

Deep Active Learning for De Novo Peptide Sequencing from Data-independent-acquisition Mass Spectrometry

Anonymous Authors¹

Abstract

De novo peptide sequencing from mass spectrometry data has been proved as one of the promising methods for the accurate identification of novel peptides. Recently, deep learning has been applied to de novo peptide sequencing using mass spectrometry data. Although numerous mass spectrometry dataset are publicly available, annotating a large amount of data is too expensive and time-consuming. Therefore, we need a solution for acquiring ms spectra with high quality rather than large number of them. By integrating active learning with deep learning, we can reduce the cost of annotation. In this work, we mainly focused on one of the state-of-the-art models, DeepNovo-DIA, which uses convolutional neural networks to extract features and long short-term memory to learn the language models of peptides. Instead of selecting spectra randomly to train the DeepNovo-DIA model, we utilized active learning algorithm to acquire the most informative spectra. We used the random selection as the base-line and compared it with two other acquisition methods. The experiments showed that by integrating active learning with de novo sequencing, we achieve better performance compare to DeepNovo-DIA model.

1. Introduction

Personalized cancer vaccines, as a promising cancer immunotherapy, can be developed based on identifying and validating neoantigens. Neoantigens are peptides produced from digested proteins that exist on the surface of tumor cells, so patient-specific peptides on the cancer cells can be targeted for producing personalized cancer vaccines (Lynn et al., 2020), (Qiao et al., 2020), (Sahin U, 2018). For identifying peptide sequences from tumor samples, we need a

powerful technique to enable sensitive peptide detection with low abundance. Liquid chromatography-tandem mass spectrometry (LC-MS/MS)-based proteomics is a powerful analytical tool for identifying and quantifying biological molecular such as peptide and protein (Jensen, 2006), (Beretta, 2007). Via LC-MS/MS, the enzymatically digested peptides elute from the LC column one by one and the mass spectrometer records mass spectra over time. The mass spectrometer records the mass-to-charge ratio of the charged peptides, termed MS1 spectrum. Then it selects peptides for fragmentation by using different approaches. The mass spectra of the charged fragments MS2 are recorded at the final step. Different approaches have been proposed for selecting peptides to be fragmented to MS2. One of these approaches is data-dependent acquisition (DDA) (Tran et al., 2017) that uses a narrow precursor m/z windows which contains a single peptide for each MS2 spectrum. In contrast, data-independent acquisition (DIA) partitions the entire m/z range of the MS1 spectrum into wide intervals and considers all the intervals. The goal of the data-independent acquisition (DIA) approach is to analyze all peptides in the sample. After acquiring MS1 and MS2 spectra containing sets of m/z and intensity via LC-MS/MS experiment, proper algorithms are required to interpret data to meaningful information. Basically, there are two main approaches for translating spectra to peptide composed of amino acid sequences: i) database search engine that uses databases containing known sequences, ii) de novo sequencing that decodes data from scratch. De novo sequence methods try to reconstruct the amino acid sequence a peptide is composed of from scratch without searching in a database and any prior knowledge of the amino acid sequence. Applying de novo sequencing methods on DIA mass spectrometry data for analyzing masses and inferring peptides is a challenging task due to the complex mixture of spectra containing multiple coeluting peptides. To interpret high multiplex spectra, we need computational methods which consider all possible combination of peptide candidates for a given spectrum. Deep learning models are good candidates for developing de novo sequencing to learn coeluting patterns. Neural networks can extract features from highly multiplex and noisy DIA spectra and learn the language model of peptides (Tran et al., 2018). To improve the DeepNovo-DIA as

¹University of North Texas, Denton, Texas, USA. Correspondence to: Anonymous Author <anon.email@domain.com>.

a de novo sequencing approach based on neural networks, we utilize active learning (AL) algorithm to acquire the most informative spectrum instead of random selection. In this paper, we combine active learning algorithm with de novo peptide sequencing enabled by deep learning model. The experiments show that by selecting less than half spectrum, we can get close performance to DeepNovo-DIA model.

2. Related Works

Due to the importance of de novo peptide sequencing in proteomics, numerous computational methods have been proposed (Hao Chi & Dong, 2013), (Yan Y, 2014), (Yan Y, 2017), and (Tschager, 2018). Deep neural networks to de novo peptide sequencing proposed by (Tran et al., 2017) as DeepNovo that outperforms the other proposed de novo sequencing algorithms without searching databases. DeepNovo achieves this by integrating Convolutional Neural Networks (CNNs) and Long short-term memory (LSTM) to learn features of MS/MS, fragments, and language model of peptides. In 2019, DeepNovo was upgraded to identify peptides from DIA MS data (Tran et al., 2018). Again CNNs used to map precursor and fragment ion profiles to embedding vector as encoder, and LSTM used as decoder for predicting next amino acids. DeepNovo-DIA achieved better performance compared to the other database search methods (Bruderer R, 2015), and (Ting, 2017). On the other hand, integrating deep learning with active learning enables processing high dimensional data plus automatic feature extraction with selecting data points more efficiently. The goal of this combination is to choose good data instead of big data to reduce the annotation cost, especially in health and biology domain. Deep active learning has been used in wide various domains and applications. Applications in NLP such as question answering (Nabiha Asghar & Li, 2017), information extraction (Jungo Kasai & Popa, 2019), (Maldonado & Harabagiu, 2019), and (Matthew Shardlow & Ananiadou, 2019), semantic analysis (S. Das Bhattacharjee & Balantrapu, 2017), text classification (Bang An & Han, 2018), (Ameya Prabhu & Singh, 2019), machine translation (Pei Zhang & Xiong, 2018), wearable device (SGautham Krishna Gudur & Umaashankar, 2019), and (Hossain & Roy, 2019), gene expression (Rania Ibrahim & El-Makky, 2014), in Electrocardiogram (ECG) signal processing (Hanbay), in computer vision such as image classification, object detection, video processing, semantic segmentation (Baolin Du & Ding, 2018), (Xuhui Chen & Li, 2018), (Samuel Budd & Kainz, 2021), (Cheng Deng & Tao, 2019), (Gal & Ghahramani, 2015), (Gal & Ghahramani, 2016), and (Yarin Gal & Ghahramani, 2017).

3. Approach

In this work, We develop an active learning framework for extending DeepNovo-DIA proposed by (Tran et al., 2018). Through active learning framework, the model could learn from small amount of spectrum. Unlike DeepNovo-DIA which selects the spectrum randomly, we utilize different acquisition function to select the most informative spectrum. In the following sections, 3.1 we used three datasets previously used by DeepNovo-DIA, 3.2 we demonstrate how we integrate active learning framework with de novo peptide sequencing model. Finally, 4 we explain our experiment results.

3.1. Dataset

We use the same DIA mass spectrometry dataset obtained and used previously (Tran et al., 2018) to train and test our model. The dataset used for training includes urine samples from different subjects (Muntel). We used ovarian cyst (OC; six subjects) as validation set, and evaluate on previously used dataset of plasma sample (Ting, 2017). Before feeding the data to the neural network, the first step is to process the dataset to extract required features. Each feature is about a precursor including its mass-to-charge (m/z), charge, retention-time, intensity profile obtained from LC-MS map (Jing Zhang 1, 2012). The pairs of (m/z , intensity) are collected from mgf file based in the center of retention-time and scan number for each precursor. The closer spectra to the center of precursor's retention time are selected because their fragment ion signals are stronger to do the de novo sequencing. After extracting the pair of signals for each precursor, we construct MS2 fragment ions. Now, the prepared dataset including precursor's features with its collected spectra are ready to lunch to DeepNovo-DIA model.

3.2. Methods

In this section, we illustrate the integration of active learning framework with de novo peptide sequencing. To integrate active learning with DeepNovo-DIA, we use pool-based methods to acquire the spectra contains more information rather than random selection. We utilize the uncertainty strategies for measuring informativeness of spectra. In the following sections, first we demonstrate DeepNovo-DIA model, then explain active learning algorithm, and finally propose the combined framework of active learning with DeepNovo-DIA.

3.2.1. DEEPNOVO-DIA

Figure 1 presents the architecture of DeepNovo-DIA which enables de novo sequencing using neural networks (Tran et al., 2018). DeepNovo-DIA model consists of encoder and

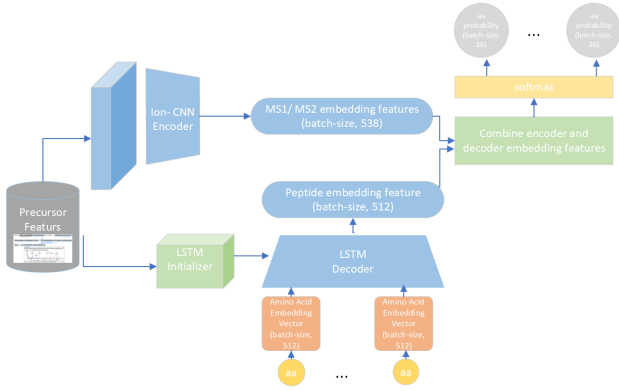


Figure 1. DeepNovo-DIA model. Precursor ion features detected by LC-MS map include m/z, charge, ms1, retention time. MS1 and MS2 collected spectra then feed to the deep learning model. Ion-CNN learns the and encodes the MS1 and MS2 into an embedding vector. The amino acid chains extracted by LC-MS are feed to LSTM model to learn amino acid language. Spectrum-CNN consisting two convolution and max pooling layers used to initialized LSTM.

decoder to learn both features from MS1 and MS2 spectra and peptide languages. It has three main modules, Ion-CNN for encoding MS1 and MS2 high dimensional input to a feature vector, CNN spectrum with LSTM for decoding peptide language and a module for combining the outputs of encoder and decoder. Since DeepNovo-DIA is proposed for interpreting highly multiplexed DIA MS/MS spectrum where fragment ions emerged from multiple peptides, for each precursor, 5 spectrum vectors are collected. Each spectrum vector is an intensity vector, where the index of each entity is m/z. This vector is feed to CNN spectrum to be encoded for initializing the LSTM decoder. To improve peptide accuracy, Focal loss is used instead of cross entropy.

3.2.2. ACTIVE LEARNING

Active learning framework as a solution for reducing the cost of annotating data can be considered as a smarter way of selecting informative data points for annotation. For developing active learning, one of the most of important parts is to define a measurement for measuring the informativeness of data points. Consequently, active learner aims to achieve precise accuracy using as few annotated data points as possible. From three scenarios that active learner queries unlabeled data points including (i) membership query synthesis (ii) stream-based selective sampling, and (iii) pool-based sampling, we choose pool-based sampling because of advantages over others (Settles, 2009). The informativeness of data points can be measured by calculating their uncertainty. More uncertainty about a data point is equal to be more informative. Thus, it is more efficient to train a model

Algorithm 1 Active Learning for De novo peptide sequencing

Input: pool set of unlabeled MS/MS DIA data
 $ps = \text{poolset}$, $t_{\text{denovo}} = \text{unlabeledtestset}$, $vs = \text{validationset}$, $ts = \text{trainset}$,
 $k = 50$, $q = 0$, $max - q = 200$
Initialization:
 $ts = \text{RandomSpectraSelection}(ps, k)$
 $\text{TrainDeepNovoDIA}(ts, vs)$
 $ps = ps - ts$
repeat
 for acquisitionF **to** AcquisitionFList **do**
 $probs - ps = \text{TestDeepNovoDIA}(ps)$
 $uncertain - spectra = \text{Select}(\text{acquisitionF}, probs - ps, k)$
 $ts = ts + uncertain - spectra$
 $ps = ps - uncertain - spectra$
 $q = q + k$
 $\text{TrainDeepNovoDIA}(ts, vs)$
 $\text{DeNovoPepSeq}(t_{\text{denovo}})$
 TestDeNovoPepSeq
 end for
until $q < max - q$

with most uncertain data points because they contain more information. There are different methods for calculating uncertainty, we choose two of the most popular with fewer drawbacks: *Margin Sampling* chooses the data points from the pool set with the lowest margin between the first and second most probable labels under the model (cheffer & Wrobel, 2011):

$$X_M = \underset{x}{\operatorname{argmin}} P_{\theta}(y_1^*|x) - P_{\theta}(y_2^*|x)$$

y_1^* is the most probable and y_2^* is the second most probable labels under the model θ . Through the margin selection, we calculate the smallest difference between first and second most probable labels. The intuition is that the larger the difference, the more confident the model is for the predicted labels. So, for selecting the most uncertain labels, we need the least difference or margin. Another method is *Maximum Entropy* known as the most popular strategy for measuring uncertainty, chooses the points from a given pool set with the maximum entropy (ZHAO, 2017), (Dagan & Engelson, 1995):

$$X_E = \underset{x}{\operatorname{argmax}} - \sum_y P_{\theta}(y_i|x) \log P_{\theta}(y_i|x)$$

$P_{\theta}(y_i|x)$ is the probability that point x belongs to a class of y_i under the model θ , where y_i is the ranges of all possible classes predicted by model θ

3.2.3. AL FOR DEEP DE NOVO PEPTIDE SEQUENCING

To integrate AL with DeepNovo-DIA as a state-of-the-art de novo peptide sequencing model using neural network, we develop the algorithm 1. The DeepNovo-DIA is initialized with a set pool of unlabeled MS/MS DIA data. To train the model, the algorithm selects k random spectra from the pool set containing MS/MS DIA mass spectrometry, and queries their labels. After training the model with small amount of data ts , we measure the performance of model on unseen/unlabeled from pool set ps and generates a probability vector $probs - ps$ used as the input of acquisition function. Acquisition functions is the core of active learning algorithm, in this work we select the most uncertain spectra based on two mentioned strategies in the 2 and use random selection as a baseline. In this work, we used margin sampling and maximum entropy to acquire the most uncertain spectra. After selecting the most uncertain spectra, the algorithm moves them from pool set to train set, then train the model with new train set again. The AL repeats this process until it meets the stopping criteria. Here we set the max number of query as stopping criteria which is equal to 200. Figure 2 illustrates the framework of combining active learning with DeepNovo-DIA model.

4. Results

To evaluate the performance of the proposed framework, we used accuracy as the same metric defined in DeepNovo-DIA model. We compared the accuracy of two main acquisition functions with random selection used by DeepNovo-DIA. The amino acid accuracy is defined as the ratio of the number of amino acids matched to the total amino acid number in the real peptide sequence. The definition for peptide level is the number of fully correctly predicted to the real ground truth peptide. Figure 3 and Figure 4 illustrate the average accuracy of each acquisition strategy used to choose the labeled spectrum. Figure 3 and 4 show the performance AL-DeepNovo-DIA framework using margin selection, maximum entropy, and random selection at amino acid level and peptide level respectively. Margin selection outperforms the maximum entropy and random selection. The experiment shows that for both peptide and amino acid level, even training the model with a small amount of spectra acquired by margin selection and maximum entropy, we can achieve the same performance compare to the gradually increased training data. Compare to the random selection that the accuracy increases significantly by increasing the number of added spectra, two other selection functions can achieve high performance with a small amount of spectra.

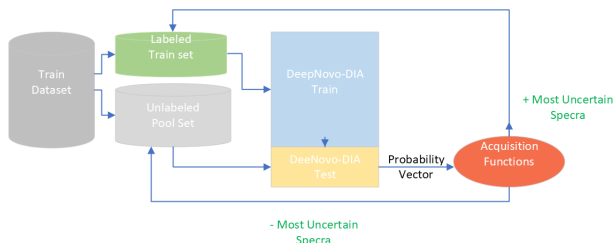


Figure 2. Active Learning framework for DeepNovo-DIA.

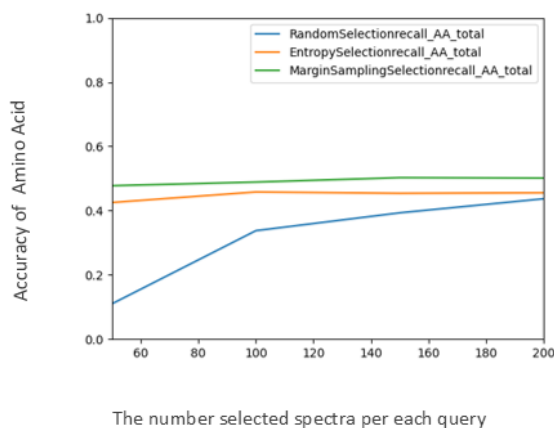


Figure 3. Performance comparison for three different acquisition functions of the proposed AL-DeepNovo-DIA framework at amino acid level.

5. Conclusion

In this paper, we proposed the deep learning model based on active learning to de novo sequence the peptides using less annotated mass spectrometry data. Instead of training the model on large, expensive annotated data, the proposed framework is able to select the data points which are more informative measured by uncertainty. The experimental results demonstrate that the proposed active learning framework can enable de novo sequencing model measuring the uncertainty of spectra and produce comparable results without trained on the original large dataset.

References

Ameya Prabhu, C. D. and Singh, M. Sampling bias in deep active classification: An empirical study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-*

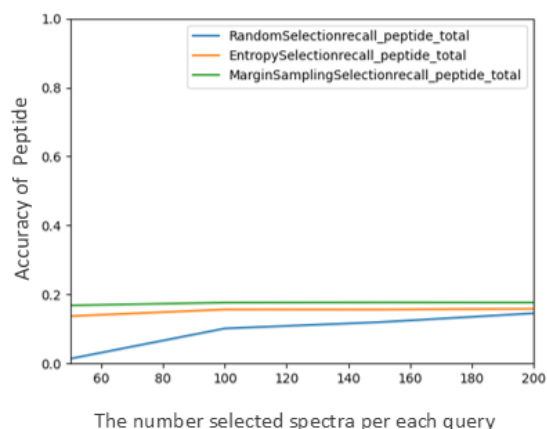


Figure 4. Performance comparison for three different acquisition functions of the proposed AL-DeepNovo-DIA framework at peptide level

IJCNLP 2019, pp. 4056–4066, Hong Kong, China, 2019. Association for Computational Linguistics.

Bang An, W. W. and Han, H. Deep active learning for text classification. In *Proceedings of the 2nd International Conference on Vision, Image and Signal Processing, ICVISIP 2018*, Las Vegas, NV, USA, 2018. ACM.

Baolin Du, Qi Qi, H. Z. Y. H. and Ding, X. Breast cancer histopathological image classification via deep active learning and confidence boosting. In *27th International Conference on Artificial Neural Networks*, pp. 109–116, Rhodes, Greece, 2018. Springer.

Beretta, L. Proteomics from the clinical perspective: many hopes and much debate. *Nature Methods*, (4):785–786, 2007.

Bruderer R, Bernhardt OM, G. T. M. S. C. L. M. S. E. T. Z. V. B. Y. E. C. V. O.-R. O. R. L. Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol Cell Proteomics*, (14), 2015.

cheffer, C. D. and Wrobel, S. Active hidden markov models for information extraction. In *Proceedings of the International Conference on Advances in Intelligent Data Analysis (CAIDA)*, 2011.

Cheng Deng, Yumeng Xue, X. L. C. L. and Tao, D. Active transfer learning network: A unified deep joint spectral-spatial feature learning model for hyperspectral im-

age classification. *IEEE Transactions on Geoscience and Remote Sensing*, 57, 2019.

Dagan and Engelson, S. Committee-based sampling for training probabilistic classifiers. In *Proceedings of the International Conference on Machine Learning (ICML)*, 1995.

Gal, Y. and Ghahramani, Z. Bayesian convolutional neural networks with bernoulli approximate variational inference. *ArXiv*, abs/1506.02158, 2015.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning, ICML*, pp. 1050–1059, New York City, NY, USA, 2016. JMLR.org.

Hanbay, K. Deep neural network based approach for ecg classification using hybrid differential features and active learning. *IET Signal Processing*, (13):165–175.

Hao Chi, Haifeng Chen, K. H. L. W. B. Y. R.-X. S. J. L. W.-F. Z. C.-Q. S. S.-M. H. and Dong, M.-Q. pnovov+: De novo peptide sequencing using complementary hcd and etd tandem mass spectra. *Proteome Research*, (12): 615–625, 2013.

Hossain, H. M. S. and Roy, N. Active deep learning for activity recognition with context aware annotator selection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining, KDD*, Anchorage, AK, USA, 2019.

Jensen, O. Interpreting the protein language using proteomics. *Nat Rev Mol Cell Biol*, pp. 391–403, 2006.

Jing Zhang 1, Lei Xin, B. S. W. C. M. X.-D. Y. W. Z. Z. G. A. L. B. M. Open access peaks db: De novo sequencing assisted database search for sensitive and accurate peptide identification. *Molecular Cellular Proteomics*, 2012.

Jungo Kasai, Kun Qian, S. G. Y. L. and Popa, L. Low-resource deep entity resolution with transfer and active learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5851–5861, pp. 5851—5861, Florence, Italy, 2019. Association for Computational Linguistics.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lynn, G., Sedlik, C., Baharom, F., and et al. Peptide-tlr-7/8a conjugate vaccines chemically programmed for nanoparticle self-assembly enhance cd8 t-cell immunity to tumor antigens. *Nat Biotechnol*, (38):320–332, 2020.

- Maldonado, R. and Harabagiu, S. M. Active deep learning for the identification of concepts and relations in electroencephalography reports. *Journal of Biomedical Informatics*, (98), 2019.
- Matthew Shardlow, Meizhi Ju, M. L. C. O. E. I. J. M. and Ananiadou, S. A text mining pipeline using active and deep learning aimed at curating information in computational neuroscience. *Neuroinformatics*, (17):391–406, 2019.
- Muntel, J. e. a. Advancing urinary protein biomarker discovery by data-independent acquisition on a quadrupole-orbitrap mass spectrometer. *J. Proteome Res*, 14.
- Nabiha Asghar, Pascal Poupart, X. J. and Li, H. Deep active learning for dialogue generation. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics, *SEM @ACM 2017*, Vancouver, Canada, 2017. Association for Computational Linguistics.
- Pei Zhang, X. X. and Xiong, D. Active learning for neural machine translation. In *2018 International Conference on Asian Language Processing, IALP*, pp. 153–158, 2018.
- Qiao, R., Tran, N. H., Shan, B., Ghodsi, A., and Li, M. Personalized workflow to identify optimal t-cell epitopes for peptide-based vaccines against covid-19. *arXiv: Populations and Evolution*, 2020.
- Rania Ibrahim, Noha A Yousri, M. A. I. and El-Makky, N. M. Multi-level gene/mirna feature selection using deep belief nets and active learning. In *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 3957–3960. IEEE, 2014.
- S. Das Bhattacharjee, A. T. and Balantrapu, B. V. Active learning based news veracity detection with feature weighting and deep-shallow fusion. In *2017 IEEE International Conference on Big Data (Big Data)*, pp. 556–565. IEEE, 2017.
- Sahin U, T. Personalized vaccines for cancer immunotherapy. *Science*, (359(6382)):1355–1360, 2018.
- Samuel Budd, E. C. R. and Kainz, B. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, 71, 2021.
- Settles, B. Active learning literature survey. Technical report, Department of Computer Sciences, University of Wisconsin-Madison, 2009.
- SGautham Krishna Gudur, P. S. and Umaashankar, V. Activeharnet: Towards on-device deep bayesian active learning for human activity recognition. In *The 3rd International Workshop on Deep Learning for Mobile Systems and Applications (EMDL’19)*. ACM, 2019.
- Ting, Y. S. e. a. Pecan: library-free peptide detection for data-independent acquisition tandem mass spectrometry data. *Nat. Methods*, (14):903–908, 2017.
- Tran, N. H., Zhang, X., Xin, L., Shan, B., and Li, M. De novo peptide sequencing by deep learning. *Proceedings of the National Academy of Sciences*, 114:8247 – 8252, 2017.
- Tran, N. H., Qiao, R., Xin, L., Chen, X., Liu, C., Zhang, X., Shan, B., Ghodsi, A., and Li, M. Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nature Methods*, 16:63–66, 2018.
- Tschager, T. *Algorithms for Peptide Identification via Tandem Mass Spectrometry*. PhD thesis, Magazin Höggerberg Diss ETH 24870, 2018.
- Xuhui Chen, Jinlong Ji, T. J. and Li, P. Cost-sensitive deep active learning for epileptic seizure detection. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB*, pp. 226–235, Washington, DC, 2018. ACM.
- Yan Y, Kusalik AJ, W. F. Novohcd: de novo peptide sequencing from hcd spectra. *IEEE Trans Nanobioscience*, (13):65–72, 2014.
- Yan Y, Kusalik AJ, W. F. Novoexd: De novo peptide sequencing for etd/ecd spectra. *IEEE/ACM Trans Comput Biol Bioinform*, (14):337–344, 2017.
- Yarin Gal, R. I. and Ghahramani, Z. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning, ICML*, pp. 1183–1192, Sydney, NSW, Australia, 2017. Proceedings of Machine Learning Research, Vol. 70. PMLR.
- ZHAO, W. *Deep Active Learning for Short-Text Classification*. PhD thesis, COMPUTER SCIENCE AND ENGINEERING, KTH Royal Institute of Technology, 2017.