed evidence. In this procedure, particles of similar $\frac{m}{z}$ ratio

are selected for fragmentation in a collision cell after the first round of mass measurement [14]. In there, the substance collides with a gas to be broken down into smaller molecules. For proteins, fragmentation happens predominantly in their backbone, producing all possible sub-sequences of the peptide. This yields a spectrum that is almost unique for its protein, which allows for peptide identification using bioinformatics tools [2].

1.1.3 Bioinformatics Tool

Algorithms like Sequest [6] or X! Tandem [5] compare the resulting spectra with theoretical spectra calculated from a list of possible peptides and compute a score based on their similarity. The peptides are generated by obtaining a list of proteins expected in the sample and calculating the peptides resulting from the, for example enzyme-based, degradation of the proteins. The best scoring peptide is then considered a peptide-spectrum-match (PSM). The scores produced by those algorithms often do not distinguish well enough between correct and incorrect matches [10], but they enable FDR estimation using decoy databases and serve as a basis for score re-calibration with the Percolator algorithm [10, 8].

Decoy databases are created from the target database, contain usually as many peptides [13, 12] in a reversed or shuffled order with respect to the amino acid sequence [1]. They are presented to the scoring algorithm either separately [8] or mixed with the target database [13]. It is assumed, that decoy and target peptides have similar features [12] and are not easily distinguishable by a scoring algorithm. When the actually fitting peptide for a given spectrum is not in the target database, and thus a wrong one will be chosen, the best scoring peptide will be a decoy approximately half of the time. This allows for an estimation of wrongly assigned targets, since the score distribution is assumed to be the same for decoys and false targets [1].

In practice, one estimates the probability of a PSM being a false target by counting the number of decoy-PSMs with the same or a higher score. It is then assumed, there are as many false targets and thus a false discovery rate (FDR) can be estimated [8]. This leads to the following formula¹:

$$FDR = \frac{\# \text{ false target PSMs}}{\# \text{ all target PSMs}} \approx \frac{\# \text{ decoy PSMs}}{\# \text{ all target PSMs}} \quad (1.1)$$

¹In this thesis, the following approximation is used:

$FDR \approx \frac{\# \text{ decoy PSMs}}{\# \text{ all PSMs}} = \frac{\# \text{ decoy PSMs}}{\# \text{ decoy PSMs} + \# \text{ target PSMs}}$
 It is faster to calculate and yields results differing by the FDR, so in the relevant range of FDRs of 0 to 5% up to 5%:

$$\frac{\frac{\# \text{ decoys}}{\# \text{ targets}}}{\frac{\# \text{ decoys}}{\# \text{ decoys} + \# \text{ targets}}} = \frac{\# \text{ decoys} + \# \text{ targets}}{\# \text{ targets}} = 1 + \frac{\# \text{ decoys}}{\# \text{ targets}} \approx 1 + FDR$$

The q-value as a measure for a single PSM rather than a metric for a set of PSMs is then derived from this as the minimum FDR of all PSMs with a lower or equal score [8, 1]. It will be used for estimating the credibility for any one PSM.

1.1.4 Percolator

As Käll et al. [10] say, separating correct from incorrect target PSMs with already mentioned algorithms works fine, but there is still room for improvement. This is because often not all information is used and considered jointly. Percolator [10, 8] tries to utilize as much information as possible by using scores from different algorithms, features of the peptide (like its length), of the spectrum, or the PSM itself. It joins them using a linear support vector machine (SVM) and a semi-supervised approach with cross-validation to retain as many PSMs as possible. In every iteration, the top ranking, non-decoy PSMs up to a certain threshold of q-value are chosen as positive training examples, and the decoy PSMs are used as negative training set. The PSMs are then re-ranked using the SVMs score with the intention of getting a better separation of true and false PSMs. If this holds true, the positive training set of the next iteration better is of higher quality and the SVM can be trained even better. The algorithm usually converges within the first 10 iterations [10].

As explained by Bishop [3], an SVM is a maximum margin classifier model. In the linear case, it tries to generate a hyperplane inside the multi-dimensional space that splits the given training data into the specified classes. For generalization performance, it tries to maximize the distance or margin between the hyperplane and the nearest training points, called support vectors. If the data is not linearly separable, it allows but punishes outliers by their distance to the hyperplane: An outlier inside the margin but on the correct side of the hyperplane is not penalized as badly as if it was on the wrong side of the hyperplane, and the further away from the hyperplane a point lies on the wrong side, the more this gets penalized. The strength of the penalization is also controlled by the parameter C , which can also be defined differently for points of different classes².

To avoid having to split the data into training and testing set and consequently losing possibly correct PSMs but also avoid overfitting, a nested cross-validation approach is being used. As Granholm et al. [8] explain, overfitting is a central problem in machine learning. It happens when the model fits itself onto random variation in the data, not actually connected to a general pattern and performing poorly when presented new data.

²<https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html>

Cross-validation means splitting the data into a number smaller batches (Percolator uses 3), presenting all but one to the model as training and the remaining part to validate the results, or, in the case of Percolator, to score the PSMs. For every part of the data a new model is trained by every part but the one to be scored, and the scores of different parts are normalized in the end. Additionally, Percolator performs cross-validation with every training batch, evaluating different parameters for the machine learning model applied to this specific batch of data [8].

1.1.5 ROC curves

As well explained by Fawcett [7], performance of classification machine learning models is often evaluated using receiver operating characteristics (ROC) curves. They are graphs in which the true positive rate of a model on the y-axis against its false positive rate on the x-axis. An example is shown in figure 1.2. Generally, a classifier producing an ROC curve further to the northwest than the curve of another classifier, is better, because it achieves a higher true positive rate at a lower false positive cost. But, to be able to compare classifiers more directly and by using a single scalar value, often the area under an ROC curve is computed and used as a metric, because it represents the probability that "the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance" ([7], page 868).

When dealing with the problem of rating PSMs however, there is no way of knowing the ground truth and thus precisely estimating a classifiers performance. Instead, the total number of PSMs that are estimated to be correct when accepting a certain q-value are plotted against that q-value [10, 8]. The graphs shape resembles that of an ROC curve, and is thus called pseudo ROC. However, there are significant differences and neither the number of accepted PSMs nor the q-value are the same as their ROC counterparts: The q-value of a certain PSM, being related to the FDR, is the (estimated!) portion of wrongly accepted PSMs out of every positively classified PSM, when using the PSMs score as threshold. The false positive rate on the other hand would be the percentage of wrongly accepted PSMs out of every wrong PSM in the whole dataset. With the confusion matrix notation from Fawcett [7], this can be written as:

$$FDR = \frac{FP}{Y}; \quad FPR = \frac{FP}{n} \quad (1.2)$$

With this notation, the true positive rate (TPR) is $\frac{TP}{p}$ and the total number of accepted PSMs is Y . From this, a completely different scale arises (since $Y \in \mathbb{N}$ and $TPR \in [0, 1]$). The AUC, which is still used as a metric, thus does

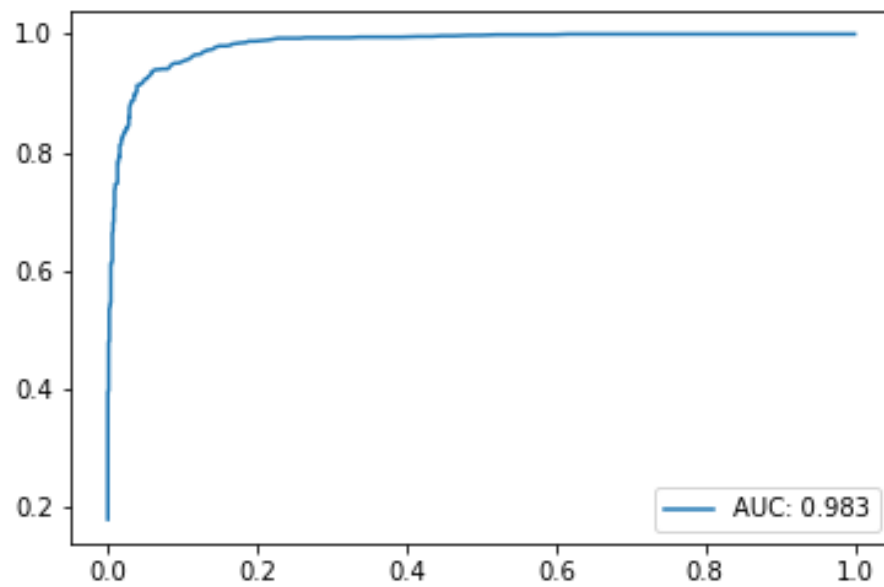
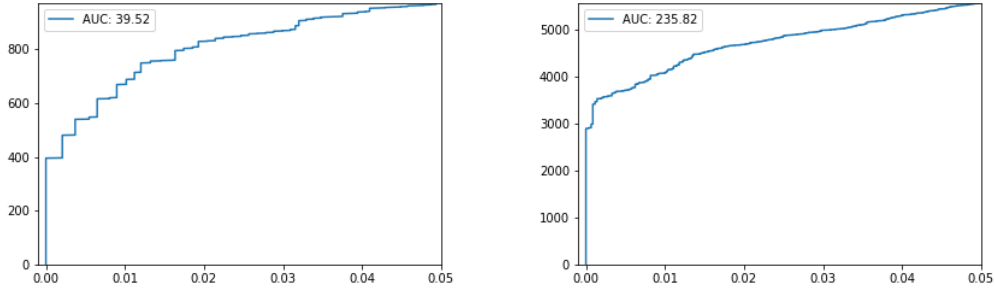


Figure 1.2: An Example for an ROC curve. It was plotted from a generated dataset with 1000 positive and negative entries respectively, being "scored" using a log normal distribution with different expected values. Calculations were done by the `pseudoROC` function as described in 2.1.



(a) This pseudo ROC was generated from the same generated dataset as was used for the ROC curve in 1.2. (b) This pseudo ROC is generated from the realistic dataset described in 2, using the given NuXL-score.

Figure 1.3: Examples for pseudo ROC curves, plotted using the function `pseudoROC` as described in 2.1.

not range from 0 to 1 and does not represent a probability, but can only be used as a relative score.

ROC curves improve when the data contains more examples with features clearly indicating the negative class membership to the classifier, because the FPR of all other examples decreases. Since pseudo ROCs are plotted only for q-values regarded interesting, often in $[0, 0.05]$, they are insensitive to such changes in the data. This allows including lower-scoring PSMs in the dataset (like is done in this thesis) without diluting the evaluation. Examples for pseudo ROCs are shown in figure 1.3.

Chapter 2

Material and Methods

To be able to test the following implementations, a dataset containing 93,219 PSMs was used. Its spectra result from UV-light cross-linked extract from the nucleus of HeLa cells, and these were matched against the nucleus proteins. For each spectrum, up to 4 matching peptides are recorded. The dataset was generated with OpenNuXL 1.0.

Of the PSMs, 51,521 are cross-linked and 41,698 linear peptides, 50,627 are target and 42,592 are decoy peptides. Every PSM has 52 features, 2 columns specifying the spectrum and 2 columns describing the peptide and the protein it originated from, as well as one column indicating to the experimenter whether one entry is a target or a decoy PSM. The features contain the NuXL-score, which was computed from the scores of different algorithms (see 1.1.3), and was used as a reference.

2.1 Implementation of the percolator algorithm

The Percolator algorithm as presented in figure 2.1 was implemented in python. The pandas¹ package was used for dataset handling, and for SVM calculations and inner cross-validation the classes LinearSVC² and GridSearchCV³ from the python sklearn package were used. The outer cross-validation was implemented utilizing the numpy array_split⁴ method. An early stopping criterion was implemented, terminating the algorithm if the area under the pseudo ROC curve does not improve over a certain number of iterations. This number is

¹<https://pandas.pydata.org/>

²<https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html>

³https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

⁴https://numpy.org/doc/stable/reference/generated/numpy.array_split.html

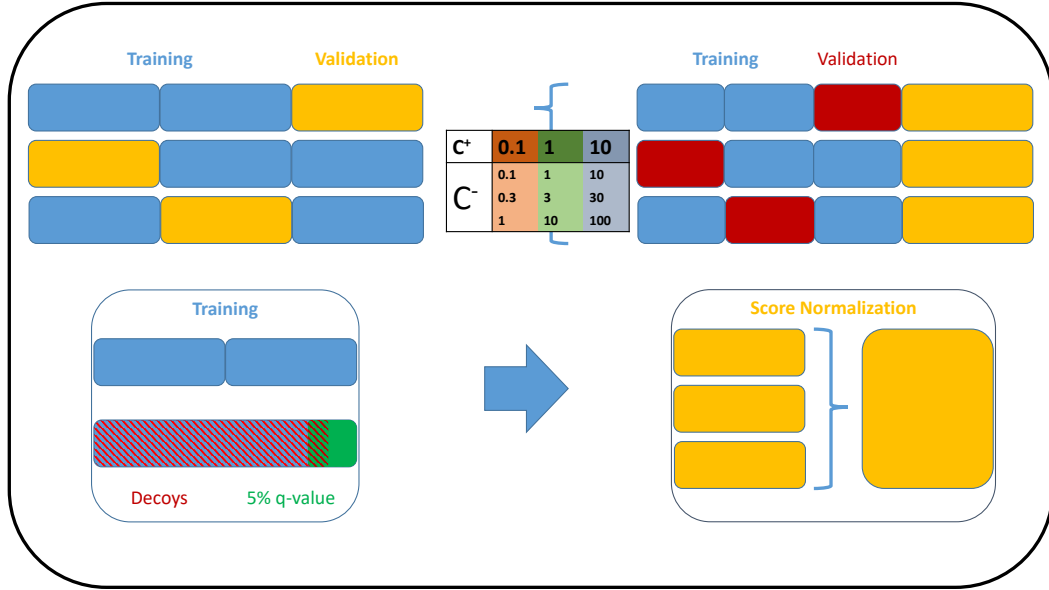


Figure 2.1: Rough outline of the Percolator algorithm. First, the dataset is split in three parts, of which two are used for training and one for scoring. Then, another cross-validation is performed on the training part to determine the best parameter combination. Percolator searches the grid spanned by the options for the C parameter for positive examples C^+ (0.1, 1 and 10) and negative examples C^- (1, 3 or $10 \times C^+$). The parameter combination achieving best results on the training set is then used to train an SVM that scores the validation set. For training, all target PSMs with a q-value of $\leq 5\%$ are used as positive and all decoy PSMs are used as negative examples. Because in the end, the dataset will have been scored by three different SVMs, these scores have to be normalized. This is done by a linear transformation mapping the score representing a q-value of 1% to 0 and the median decoy score to -1 . If **qScore** denotes the minimum score for which a q-value of 1% is estimated and **dMedian** the median score a decoy achieves in the validation set, this sets score is transformed by the following equation: $\text{new Score} = \frac{\text{old score} - \text{qScore}}{\text{qScore} - \text{dMedian}}.$

controlled by the parameter `termWorseIters` and has a default of 4.

As preprocessing, the provided file is read in as a pandas dataframe (converting strings to floating point or integer values if possible) and q-value and ranks are calculated as discussed in 1.1.3 and 2.2.1 respectively. Also, the columns containing features are normalized. This means, their value is mapped linearly onto $[0, 1]$ to avoid rounding errors in features of different orders of magnitude, caused by the representation of floating point numbers.

For the calculation of pseudo ROC curves and the area under this curve, a function `pseudoROC` was implemented. It retrieves the entries of the given datasets q-value column, enumerates them and plots the enumeration against the q-values. This results in a pseudo ROC curve. Plotting was performed using the `matplotlib.pyplot`⁵ package, and the area under the pseudo ROC curve was calculated by `sklearn.metrics.auc`⁶ method.

To distinguish the original Percolator algorithm from the version implemented in python and altered here, the latter will be called Pycolorator.

2.2 Adapting Percolator to Cross-link Identification

To be able to monitor the difference any experiment makes, especially with respect to the cross-linked or non-cross-linked PSMs, following features were implemented:

First, in addition to the q-value, which is calculated as described in 1.1, the calculation of a class-specific q-value was implemented. This is done by splitting the dataset according to the class affiliation and calculating the q-value separately for both splits. This allows a more precise performance estimation when looking at only one of the classes.

Second, a ROC curve using the `pseudoROC` function (2.1) is plotted after every iteration of Pycolorator: one for the whole dataset, one only for cross-linked, and one only for non-cross-linked PSMs. Accordingly, the respective class-specific q-value is used. Thus, three plots covering the corresponding class(es) and every iteration, as well as the area under the curve are output. This allows for fast visual detection of the impact a specific change to the algorithm has on certain classes, iterations or general sensitivity.

2.2.1 Different Ranks

As experience shows, cross-linked peptides can be harder to detect than linear peptides. This means, the possibly correct cross-linked peptide will frequently

⁵https://matplotlib.org/api/pyplot_api.html

⁶<https://scikit-learn.org/stable/modules/generated/sklearn.metrics.auc.html>

not get the highest score of all the peptides. It thus can be beneficial to not only consider the highest scoring peptide, but also the highest scoring cross-linked peptide or also some lower-scoring peptides, and assign them ranks. Then, as following experiments show, Percolator can correct the scores of some lower-ranking PSMs, possibly detecting more cross-linked PSMs. Meanwhile it is known to the experimenter, that only one of the peptides can be the correct match to a spectrum, and thus any PSM with a lower rank than 1 should be excluded at the end.

To tackle both constraints, Pycolorator first trains the SVM with every PSM available and re-calculates the ranks based upon the newly assigned score after every iteration, possibly correcting the ranks of some PSMs. When the used metric, normally the area under the pseudo ROC curve, does not improve beyond a certain threshold per iteration, every PSM with rank 2+ is dropped. The threshold is controlled by the parameter `cutOffImprove` with a default of 0.01 corresponding to a 1% increase of the used metric per iteration. Since also some of the best scoring PSMs will be dropped (even though they were not assigned rank 1), the algorithm then runs further until the maximum iterations are reached or the score does not further improve, in order to properly integrate and score the new PSMs considered confident.

The performance of this feature was tested against dropping the lower ranking PSMs once at the very end of the algorithm and once at the very beginning. Pycolorator was run on the given dataset (2) and pseudo ROCs were plotted as described in 2.2.

2.2.2 Characteristics of Cross-linked PSMs

Apart from being hard to detect, cross-linked peptides also have other characteristics, some of which pose problems to the computational detection of correct PSMs. As discussed in 2.2.2 "Splitting the dataset", the features of cross-linked and linear peptides are so dissimilar, splitting the dataset and training a linear SVM separately on cross-links and non-cross-links yields significantly better results than training one linear SVM. To reduce the impact of this heterogeneity, the following experiments were conducted:

Proportions of Different Classes

As discussed in 1.1.4, Percolator employs a nested cross-validation approach, splitting the dataset, training on all parts than one and testing/scoring on the remaining part. If the splitting was uneven by chance, the SVM would be trained badly and the scoring inaccurate. Having different classes with significant differences in the dataset, like in our case cross-linked and linear PSMs or targets and decoys, increases this problem. For example, if there was a testing split with many cross-linked PSMs, the SVM had to be trained on

the remaining data with few cross-linked PSMs, resulting in probably poor scoring of the many cross-linked PSMs in the test set. In the average case, this should not be a problem, but it can occasionally produce worse or better results, which would follow from overfitting.

To solve this problem, a mechanism of maintaining the proportion of the classes in the whole dataset for every inner and outer split was implemented. It can be toggled for targets and decoys or cross-linked and non-cross-linked PSMs as well as for inner and outer split independently. The impact has been measured by running the algorithm 10 times, plotting and recording the results of the best, worst and median run w.r.t. to the q-value any number of PSMs was identified at. Because this took approximately 90 minutes, the test was performed in a Google Colaboratory notebook⁷.

Imputation

Cross-linked PSMs naturally have features linear PSMs do not have, which can however be used for training the SVM. An example would be the nucleotide it was linked to, or the partial loss score, which is the X! Tandem [5] hyper score for the cross-linked peaks. In the given dataset (2), 16 out of 61 features were only given for cross-linked PSMs. Optimally, these should not influence the score a non-cross-linked PSM gets. However, 0 was filled in for the missing values and because that is a valid value for the linear SVM, it biases the decision made. For example, if a high value in a feature leads the SVM to a decision against the PSM, 0 as the lowest value possible after feature normalization (2.1) will tell the SVM to give the PSM a higher score. The procedure of replacing missing values with numbers minimizing the bias is called imputation. The scikit-learn package provides methods for the implementation in python⁸, and one of these, the `IterativeImputer` function, was tested.

Splitting the Dataset

Because the SVM is linear, one feature can not alter the influence another feature has on the decision. Thus, splitting the dataset into cross-linked and non-cross-linked PSMs should allow the SVM to fit more precisely onto the characteristics of each. This was tested by running Pycolorator on each split and the concatenate the datasets, recalculating the q-values from the Pycolorator scores. It was also tested splitting the dataset by the nucleotide the peptide was cross-linked onto, and non-cross-linked PSMs as a fifth class.

⁷The Google Colaboratory notebook used:
https://colab.research.google.com/drive/1VqZAmtdta57YhgobA0WkQMqe_U9YIUUnD1?usp=sharing

⁸<https://scikit-learn.org/stable/modules/impute.html>

2.2.3 Small datasets

The portion of cross-linked peptides in a dataset is often very small. To evaluate when splitting as proposed in 2.2.2 "Splitting the Dataset" is possible and obtain general insight into the scalability of Pycolorator, two experiments were conducted to conclude how small a dataset may be, so that the Percolator algorithm still works. In each case, Pycolorator was applied to a dataset sampled from the given one (2). The area under the pseudo ROC curve and the number of PSMs identified at 1% q-value were recorded and used as metrics. The results achieved by Pycolorator were compared to the effect the splitting had on the q-value estimation utilizing the given, precomputed NuXL score.

First, the smaller dataset was sampled using every 2^i -th PSM from the whole dataset with $i \in \mathbb{N} \wedge i \in [0, 13]$. Second, the smaller dataset was sampled randomly using a uniform distribution from all PSMs with a q-value of $< 10\%$. The portions sampled were $\frac{1}{2^i} : i \in \mathbb{N} \wedge i \in [0, 12]$. Additionally to the metrics mentioned above, also the expected number of identifications at 1% q-value for each split were calculated and recorded. This was done by counting the number of such PSMs before re-estimating the q-value or letting Pycolorator run, thus being the number of identifications by the NuXL score if the q-value is estimated based on the whole dataset. To compensate for the randomness in this experiment, Monte-Carlo-Sampling with 10 iterations was performed and the results are presented with boxplots.

Because numerical problems occur when calculating the area under a pseudo ROC curve of a very small dataset (since the curve consists of only one point), another metric was introduced: the number of identified PSMs at a q-value threshold of 1%. Per default, Pycolorator decides automatically after its first iteration if this metric is required based on the dataset, and adjusts its log, plots per iteration and functions like the early stopping criterion accordingly.

- (- Performance auf anderem Datensatz
- Vergleich mit Entrapment FDR)

Chapter 3

Results

3.1 Implementation of the percolator algorithm

The original Percolator algorithm finds 5412 cross-linked PSMs with a q-value of $\leq 1\%$ on the used dataset (2). The re-implementation in python yields 5503 such PSMs on rank 1.

The implementation of feature normalization yielded an increase in the number of confident PSMs found, as the y-axis and legend of figure 3.1 show. This type of plot is explained in section 3.2. The performance of Pycolorator with feature normalization is better from the first iteration on, and increases until iterations 7 or 8, when the results converge. Without feature normalization, the area under the pseudo ROC increases until iteration 3, when it begins to oscillate.

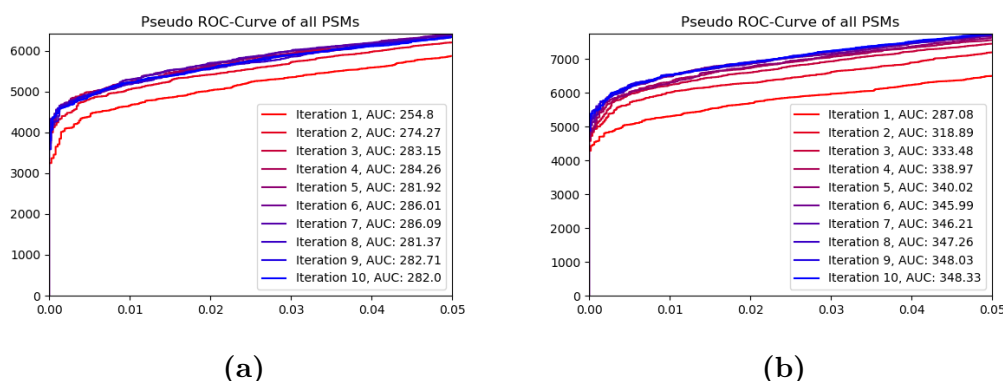


Figure 3.1: Result of Pycolorator before (3.1a) and after (3.1b) the implementation of feature normalization.

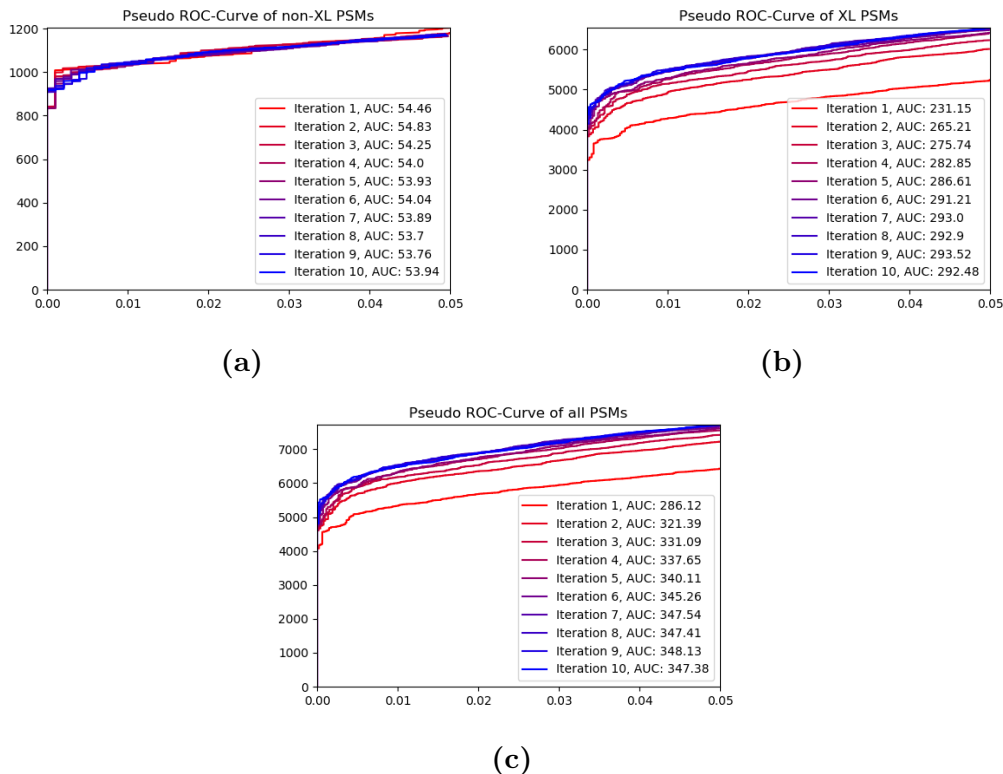


Figure 3.2: Example of the plots output by Pycolorator, generated by running with default parameters on the given dataset (2). 3.2a shows the result of non-cross-linked PSMs, 3.2b that of cross-linked and 3.2c that of all PSMs in the dataset.

3.2 Adapting Percolator to Cross-link Identification

An example of the plots created by Pycolorator are shown in figure 3.2. 3.2a shows the pseudo ROC of the non-cross-linked PSMs in every iteration, 3.2b the cross-linked and 3.2c all of the PSMs. In each, every iterations pseudo ROC curve has its own color, fading from red in iteration 1 to blue in iteration 10. The colors can be assigned from the legend, which also shows the area under the depicted pseudo ROC curve. Every curve ranges from q-values 0 to 5% and the y-axis is scaled appropriate for the data. As explained in 1.1.5, a higher and more left (pseudo) ROC curve represents better classification. Thus, figure 3.2c shows that the result of Pycolorator improves over most iterations.

3.2.1 Different Ranks

Figure 3.3 shows the pseudo ROCs of Pycolorator before the implementation of the new mechanism. 3.3a was plotted when Pycolorator used all PSMs available, meaning all ranks, and only at the end lower ranking PSMs than 1 were excluded. 3.3b is the result of running Pycolorator only with rank 1 PSMs of the given dataset 2 and 3.3c shows the final result with the new feature.

As one can see, without the new mechanism Pycolorator gradually improves the area under the curve and takes 5 and 4 iterations to roughly converge when given every PSM or only rank 1 PSMs to train respectively. This can be seen by the AUC results after every iteration or the ROC curves themselves. The end result, however, is better when removing the lower ranking PSMs right at the beginning (AUC of 345.79) instead of in the end (AUC of 343.21 after dropping the lower ranking PSMs). Since in the latter case Pycolorator ended with an AUC of 343.79 as one can see in 3.3a, dropping the PSMs slightly worsened the end result. With the new feature, the algorithm takes about 5 iterations to converge, then performs a jump and again improves over two iterations. The end result, an AUC of 348.13, is higher than both alternatives. Multiple runs yielded very similar results.

3.2.2 Characteristics of Cross-linked PSMs

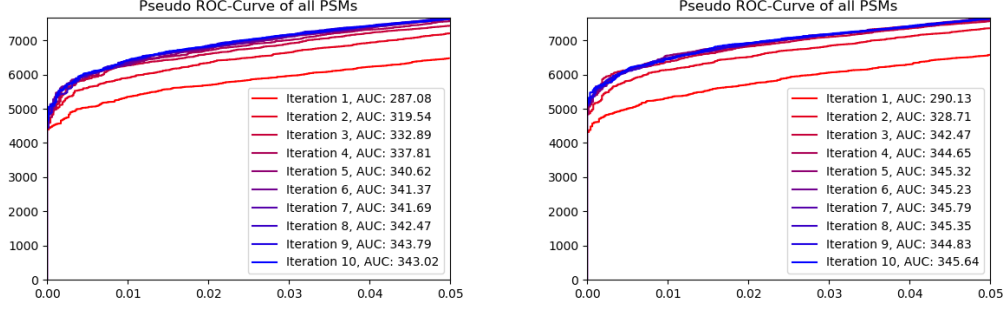
Proportions of Different Classes

Maintaining the proportion of classes works as intended, as figure 3.4 shows. It depicts the proportion of targets (t) to decoys (d) in each of the three outer splits and the whole dataset.

The evaluation produced the results shown in table 3.1.

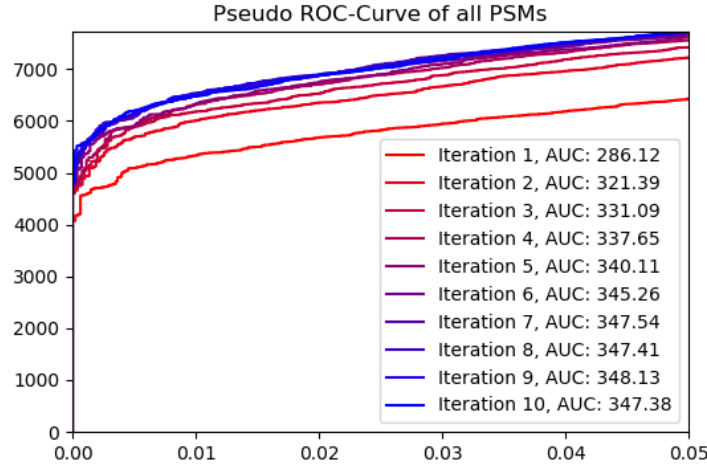
Table 3.1: Shows the area under the pseudo curves of cross-linked, non-cross-linked and all PSMs in the best, worst and median run w.r.t. the AUC of all PSMs. Pycolorator was run with and without balancing of target/decoy and cross-linked/non-cross-linked classes in outer and inner splits 10 times each.

AUC of pseudo ROC		maximum	minimum	median
balancing	cross-linked	293.91	293.12	293.58
	non-cross-linked	53.85	53.76	53.84
	all PSMs	348.65	347.48	348.19
no balancing	cross-linked	294.42	291.51	292.86
	non-cross-linked	53.63	54.34	53.78
	all PSMs	349.14	347.04	347.69



(a) The resulting pseudo ROCs (explained in 3.2) if Pycolorator is given every PSM regardless of its rank. Lower ranking PSMs are only dropped in the end, which results in a final AUC of 343.21.

(b) The resulting pseudo ROCs (explained in 3.2) if Pycolorator is given only the top scoring peptide for every spectrum. Lower ranking PSMs are dropped before running the algorithm.



(c) The resulting pseudo ROCs (explained in 3.2) if Pycolorator is run with the newly implemented feature. It first trains with every PSM available, and after the results converge, lower ranking PSMs are dropped. Then, the algorithm runs for some more iterations.

Figure 3.3: Results of running Pycolorator with different strategies as to how lower ranking PSMs are dealt with.

```
p = percolator_experimental(dSlow, idCol, features, I=2, plotEveryIter = True, fastCV = True)
```

```
t/d= 1.1899358658115442
t/d= 1.1733930195145834
t/d= 1.2028214943995463
t/d for the whole df= 1.1886504507888804
```

(a) Ratios of target to decoy PSMs in each of the three splits, ranging from 1.173 to 1.203, while the ratio is 1.8865 in the whole dataset.

```
p = percolator_experimental(dSlow, idCol, features, I=2, plotEveryIter = True, propTarDec = True, balancingOuter = True,
```

```
t/d= 1.1886181152274968
t/d= 1.1887018384165668
t/d= 1.1886314010002113
t/d for the whole df= 1.1886504507888804
```

(b) Ratios of target to decoy PSMs in each of the three splits, ranging from 1.18862 to 1.18870, while the ratio is 1.18865 in the whole dataset.

Figure 3.4: The log of two test runs of Pycolorator. It calculated and output the ratio of targets to decoys in every of the outer split and the whole dataset. In 3.4b the proportions were maintained, and in 3.4a they were not.

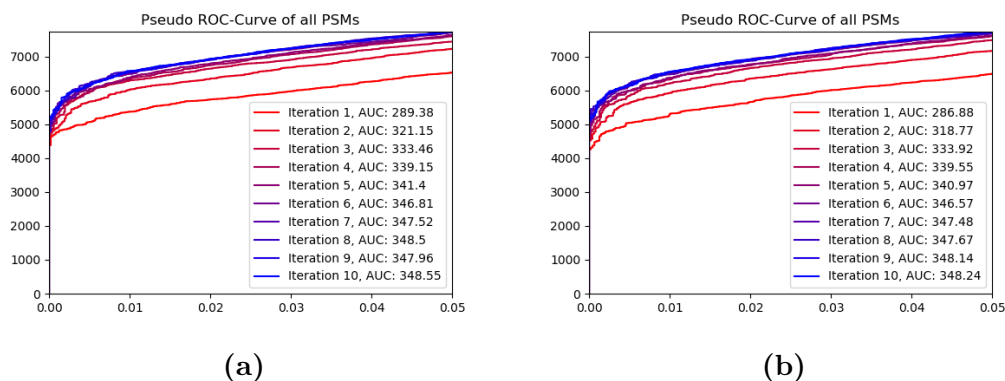


Figure 3.5: Results of running Pycolorator with (3.5b) and without (3.5a) the use of imputation.

Imputation

The resulting pseudo ROCs and their AUC for all PSMs are shown in figure 3.5. The pseudo ROCs for cross-linked and not cross-linked PSMs show similarly few differences.

Splitting the Dataset

The results of splitting the dataset before running Pycolorator are shown in figure 3.6.

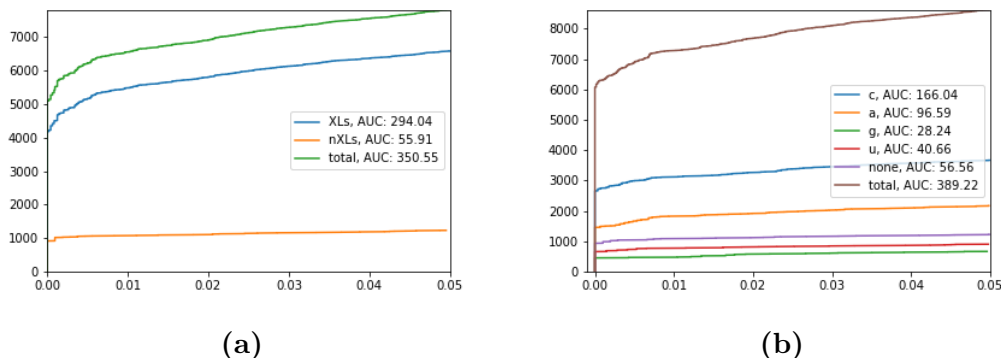


Figure 3.6: Resulting pseudo ROCs and their AUC for each split and the combined dataset when splitting the dataset by 3.6a) cross-linked and non-cross-linked PSMs, 3.6b) the cross-linked nucleotide. XLs stands for cross-linked PSMs and nXLs for non-cross-linked PSMs.

3.2.3 Small datasets

When measuring the AUC in the first experiment, a roughly linear correlation to the datasets size was found. However, the AUC could not be measured when the dataset became smaller than approximately 200 PSMs, as shown in figure 3.7. The number of PSMs at 1% FDR was larger relative to dataset size when using a smaller portion of the dataset, whether using a fixed score (3.8a) or percolator (3.8b). In the case of percolator, with smaller dataset size also more fluctuations occurred.

The second experiment yielded following results: figure 3.9 shows the portion of the dataset that was originally estimated to have a q-value of $< 1\%$. The ratio of how many PSMs were assigned such a q-value after the sampling to before, is shown in figure 3.10a without re-calculating the score and in figure 3.10b when running Pycolorator on the sampled data and estimating the q-value from the obtained score. Figure 3.11 shows how well the Pycolorator score performed compared with the precomputed NuXL score when measured by the identifications with 1% q-value or area under the pseudo ROC curve.

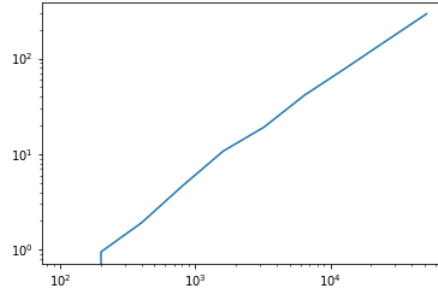


Figure 3.7: The AUC of a Pycolorator run on the y-axis is plotted against the size of the dataset sampled as described in 2.2.3 on the x-axis. Below about 200 PSMs in the dataset, the pseudo ROC consisted of data at one q-value and thus enclosed an area of 0.

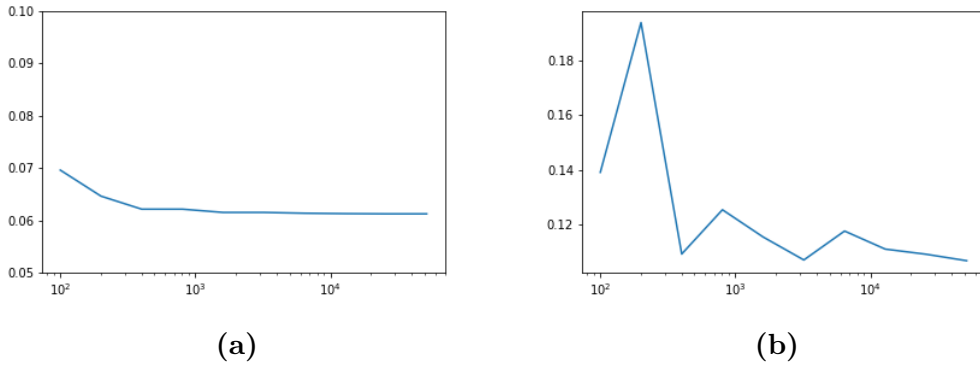


Figure 3.8: The portion of PSMs in the sampled dataset with a 1% q-value on the y-axis is plotted against the datasets size on the x-axis. Left (3.8a) when estimating the q-value based on the unchanged NuXL score, and right (3.8b) after running Pycolorator on the sampled dataset.

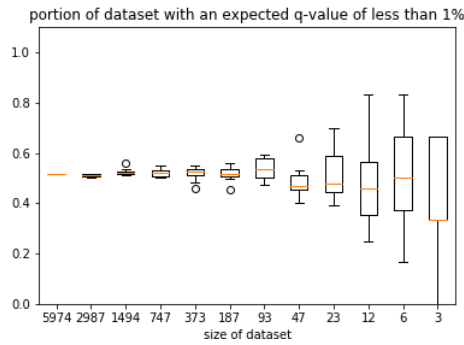


Figure 3.9: Portion of the sampled dataset from the second experiment of high confidence PSMs ($q\text{-value} \leq 1\%$) on the y-axis is plotted against the datasets size on the x-axis. The data from every of the 10 repetitions is shown combined in a boxplot.

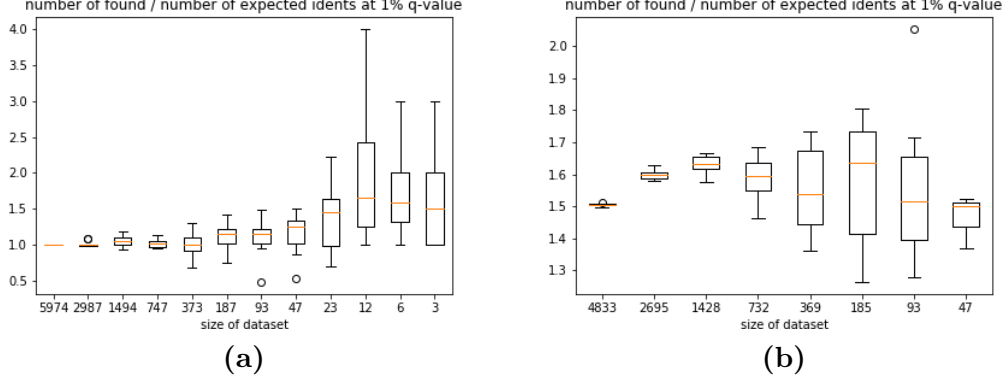


Figure 3.10: Ratio of PSMs with $q\text{-value} \leq 1\%$ found after to before recalculating the FDR in the sampled dataset on the y-axis, is plotted against the datasets size on the x-axis. For 3.10b, the newly calculated Pycolorator score was used to re-calculate the FDR. The data from every of the 10 repetitions is shown combined in a boxplot. In 3.10a, the variance but also the ratio increases for a smaller dataset, while figure 3.10a shows a larger variance, fluctuation and a slight increase of the ratio for smaller datasets.

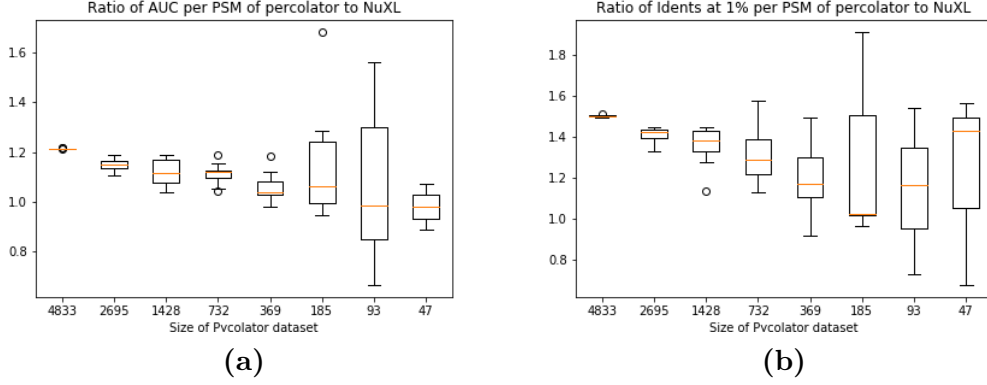


Figure 3.11: The ratio of Pycolorators performance to the unchanged NuXL score on the y-axis is plotted against the datasets size on the x-axis. Left (3.11a), to measure the performance the area enclosed by the pseudo ROC is used, right (3.11b) the number of PSMs with a $q\text{-value} \leq 1\%$. The data from every of the 10 repetitions is shown combined in a boxplot. While the median AUC increases for smaller datasets, Pycolorator does not identify more PSMs with a confidence of 1% than the NuXL score on datasets with 47 or less PSMs.

Chapter 4

Discussion

As already explained in 2.2.1, cross-linked peptides often are harder to score than linear ones. Therefore, they can get a lower score than appropriate and it can be useful to also include the best scoring cross-linked peptide when running the Percolator algorithm. It may be able to revise the PSMs scores and the actually correct cross-linked peptide may become the best scoring PSM. This thesis is also supported by the findings in 3.2.1. If only including the best scoring peptide, of which 3.3b shows the results, the end result is worse than when giving Pycolorator some iterations to re-rank the found PSMs. However, it was better than when giving Pycolorator all of the PSMs available, which is unexpected, since giving a machine learning model more information should generally improve its learning. Apparently, the lower ranking PSMs, even when having such a high score a q-value of $\leq 5\%$ is estimated, contain misleading information and thus the SVM learns patterns not valid for correct PSMs. This also explains why the algorithm converges faster when only given rank 1 PSMs. The higher quality of data lets the SVM learn the correct patterns after fewer iterations.

The pseudo ROC generated when using the newly implemented mechanism (3.3c) shows a convergence after 5 iterations, just like when using every PSM (3.3a). Then, as the log shows, lower ranking PSMs are dropped and the next iteration has a much higher AUC, probably as a result of the better quality of the PSMs. Letting the algorithm run on the new dataset again improves the AUC even beyond that of Pycolorator when only using rank 1 PSMs. This suggests, that indeed a re-ranking takes place in the first half of iterations.

Comparing the AUC of Pycolorator when run with the new mechanism (3.3c) after iteration 6 (345.26) with the end result of running Pycolorator with every PSM available (3.3a, 343.21), yields the following insight: Dropping all PSMs with rank 2+ yields a worse result when the Percolator algorithm has been running for 10 rather than 6 iterations. This implies an overfitting onto the PSMs with lower quality and thus a worse scoring.

Although implementing the balancing of different classes in the nested cross-validation splits had no big impact, it reduced the spread. On other datasets, containing much fewer cross-linked than non-cross-linked PSMs, this procedure could have greater impact. This needs to be tested in further experiments. The feature has no disadvantages except negligible worse performance and can be generalized to other classes that could be present in a dataset.

Using imputation did not improve the performance of Pycolorator. Neither for a single class nor for both. It is thinkable that imputation has a different effect on other datasets, which contain far fewer cross-links than non-cross-links. It could then prevent that the SVM fits onto the characteristics of non-cross-links because they are more frequent and find very few cross-links, but rather fit onto both classes in the same way. This has to be tested in further experiments. Because the imputation takes a significant amount of time, this option is switched off per default.

Splitting the dataset by cross-linked and non-cross-linked PSMs slightly improved the performance of Pycolorator, and splitting by the cross-linked nucleotide yielded a great improvement. In practice, neither is applicable to most datasets, because they often only contain a very small number of cross-linked PSMs. As the experiments in section 3.2.3 show, applying Pycolorator on a small dataset bring a lot of variation.

The non-random approach in the first experiment regarding small datasets keeps the ratio of correct to incorrect PSMs and ensures the appearance of PSMs originally classified as correct. However, maintaining the ratio when sampling a very small subset means it will contain one or two PSMs, which originally were assigned a q-value of $< 1\%$, and many worse PSMs. This general bad quality of the dataset makes it difficult to obtain quantitative results for the very small datasets, which is why the second experiment was conducted. The findings however do suggest that down to a dataset size of about 500, Pycolorator does not suffer from any drawbacks. With smaller datasets, the AUC metric does not work anymore and a lot of variation happens (3.8b). Only the first problem could be solved: by implementing the new metric of identifications at 1% q-value. Figure 3.8a was generated to test if the FDR estimation would suffer from small datasets. This can not be observed, as the smallest dataset for which a point is plotted has approximately 100 PSMs, and thus the small increase from 0.061 to 0.07 is due to rounding.

In the second experiment however, this thesis seems to be supported by figure 3.10a, showing that when re-calculating the q-value after sampling, without changing the score, more PSMs receive a high confidence. But: very few decoys have a high score, aber die lassen den q-value von vielen targets unter 1% runtschen. Wenn man jetzt random samplet, ist die wahrschein-

lichkeit immer kleiner, dass viele von diesen decoys mitkommen, je kleiner der Datensatz wird. Darunter kommen aber noch viele high scoring targets, die nachrücken können. differing lengths of datasets dxl and pxl is because of filtration of ranks

- Ist Pycolator besser geworden?
- Neues Diagramm zeichnen
- Methoden hinterfragen oder begründen, Ergebnisse interpretieren, Anwendbarkeit diskutieren, z.B.:
- Falsche Formel für q-value
- C Parameter für jeden split neu optimieren führt zu overfitting? → Original-Algorithmus macht es auch so
- Wie sinnvoll ist die neue Metrik (idents bei 1%)?
- ScanNr Versuche: Gleiche Spektren (identifiziert anhand der ScanNr.) auf verschiedene splits verteilen verändert nichts, d.h. vermutlich sind die niedrigeren Ränge dann so schlecht, dass es nichts bringt die schonmal gesehen zu haben.
- Peptide Versuche: Schlechtere Ergebnisse, aber vllt ehrlicher?

4.1 Outlook

- Mögliche weiterführende Experimente: mächtigere Klassifikatoren + monotonic constraints (wie von Timo ausprobiert), Ada-Boosting, feature selection

Bibliography

- [1] Suruchi Aggarwal and Amit Kumar Yadav. False discovery rate estimation in proteomics. In *Methods in Molecular Biology*, pages 119–128. Springer New York, 2016. doi: 10.1007/978-1-4939-3106-4_7. URL https://doi.org/10.1007/978-1-4939-3106-4_7.
- [2] Thomas E. Angel, Uma K. Aryal, Shawna M. Hengel, Erin S. Baker, Ryan T. Kelly, Errol W. Robinson, and Richard D. Smith. Mass spectrometry-based proteomics: existing capabilities and future directions. *Chemical Society Reviews*, 41(10):3912, 2012. doi: 10.1039/c2cs15331a. URL <https://doi.org/10.1039/c2cs15331a>.
- [3] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer Science + Business Media, LLC, 2006. ISBN 978-0387-31073-2.
- [4] A. Burns and S. Iliffe. Alzheimer’s disease. *BMJ*, 338(feb05 1):b158–b158, February 2009. doi: 10.1136/bmj.b158. URL <https://doi.org/10.1136/bmj.b158>.
- [5] R. Craig and R. C. Beavis. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–1467, February 2004. doi: 10.1093/bioinformatics/bth092. URL <https://doi.org/10.1093/bioinformatics/bth092>.
- [6] Jimmy K. Eng, Ashley L. McCormack, and John R. Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 1994.
- [7] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, June 2006. doi: 10.1016/j.patrec.2005.10.010. URL <https://doi.org/10.1016/j.patrec.2005.10.010>.
- [8] Viktor Granholm, William Noble, and Lukas Käll. A cross-validation scheme for machine learning algorithms in shotgun proteomics. *BMC*

- Bioinformatics*, 13(Suppl 16):S3, 2012. doi: 10.1186/1471-2105-13-s16-s3. URL <https://doi.org/10.1186/1471-2105-13-s16-s3>.
- [9] Xuemei Han, Aaron Aslanian, and John R Yates. Mass spectrometry for proteomics. *Current Opinion in Chemical Biology*, 12(5):483–490, October 2008. doi: 10.1016/j.cbpa.2008.07.024. URL <https://doi.org/10.1016/j.cbpa.2008.07.024>.
- [10] Lukas Käll, Jesse D Canterbury, Jason Weston, William Stafford Noble, and Michael J MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods*, 4(11):923–925, October 2007. doi: 10.1038/nmeth1113. URL <https://doi.org/10.1038/nmeth1113>.
- [11] Min-Sik Kim, Sneha M. Pinto, Derese Getnet, Raja Sekhar Nirujogi, Srikanth S. Manda, Raghothama Chaerkady, Anil K. Madugundu, Dhanashree S. Kelkar, Ruth Isserlin, Shobhit Jain, Joji K. Thomas, Baby-lakshmi Muthusamy, Pamela Leal-Rojas, Praveen Kumar, Nandini A. Sahasrabuddhe, Lavanya Balakrishnan, Jayshree Advani, Bijesh George, Santosh Renuse, Lakshmi Dhevi N. Selvan, Arun H. Patil, Vishalakshi Nanjappa, Aneesha Radhakrishnan, Samarjeet Prasad, Tejaswini Subbannayya, Rajesh Raju, Manish Kumar, Sreelakshmi K. Sreenivasamurthy, Arivusudar Marimuthu, Gajanan J. Sathe, Sandip Chavan, Keshava K. Datta, Yashwanth Subbannayya, Apeksha Sahu, Soujanya D. Yelamanchi, Savita Jayaram, Pavithra Rajagopalan, Jyoti Sharma, Krishna R. Murthy, Nazia Syed, Renu Goel, Aafaque A. Khan, Sartaj Ahmad, Gourav Dey, Keshav Mudgal, Aditi Chatterjee, Tai-Chung Huang, Jun Zhong, Xinyan Wu, Patrick G. Shaw, Donald Freed, Muhammad S. Zahari, Kanchan K. Mukherjee, Subramanian Shankar, Anita Mahadevan, Henry Lam, Christopher J. Mitchell, Susarla Krishna Shankar, Parthasarathy Satishchandra, John T. Schroeder, Ravi Sirdeshmukh, Anirban Maitra, Steven D. Leach, Charles G. Drake, Marc K. Halushka, T. S. Keshava Prasad, Ralph H. Hruban, Candace L. Kerr, Gary D. Bader, Christine A. Iacobuzio-Donahue, Harsha Gowda, and Akhilesh Pandey. A draft map of the human proteome. *Nature*, 509(7502):575–581, May 2014. doi: 10.1038/nature13302. URL <https://doi.org/10.1038/nature13302>.
- [12] Roger E. Moore, Mary K. Young, and Terry D. Lee. Qscore: An algorithm for evaluating SEQUEST database search results. *Journal of the American Society for Mass Spectrometry*, 13(4):378–386, April 2002. doi: 10.1016/s1044-0305(02)00352-5. URL [https://doi.org/10.1016/s1044-0305\(02\)00352-5](https://doi.org/10.1016/s1044-0305(02)00352-5).
- [13] Junmin Peng, Joshua E. Elias, Carson C. Thoreen, Larry J. Licklider, and Steven P. Gygi. Evaluation of multidimensional chromatography coupled

- with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *Journal of Proteome Research*, 2(1):43–50, February 2003. doi: 10.1021/pr025556v. URL <https://doi.org/10.1021/pr025556v>.
- [14] Timo Sachsenberg. Computational methods for mass spectrometry-based study of protein-rna or protein-dna complexes and quantitative metaproteomics. 2017. URL <http://hdl.handle.net/10900/83311>.
- [15] Mathias Wilhelm, Judith Schlegl, Hannes Hahne, Amin Moghaddas Gholami, Marcus Lieberenz, Mikhail M. Savitski, Emanuel Ziegler, Lars Butzmann, Siegfried Gessulat, Harald Marx, Toby Mathieson, Simone Lemeer, Karsten Schnatbaum, Ulf Reimer, Holger Wenschuh, Martin Mollenhauer, Julia Slotta-Huspenina, Joos-Hendrik Boese, Marcus Bantscheff, Anja Gerstmair, Franz Faerber, and Bernhard Kuster. Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502):582–587, May 2014. doi: 10.1038/nature13319. URL <https://doi.org/10.1038/nature13319>.

Selbständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe und dass alle Stellen, die dem Wortlaut oder dem Sinne nach anderen Werken entnommen sind, durch Angaben von Quellen als Entlehnung kenntlich gemacht worden sind. Diese Bachelorarbeit wurde in gleicher oder ähnlicher Form in keinem anderen Studiengang als Prüfungsleistung vorgelegt.

Ort, Datum

Unterschrift