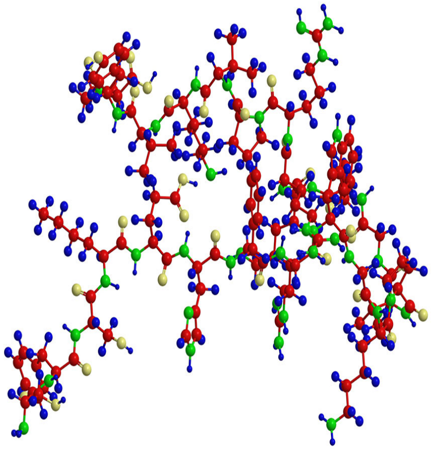
Peptide Detection Using Convolutional Neural Network

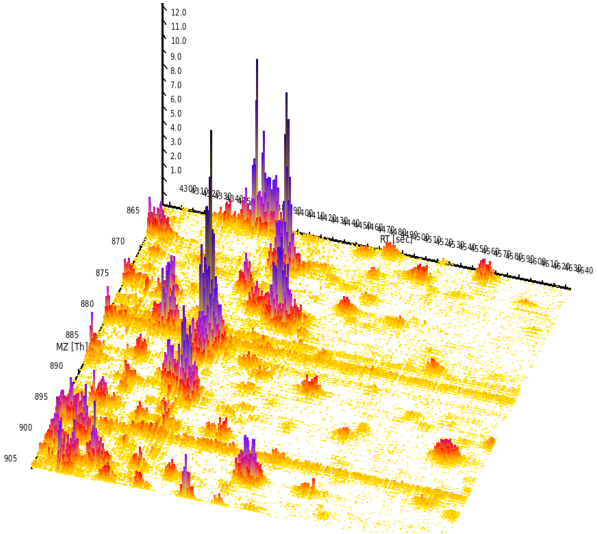
Zaid Ur Rehman

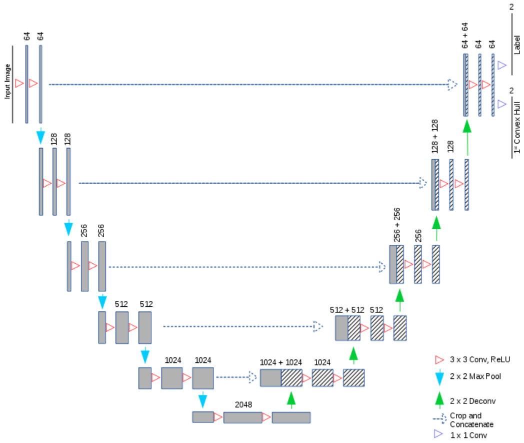
Albert-Ludwigs-Universität Freiburg

Motivation



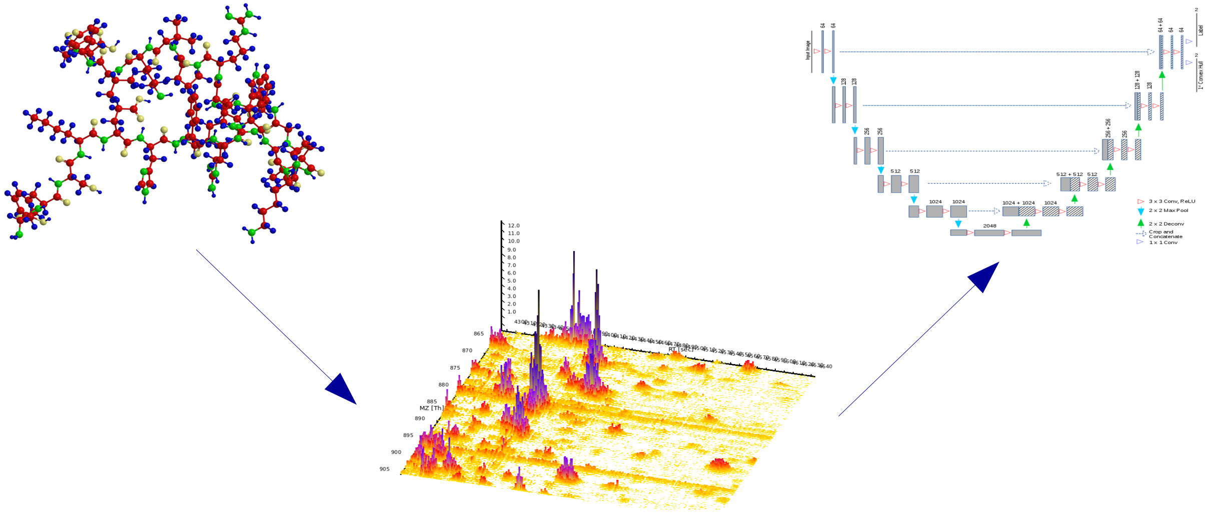
Complete molecular characterization by Mass Spectrometry of Proteomic samples





Neural Network can be trained to detect peptides in spectra

Motivation (contd.)

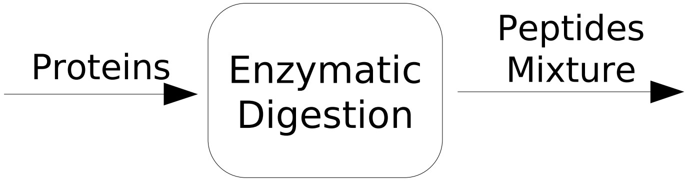


- Mass spectrometer produce huge amounts of data
- Without sequence information, characteristic patterns in protein fragments have to be identified
- Feature finding algorithms need parameter tuning by an expert
- A neural network can leverage huge amount of data without any manual tuning

Outline

1. Background
   a) Mass Spectrometry Based Proteomics Experiment
   b) MS Spectra
   c) MS Data Analysis Tools
   d) Convolutional Neural Network
2. Data Insight
3. Data Preprocessing
4. Rendered Images
5. Evaluation
   a) Binary Classification
   b) Evaluation Metrics
6. Results
   a) Quantitative Results
   b) Qualitative Results
   c) Sparse Region Analysis
7. Conclusion
8. Future Work

Mass Spectrometry Based Proteomics Experiment

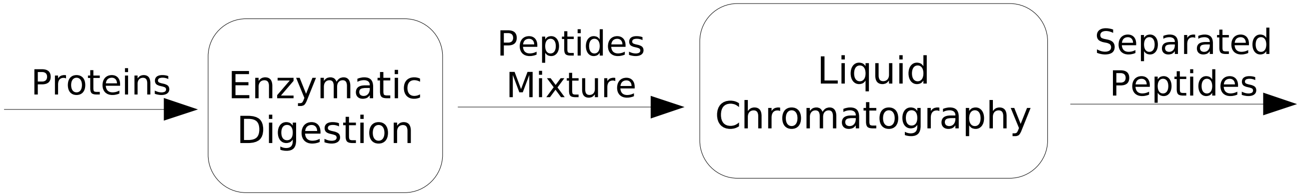


> CA146_HUMAN (protein accession Q5VVC0)

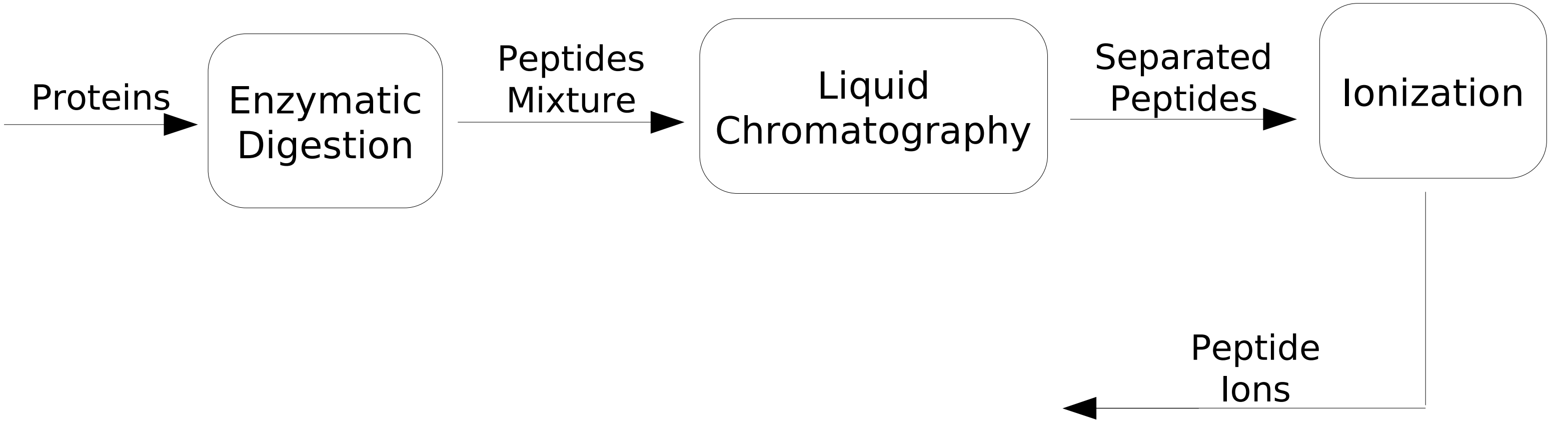
```
MAESGKEKIKWTTTIIISSSLKSYEVATALENRSHKVRYSDSV
ENGSIIFSLSGVAFLLMDTKECLLSTEEIFLAKIEKFINIHQN
SFLVLSAALHGPEEWKLMFRIQQRFLGCNLRILPVHNTVNAIN
LMCTIAKTTSKPYIDSICYRMITAKAYIIEQSPVWKTLQKIKL
NSDSVNPN
```

```
MAESGKEK
IKWTTTIIISSSLK
SYEVATALENRSHK
VR
YSDSVENGSIIFSLSGVAFLLMDTK
ECLLSTEEIFLAK
IEK
FINIHQNSFLVLSAALHGPEWK
LMFR
IQQRFLGCNLR
ILPVHNTVNAINLMCTIAK
TTSKPYIDSICYR
MITAK
AYIIEQSPVWK
TLQK
IK
LNSDSVNPN
```

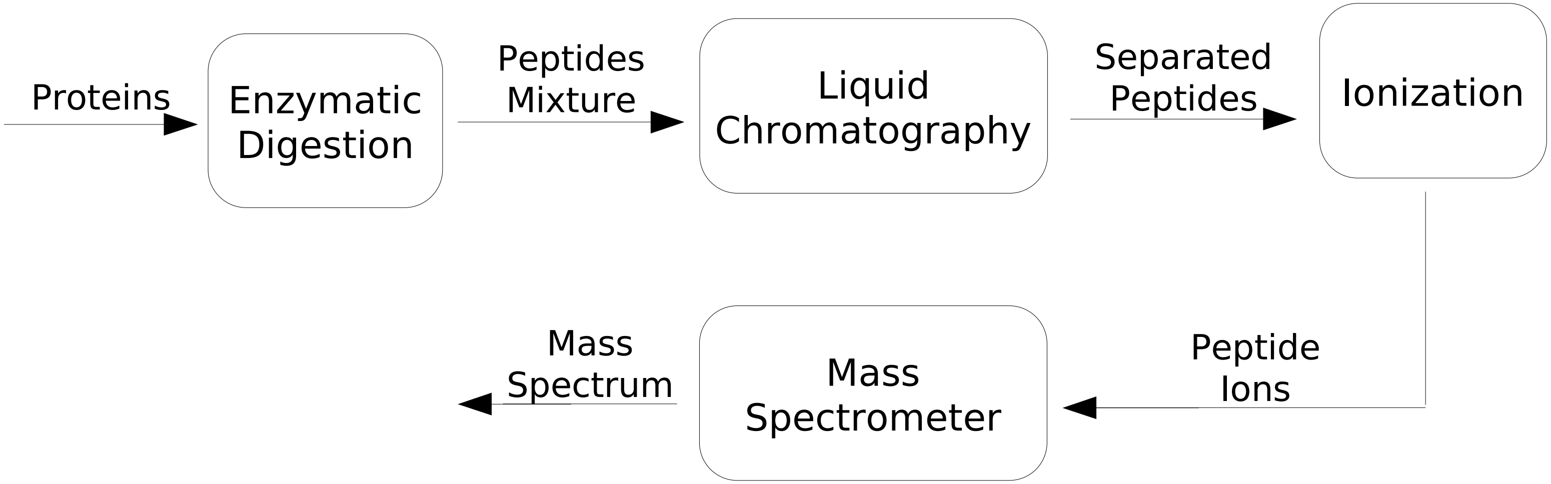
MS Based Proteomics Exp (contd.)



MS Based Proteomics Exp (contd.)



MS Based Proteomics Exp (contd.)



Proteins

Enzymatic Digestion

Peptides Mixture

Liquid Chromatography

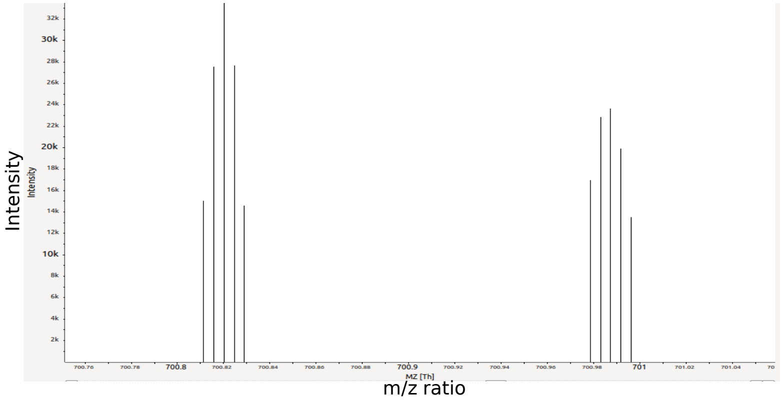
Separated Peptides

Ionization

Peptide Ions

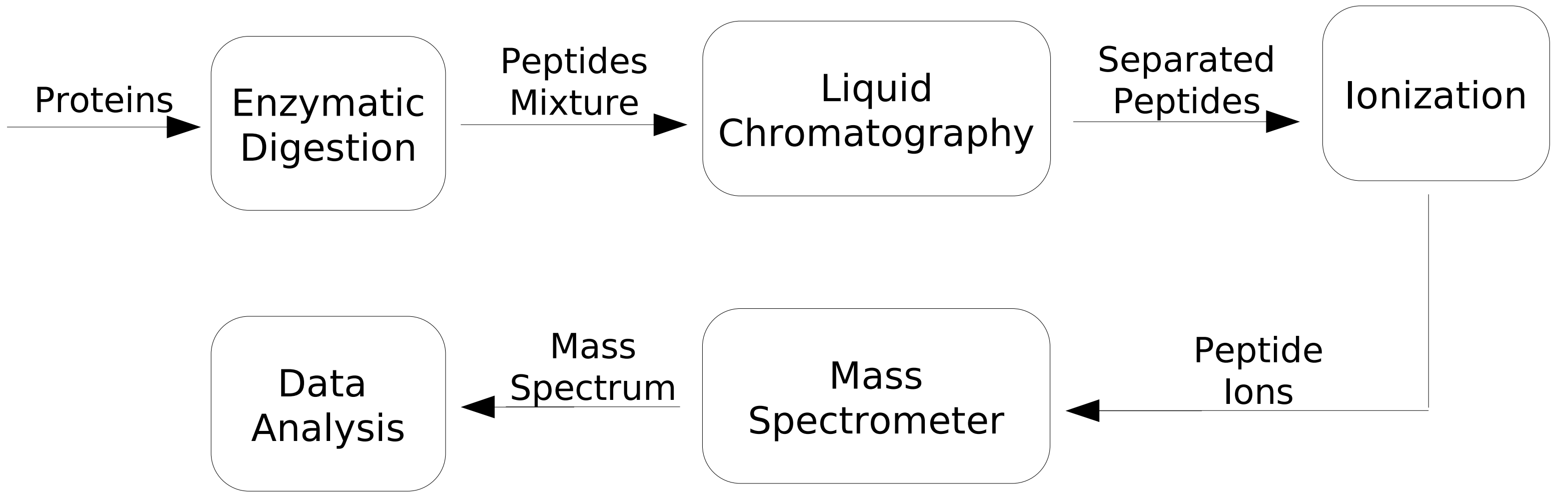
Mass Spectrometer

Mass Spectrum

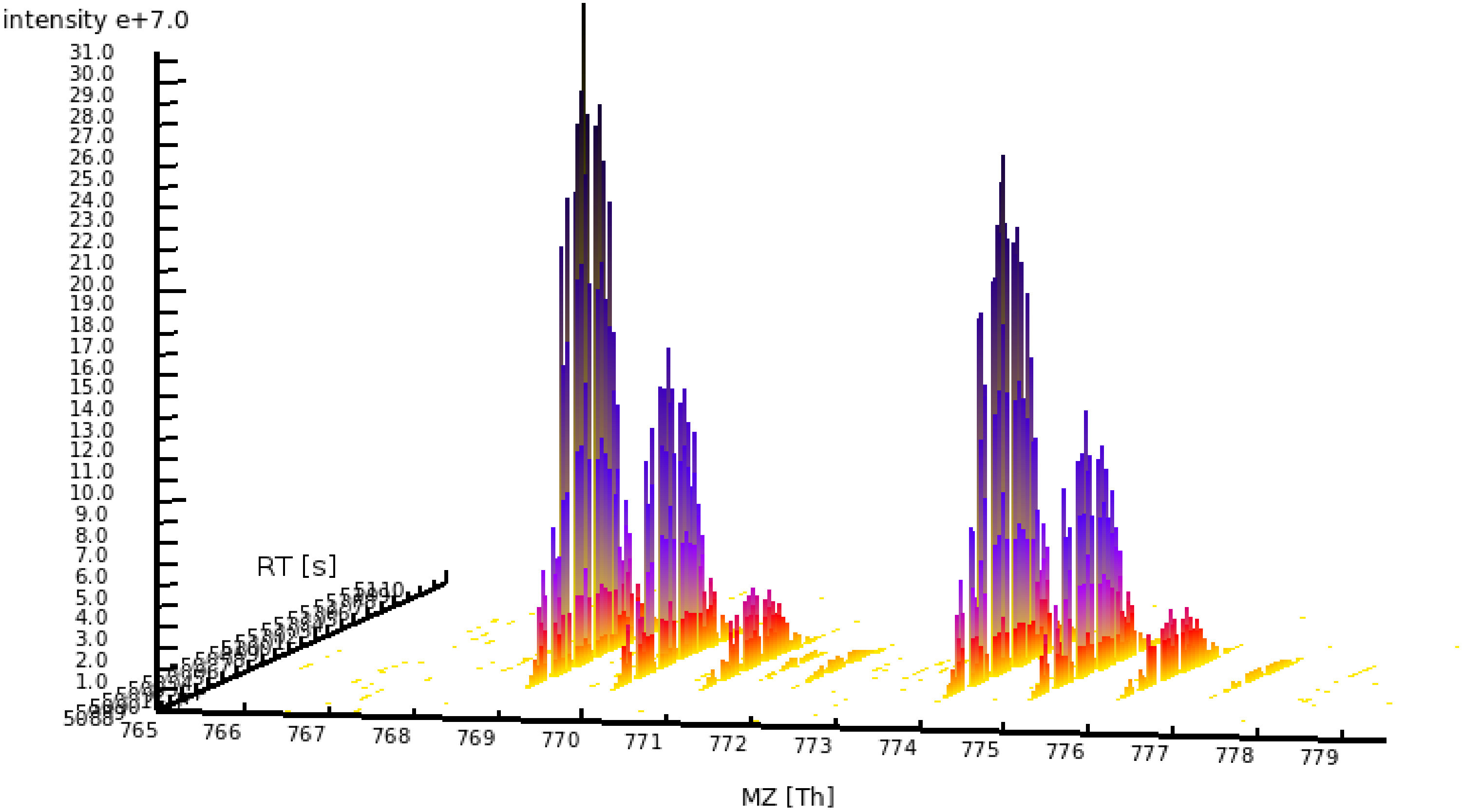


A single mass spectrum measures signal intensity of ions at each *m/z* value

MS Based Proteomics Exp (contd.)

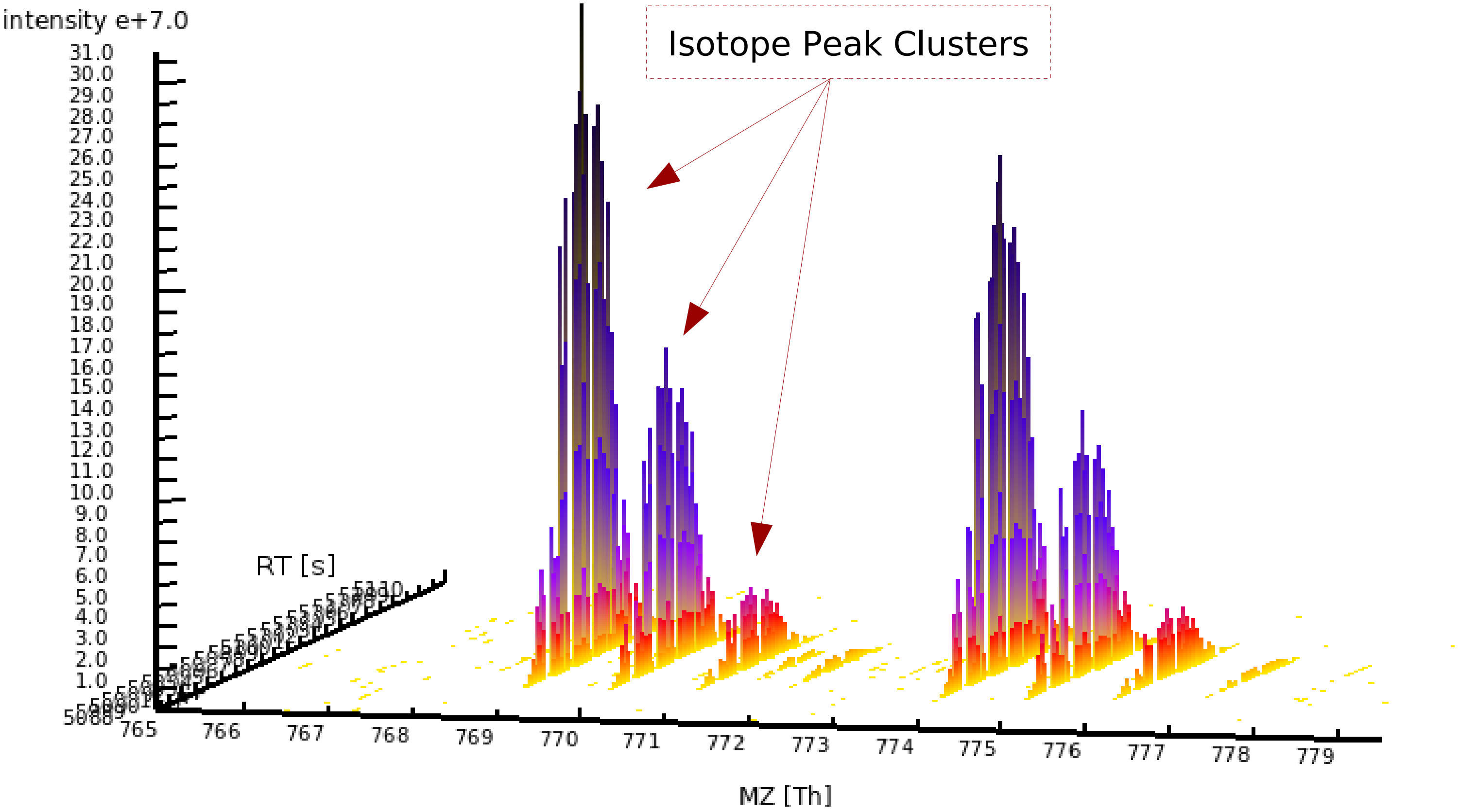


MS Spectra



- A 3D plot represents a MS Experiment
- RT: retention time, measured in seconds
- MZ: mass-to-charge ratio, measured in Thomson (Th)

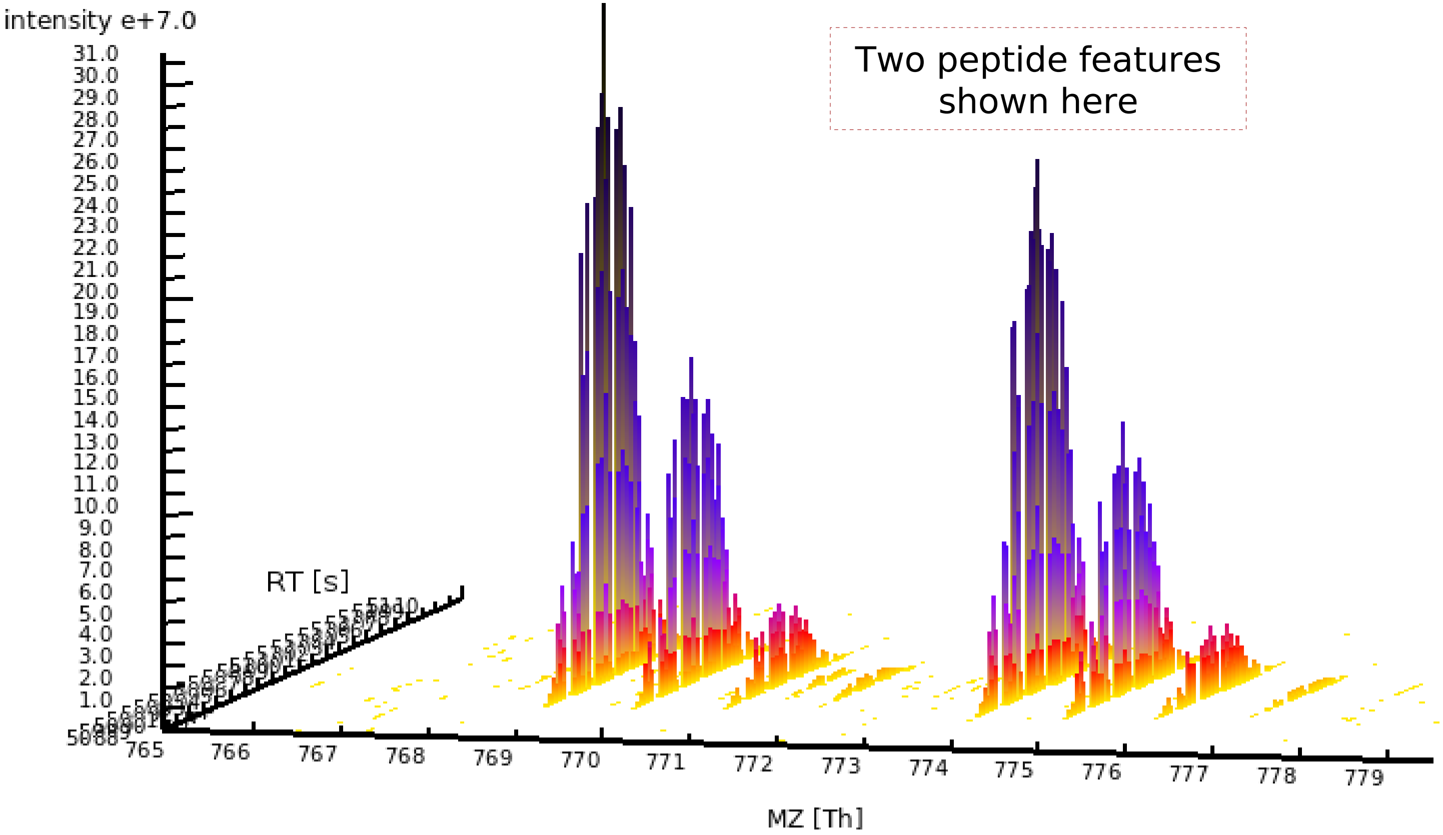
MS Spectra (contd.)



Isotope Peak Clusters

- Presence of C^{13} isotope adds a mass of 1 Da
- MZ offset depends on charge

MS Spectra (contd.)



Two peptide features shown here

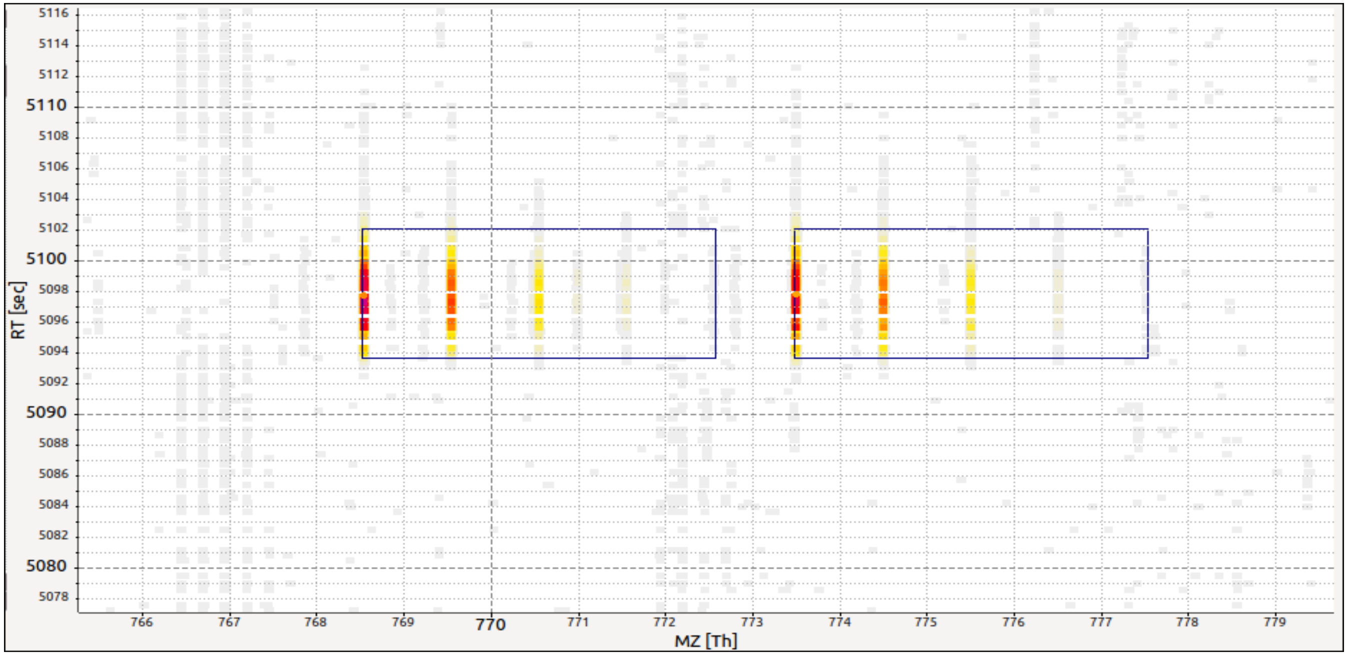
- Peptides features can be detected due to the characteristic isotope peaks pattern

OpenMS [1]

- An open source software platform for mass spectrometry data analysis
- The tools used during this project:
  - TOPPView : for viewing mass spectra [2]
  - FeatureFinderMultiplex : for detecting peptide features in mass spectra
  - MSSimulator : for generating simulated mass spectra [3]
  - pyOpenMS: a python interface for data pre- and post-processing [4]

FeatureFinderMultiplex (FFM)

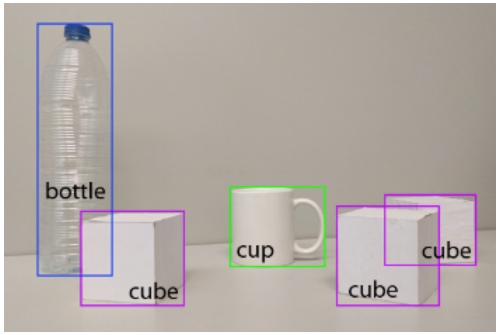
- Classical feature finding algorithm in OpenMS
- Does not rely on peptide sequence information
- Parameters were tuned for available MS data
- Used for generating ground truth for real data



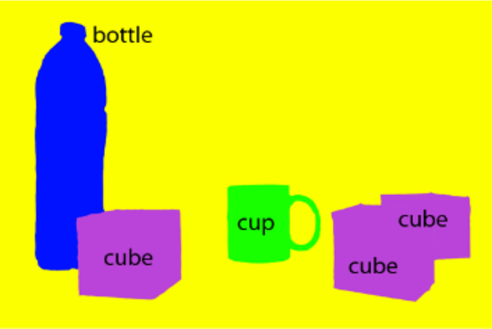
Convolutional Neural Networks (CNN)

- Take advantage of spatial structure of data
- Widely used for visual recognition tasks, examples shown below
- Reformulate the peptide detection problem as semantic segmentation

Image Classification



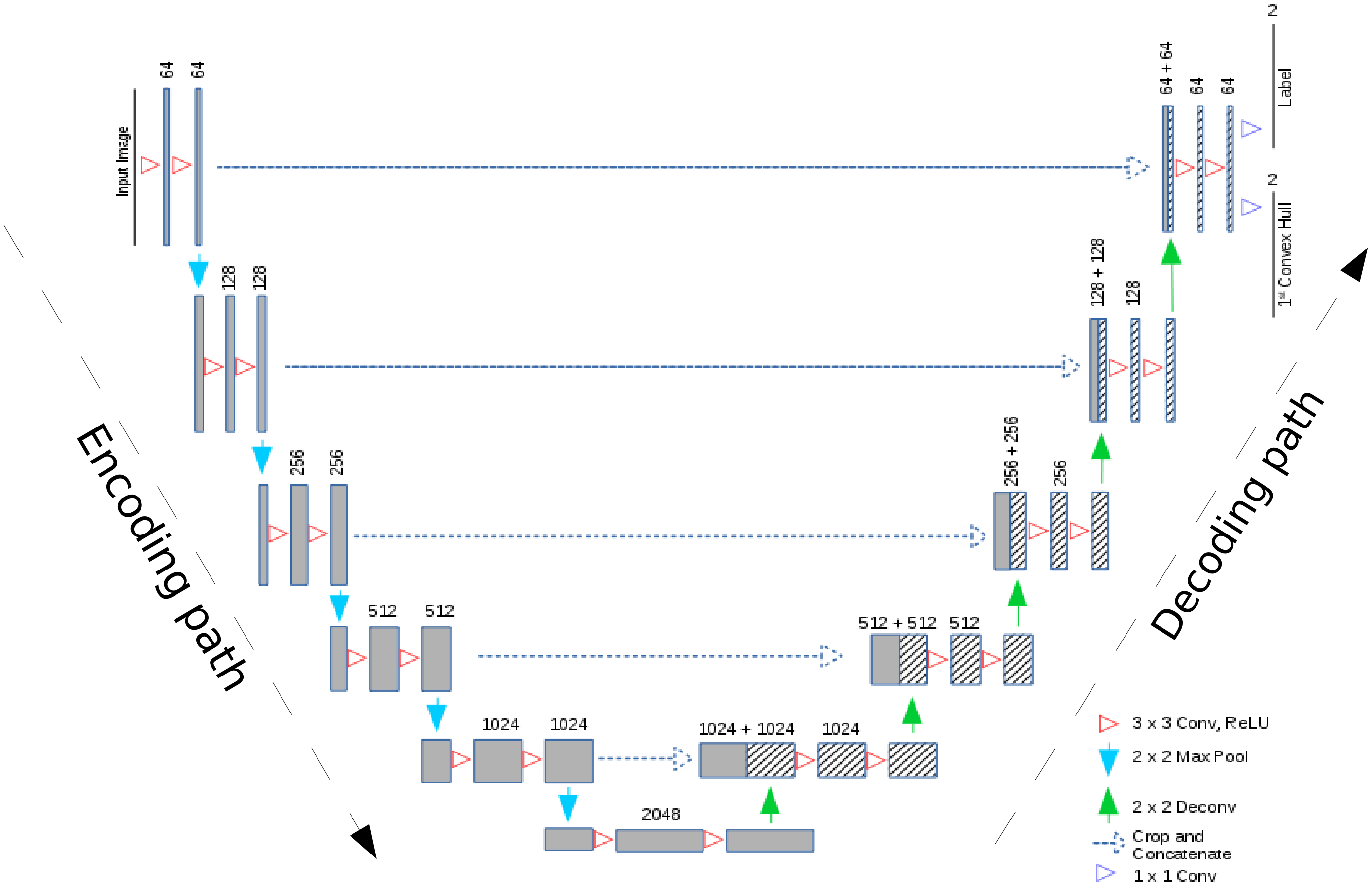
Object Localization



Semantic Segmentation

Images from [5]

U-Net



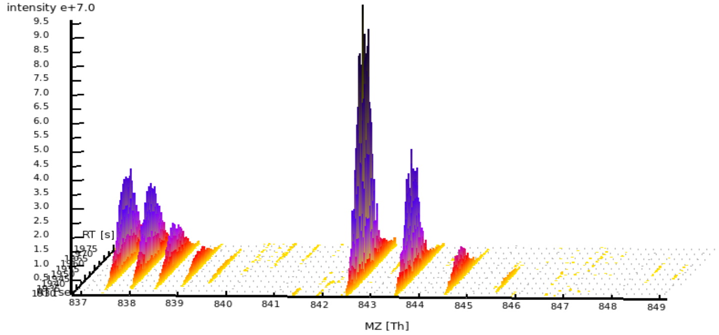
- CNN architecture referred to as U-Net [6]

CNN Training

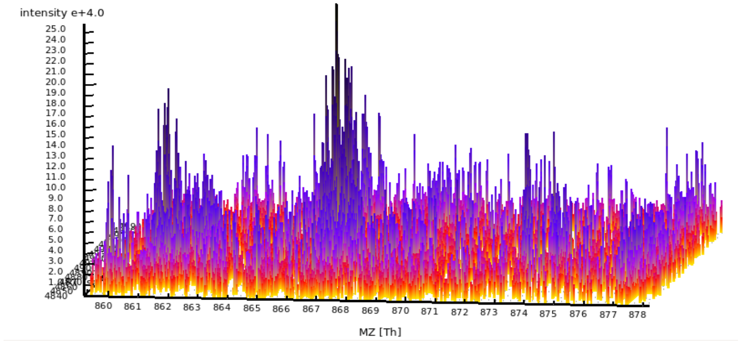
- Objective Function: Cross-entropy of the pixel-wise softmax over the final output channels
  - Softmax score gives probability of class label for each pixel
  - Cross-entropy measures similarity between predictions and ground truth distributions
- Optimizer: Stochastic Gradient Descent
  - Fixed Learning Rate: 0.01
  - Momentum: 0.9
  - Iterations: ~ 380000

Data Insight

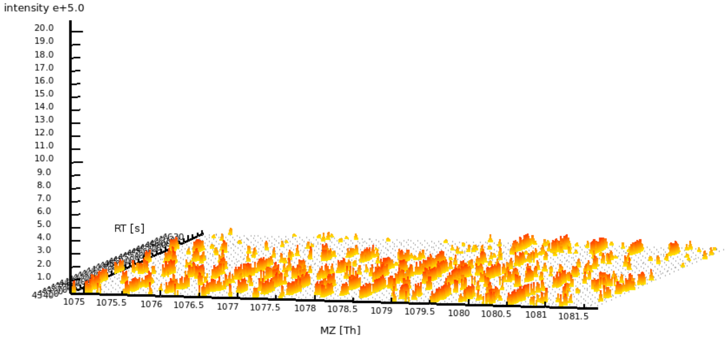
Intensities comparison



a) High Intensity Features



b) Low Intensity Features

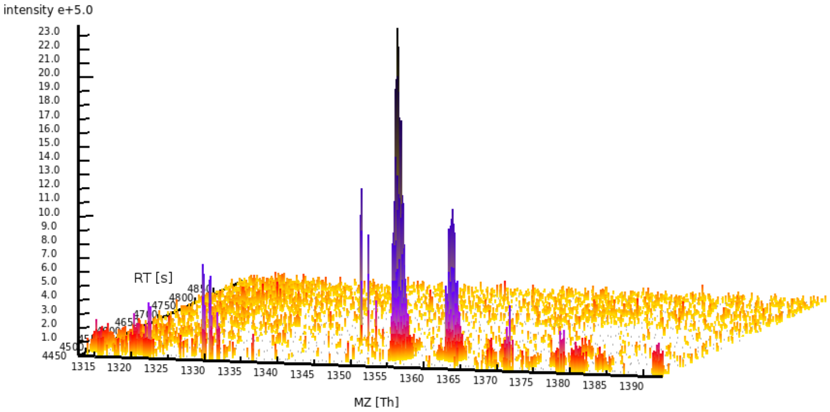


c) No feature / noise

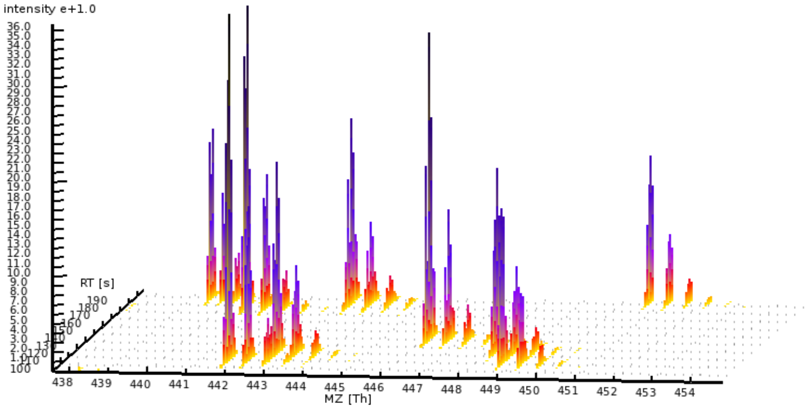
- Higher intensity features are easier to detect due to a higher S/N ratio

Data Insight

Real vs. Simulated Data



Real MS Spectra

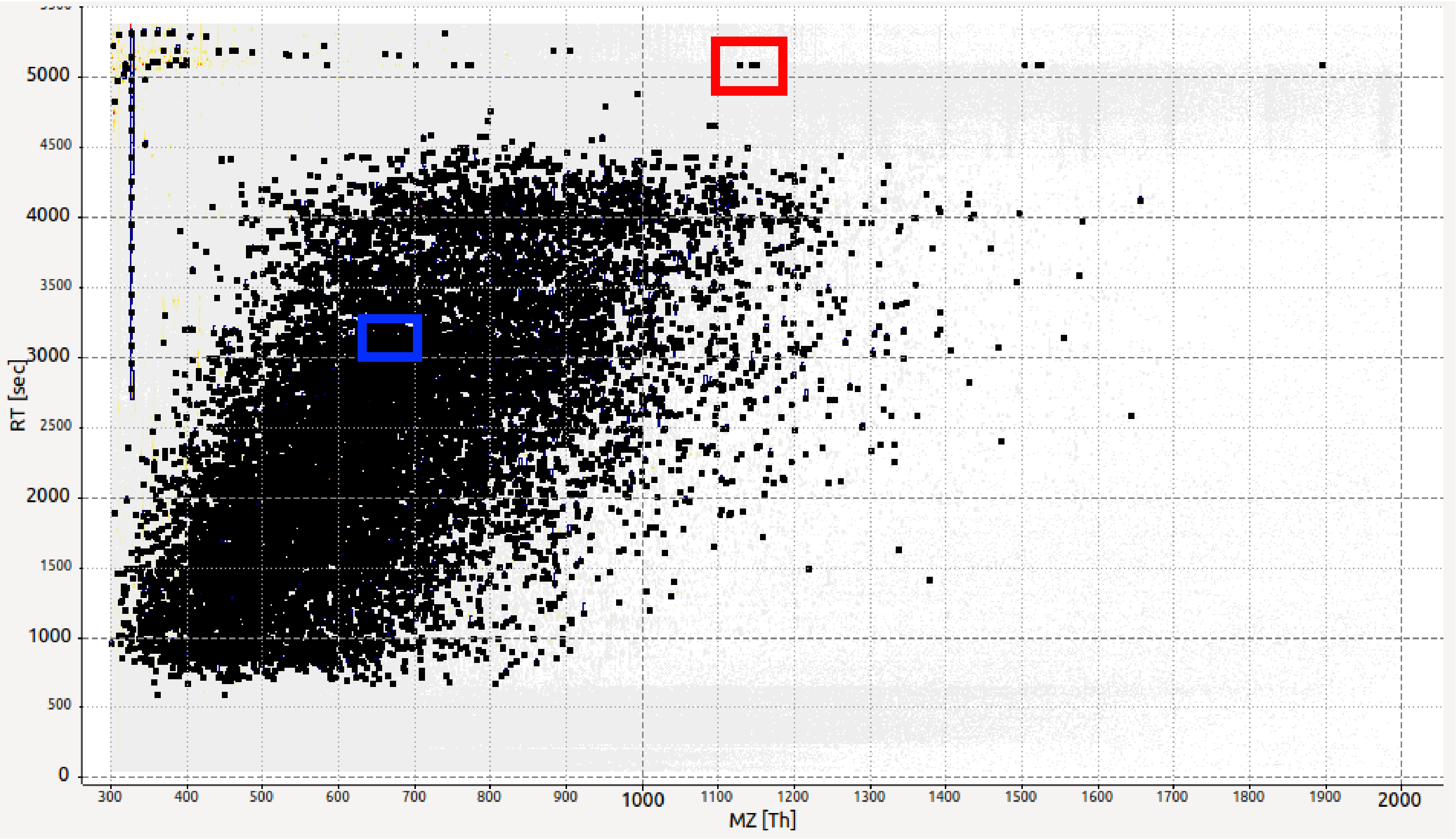


Simulated MS Spectra

- Experimental spectra contain irregular mass traces at lower intensities
- Simulated spectra does not contain any noise

Data Insight

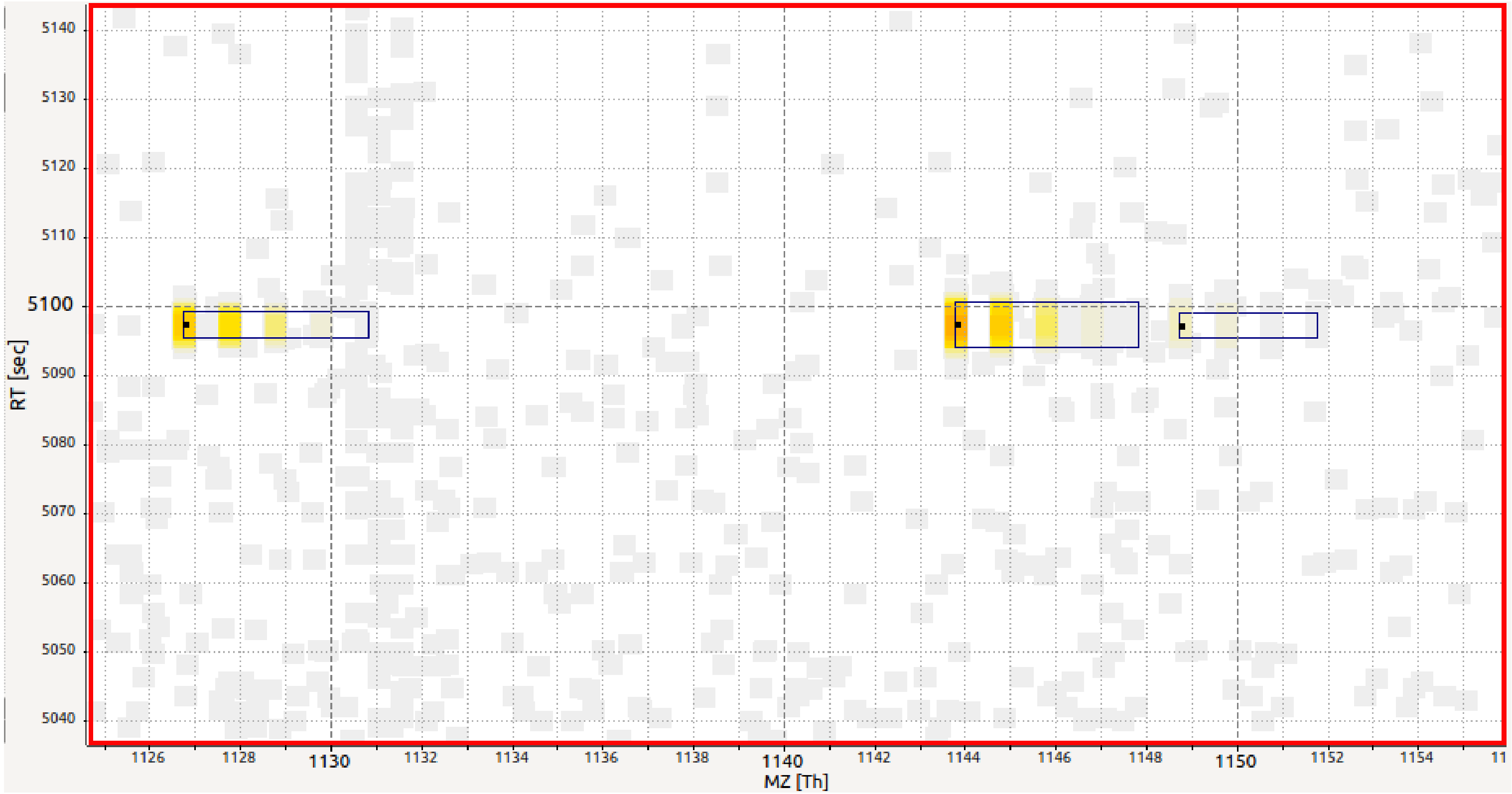
Real Data Overview



Peptides detected by FFM are marked by black squares.
Dense and sparse regions are marked with rectangles.

Data Insight

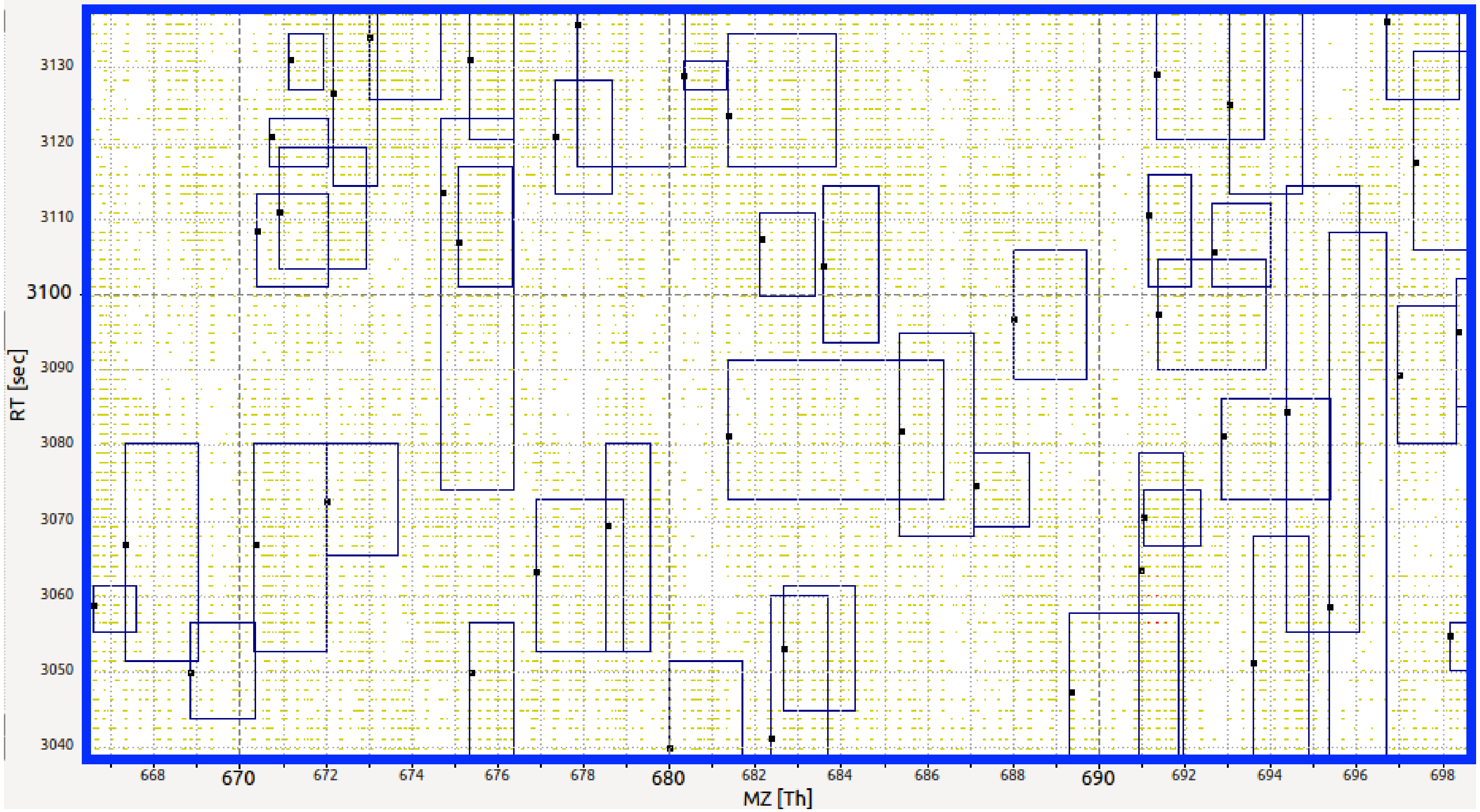
Sparse Region In Experimental Data



- Signal-to-noise ratio drops in low-intensity regions
- Shorter RT times for peptides

Data Insight

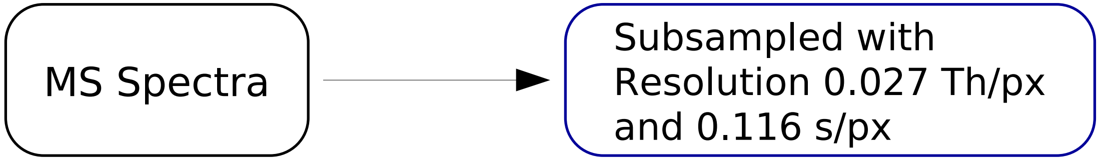
Dense Region In Experimental Data



- Dense regions have overlapping peptide features
- Peptides have longer RT times in dense regions

Data Preprocessing

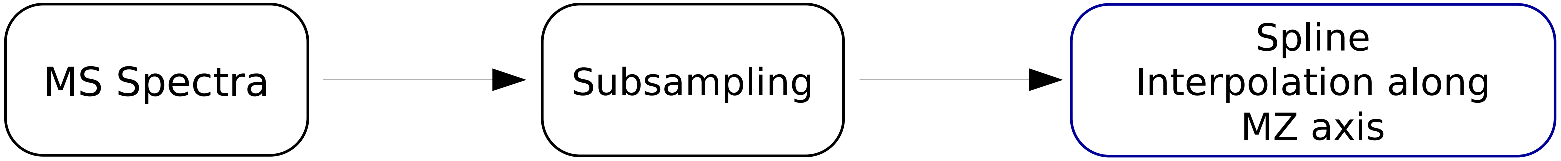
Subsampling Step



- Higher rendering resolution allows overlapping spectra remain distinguishable
- Full MS Spectra becomes a 50000x75000 pixels image
- Trade-off: Higher resolution needs more computational resources for preprocessing and network training
- CNN can not “look” at full spectra at the same time
- U-Net has a receptive field of 44 s across 10 Th

Data Preprocessing

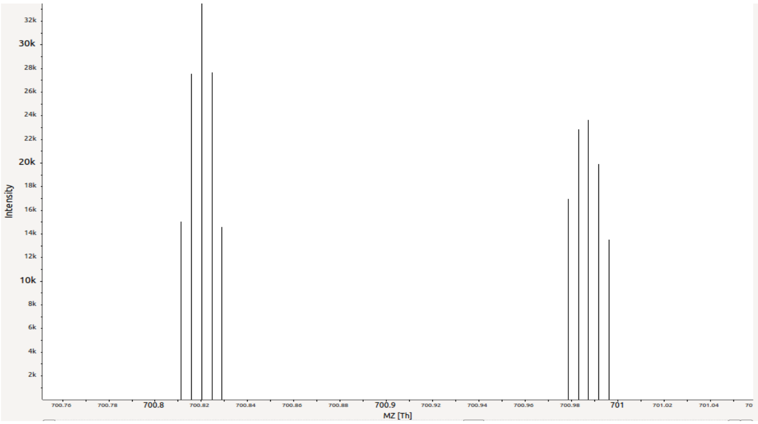
Interpolation along MZ axis

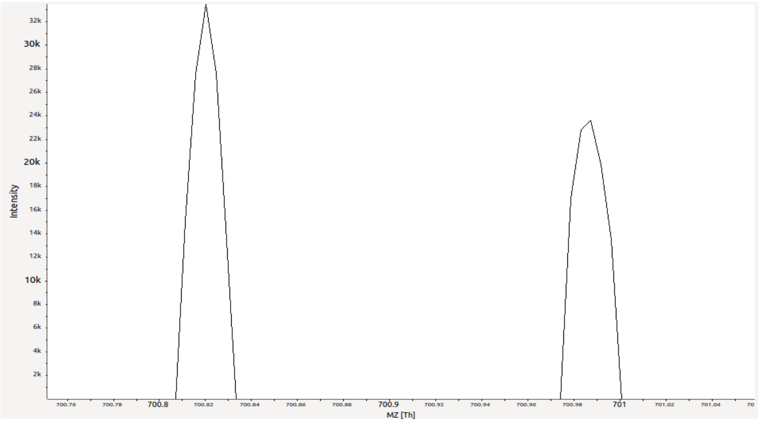


MS Spectra

Subsampling

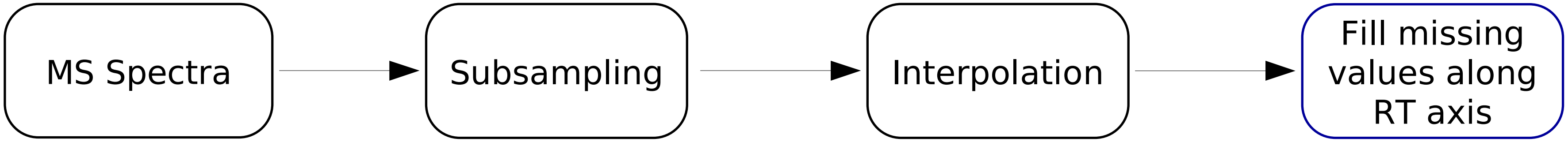
Spline Interpolation along MZ axis





Data Preprocessing

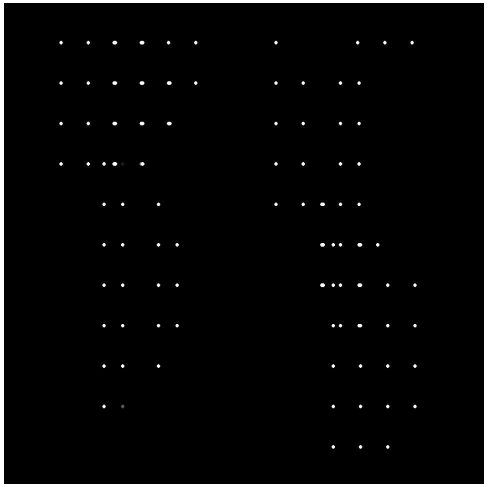
Repetition along RT axis

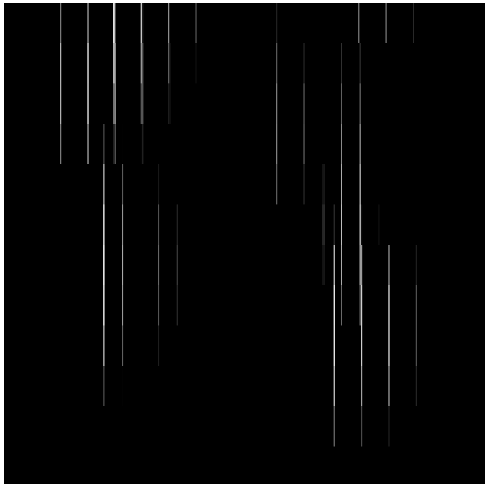

MS Spectra

Subsampling

Interpolation

Fill missing values along RT axis

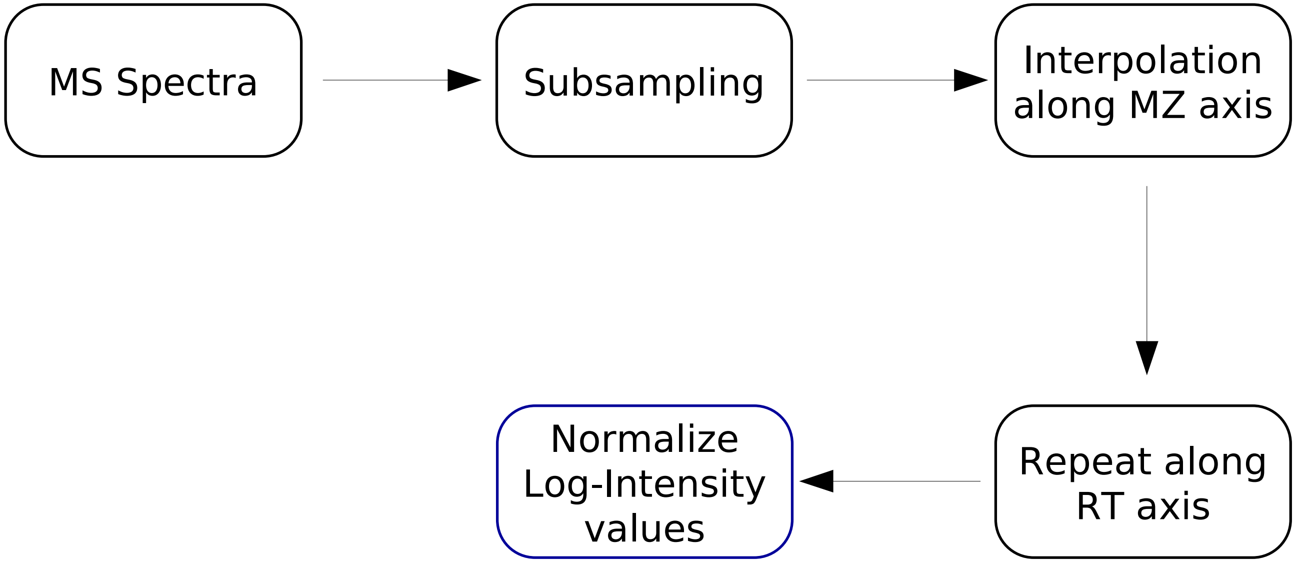

Each point represents a peak


Each line represents a peak cluster

- Intensity values are repeated along RT axis for every spectrum to fill missing values

Data Preprocessing

Normalization


MS Spectra

Subsampling

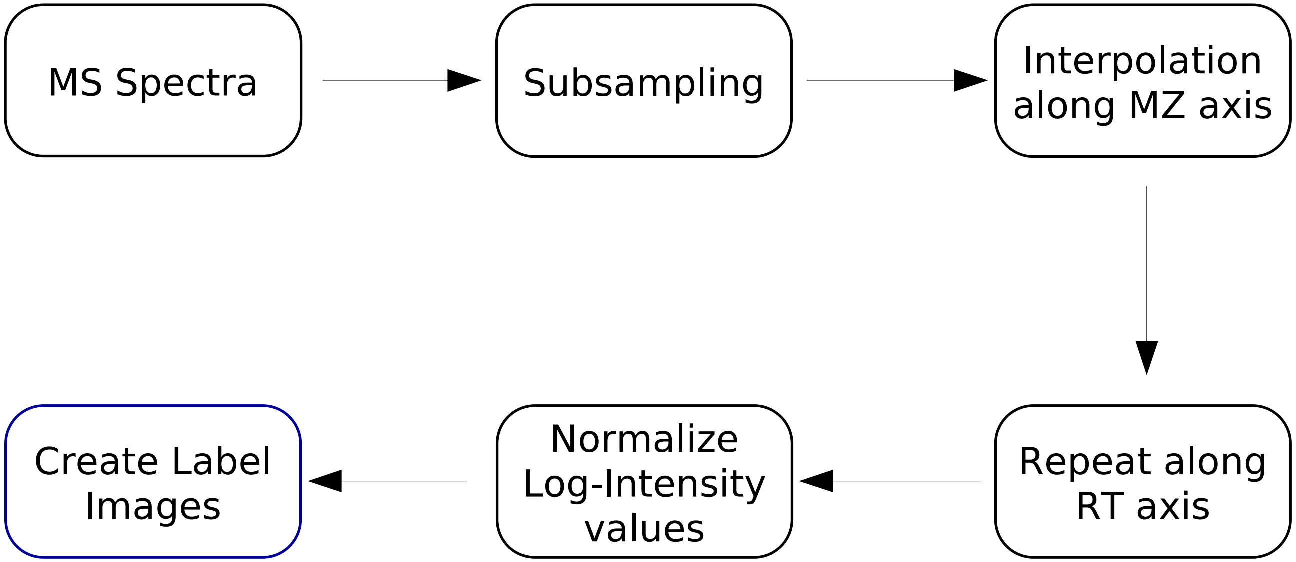
Interpolation along MZ axis

Repeat along RT axis

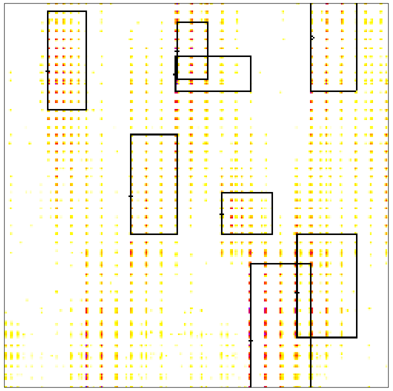
Normalize Log-Intensity values

Data Preprocessing

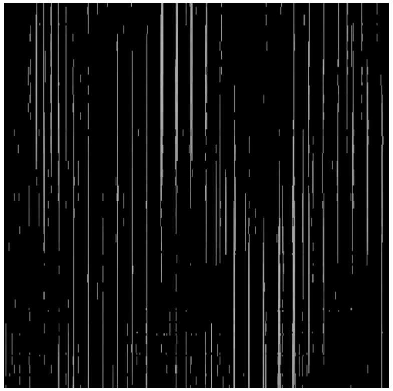
Label Images



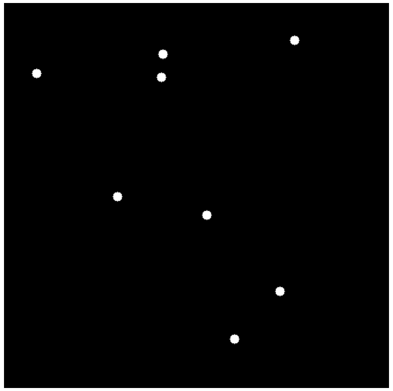
Rendered Images



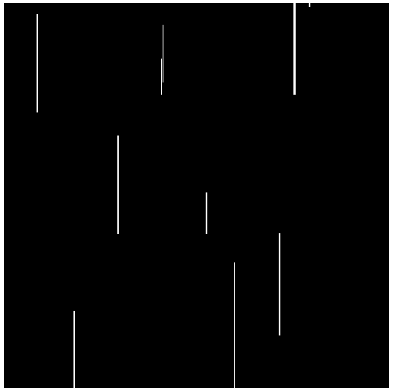
Spectra with annotations for peptides



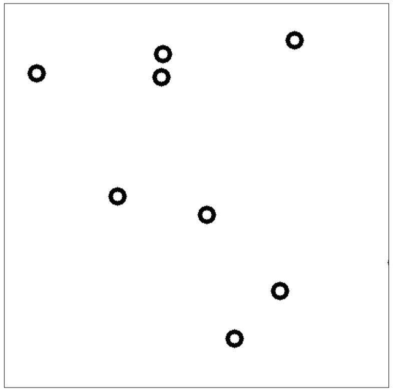
Rendered Spectra Image



Dot Labels on highest peak of 1st mass trace



Line Labels for 1st mass trace

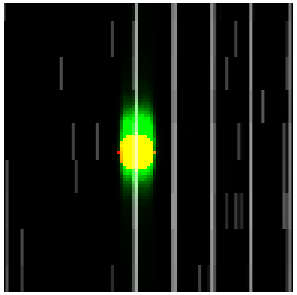


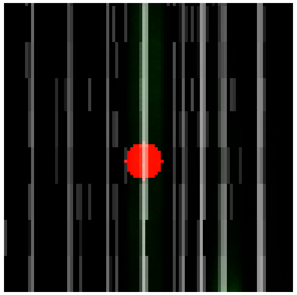
Weights for Dot Labels

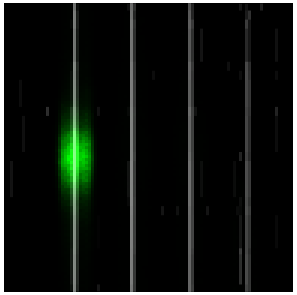
Binary Classification

- True Positive: a label correctly predicted
- False Negative: a label missed in predictions
- False Positive: a predicted label with no corresponding ground truth label

- Red: Ground truth, Green: Predicted label


True Positive


False Negative


False Positive

Red:
Ground truth

Green:
Predicted label

Evaluation Metrics

$$\text{Precision} = \frac{|\text{True Positives}|}{|\text{True Positives}| + |\text{False Positives}|}$$

$$\text{Recall} = \frac{|\text{True Positives}|}{|\text{True Positives}| + |\text{False Negatives}|}$$

- Precision measures the efficiency of neural network's predictions
- Recall measures the relevancy of the predictions

Quantitative Results

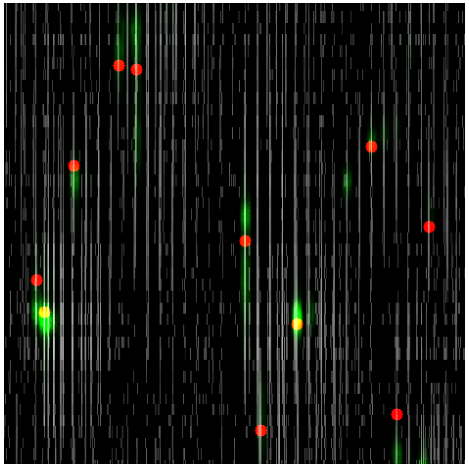
Exp	Training Data	Valid Data	Results			
			Training Set		Test Set	
			Precision	Recall	Precision	Recall
1	Sim	Sim				
2	Real + Sim	Sim				
3	Real + Sim	Real				
4	Real	Real				

Quantitative Results

Exp	Training Data	Valid Data	Results			
			Training Set		Test Set	
			Precision	Recall	Precision	Recall
1	Sim	Sim	92%	91%	92%	92%
2	Real + Sim	Sim	94%	88%	94%	89%
3	Real + Sim	Real	55%	7%	54%	4%
4	Real	Real	53%	50%	50%	10%

Qualitative Results

Dense Region Example



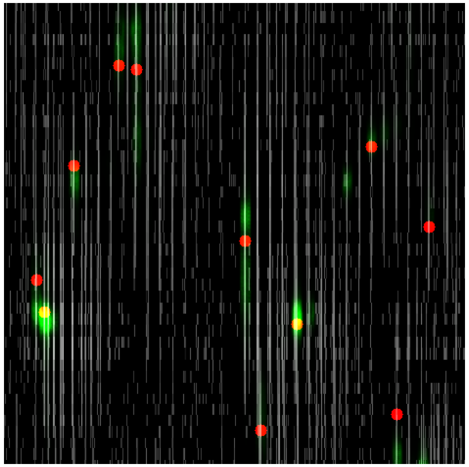
Predictions in dense region

- Red: Ground truth labels, Green: U-Net predictions
- Predictions shown here are from the model trained on real data only

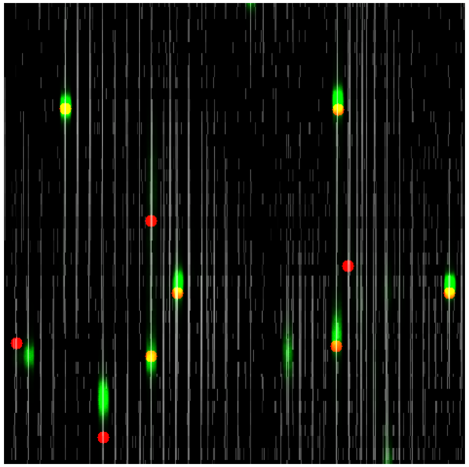
- Few true positives in dense regions, many false negatives

Qualitative Results

Medium Density Region



Predictions in dense region



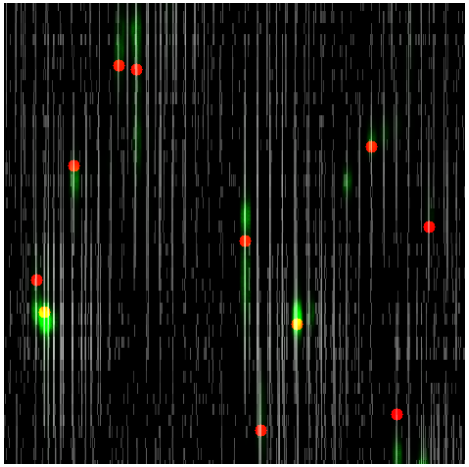
Predictions in medium density region

- Red: Ground truth labels, Green: U-Net predictions
- Predictions shown here are from the model trained on real data only

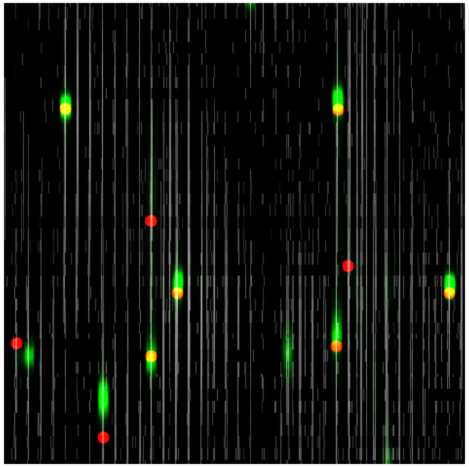
- Few true positives in dense regions, many false negatives
- More true positives and less false negatives in medium density regions

Qualitative Results

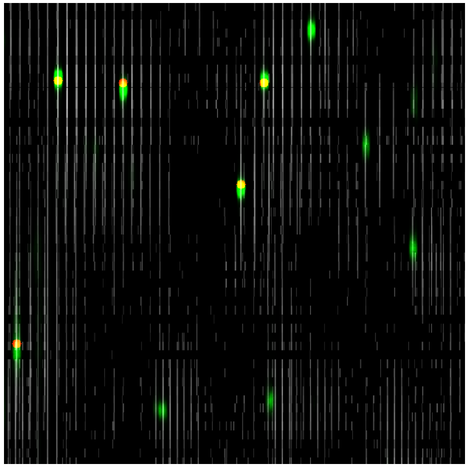
Sparse Region Example



Predictions in dense region



Predictions in medium density region

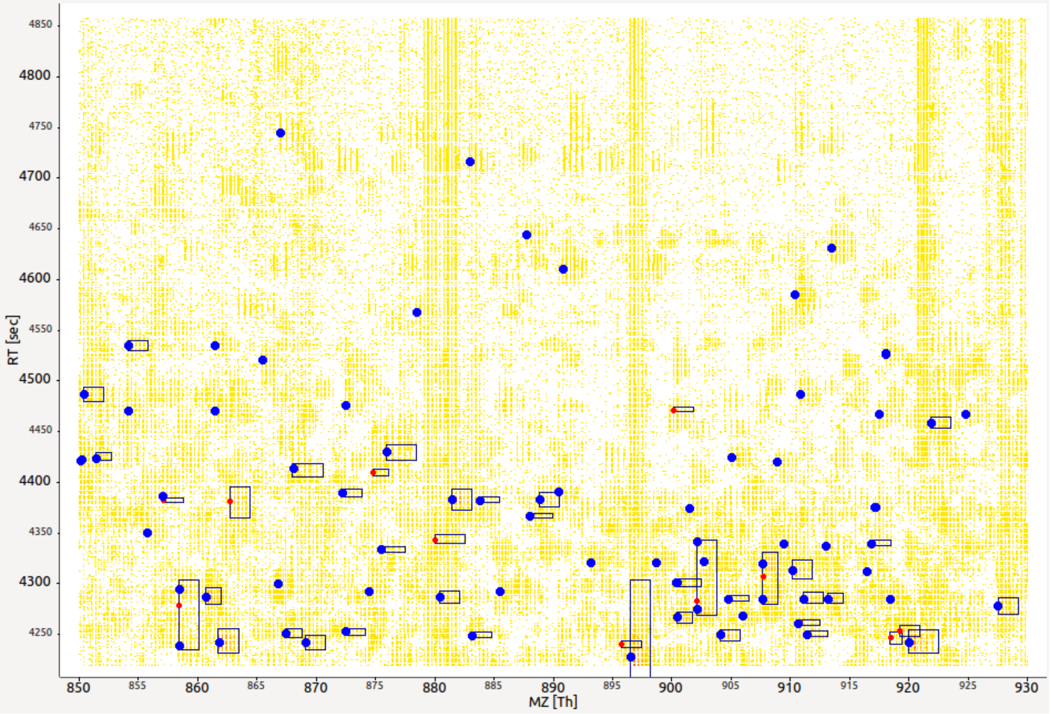


Predictions in sparse region

- Red: Ground truth labels, Green: U-Net predictions
- Predictions shown here are from the model trained on real data only

- Few true positives in dense regions, many false negatives
- More true positives and less false negatives in medium density regions
- High recall in low intensity regions along with many false positives

Sparse Region Analysis



Blue circles:
Predicted labels

Rectangles:
Ground Truth

- Mass Spectra for False Positives have to be analyzed in TOPPView
- Mass offset between two isotope peaks should be 1 Da
- Dimethyl labeled peptides appear in pairs, and CNN should detect them both
- FFM can be fine tuned for sparse region

Conclusion

- CNN did not detect overlapping peptides efficiently mainly due to receptive field limitation
- It detects peptides in low-intensity regions where FFM fails
- CNN can not outperform FFM without any human control since both data and labels had noise
- Data preprocessing is crucial and dictates network configuration as well

Future Work

- Neural Network should be optimized for bigger receptive field
- Sparse encoding and decoding based network should perform better on MS spectra
- Reformulate as Localization problem to draw bounding box around detected peptides
- Data gathered using other brands of mass spectrometers should also be utilized
- Domain adaptation techniques should be used when mixing real and synthetic data

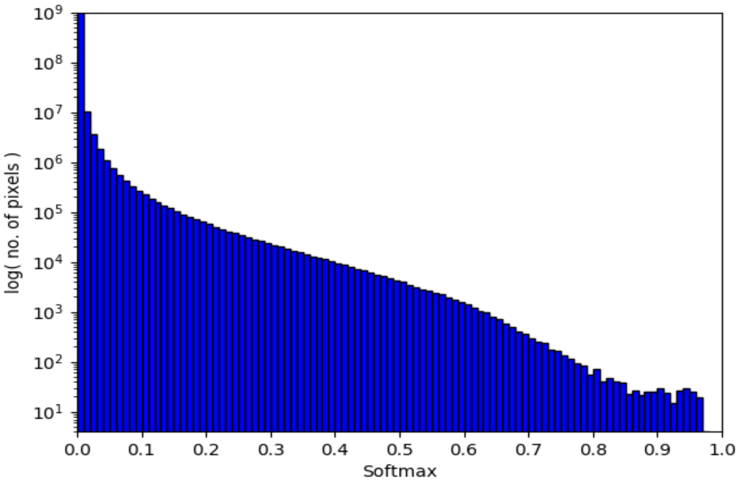
Questions

References

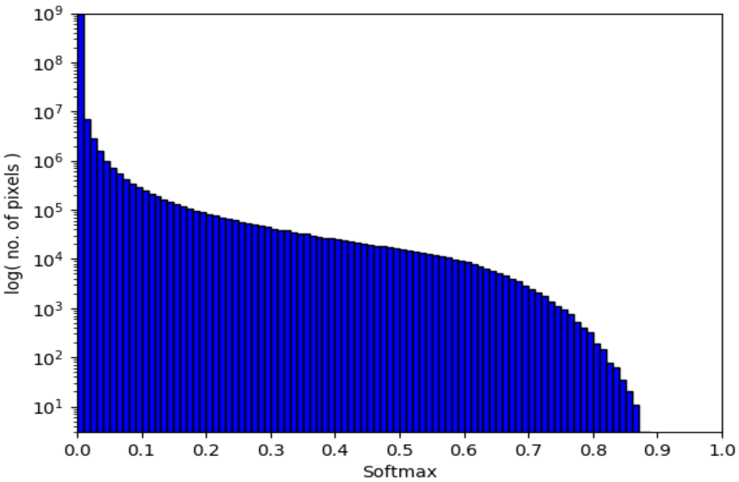
1. H. L. Röst, T. Sachsenberg, S. Aiche, C. Bielow, H. Weisser, F. Aicheler, S. Andreotti, H.-C. Ehrlich, P. Gutenbrunner, E. Kenar, et al., “Openms: a flexible open-source software platform for mass spectrometry data analysis,” *Nature methods, vol. 13, no. 9, pp. 741–748*, 2016.
2. M. Sturm and O. Kohlbacher, “Toppview: an open-source viewer for mass spectrometry data,” *Journal of proteome research, vol. 8, no. 7, pp. 3760–3763*, 2009.
3. O. Kohlbacher, K. Reinert, C. Gröpl, E. Lange, N. Pfeifer, O. Schulz-Trieglaff, and M. Sturm, “Topp—the openms proteomics pipeline,” *Bioinformatics, vol. 23, no. 2, pp. e191–e197*, 2007.
4. H. L. Röst, U. Schmitt, R. Aebersold, and L. Malmström, “pyopenms: A python-based interface to the openms mass-spectrometry algorithm library,” *Proteomics, vol. 14, no. 1, pp. 74–77*, 2014.
5. A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, “A review on deep learning techniques applied to semantic segmentation,” *arXiv preprint arXiv:1704.06857*, 2017.
6. O. Ronneberger, P. Fischer and T. Brox. (2015, October). “U-net: Convolutional networks for biomedical image segmentation”. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Springer, Cham.

Appendix - I

Softmax Score Distributions



Predictions from a model trained on real and simulated data

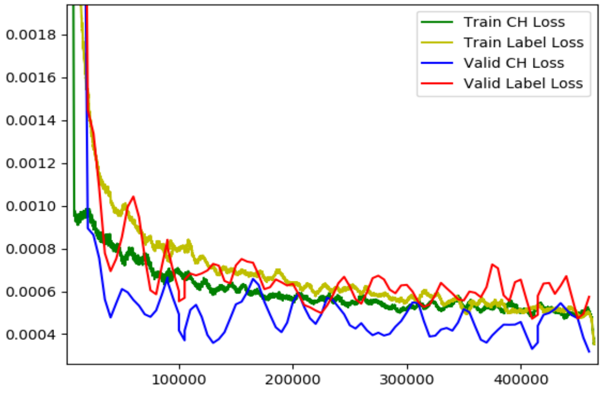


Predictions from a model trained on real data only

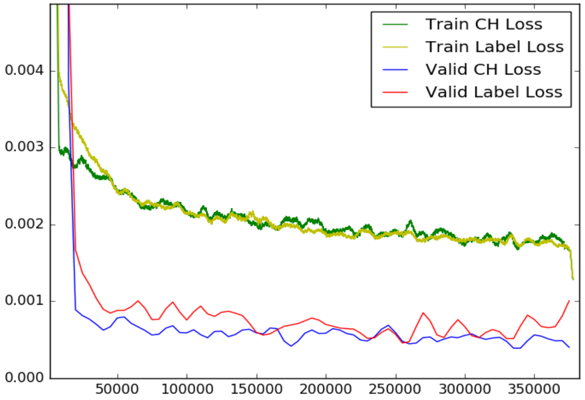
Using simulated data along with real data yields higher softmax scores in predicted labels

Appendix - II

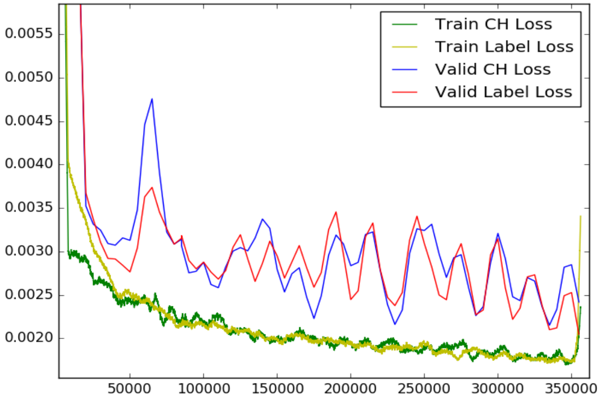
Loss Plots



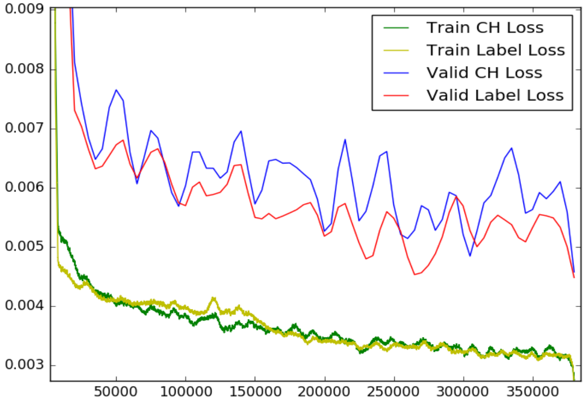
Training and validation on simulated data



Training on real and sim data, validation on simulated data



Training on real and sim data, validation on real data



Training and validation on real data

Label: Dot Label
CH: Line Label