Machine learning-based peptide-spectrum match rescoring opens up the immunopeptidome

Charlotte Adams[1,2], Kris Laukens[1], Wout Bittremieux[1], Kurt Boonen[2,3,*]

1 Department of Computer Science, University of Antwerp, Antwerp, Belgium

2 Centre for Proteomics (CFP), University of Antwerp, Antwerp, Belgium

3 ImmuneSpec BV, Niel, Belgium

Abbreviations

APC: Antigen Presenting Cells

CCS: Collisional Cross Section

DL: Deep Learning

HLA: Human Leukocyte Antigen

LC: Liquid Chromatography

MHC: Major Histocompatibility Complex

ML: Machine Learning

MS: Mass Spectrometry

nuORFs: unannotated open reading frames

PSM: Peptide Spectrum Matches

PTM: Post-Translational Modification

SA: Normalized Spectral Contrast Angle

SVM: Support Vector Machine

Keywords

Data analysis, immunopeptidomics, machine learning, mass spectrometry

Abstract

Immunopeptidomics is a key technology in the discovery of targets for immunotherapy and vaccine development. However, identifying immunopeptides remains challenging due to their non-tryptic nature, which results in distinct spectral characteristics. Moreover, the absence of strict digestion rules leads to extensive search spaces, further amplified by the incorporation of somatic mutations, pathogen genomes, unannotated open reading frames, and post-translational modifications. This inflation in search space leads to an increase in random high-scoring matches, resulting in fewer identifications at a given false discovery rate. Peptide-spectrum match rescoring has emerged as a machine learning-based solution to address challenges in mass spectrometry-based immunopeptidomics data analysis. It involves post-processing unfiltered spectrum annotations to better distinguish between correct and incorrect peptide-spectrum matches. Recently, features based on predicted peptidoform properties, including fragment ion intensities, retention time, and collisional cross section, have been used to improve the accuracy and sensitivity of immunopeptide identification. In this review, we describe the diverse bioinformatics pipelines that are currently available for peptide-spectrum match rescoring and discuss how they can be used for the analysis of immunopeptidomics data. Finally, we provide insights into current and future machine learning solutions to boost immunopeptide identification.

Introduction

The adaptive immune system plays a crucial role in safeguarding the body against pathogen-infected and cancerous cells, by recognizing peptides bound to major histocompatibility complex (MHC) molecules on the cell surface [1]. Researchers have gained valuable insights into the immune responses by characterizing the immunopeptidome, which encompasses the repertoire of MHC-bound peptides presented for immunosurveillance. The field of immunopeptidomics has made remarkable advancements in recent years, enabling the identification of T cell targets against tumors, autoimmune diseases, and pathogens [2–5].

However, identifying immunopeptides from mass spectrometry (MS) data remains challenging, with a large portion of measured spectra that are left unannotated. In an immunopeptidomics experiment, the MHC-peptide complexes are purified and the immunopeptides are eluted [6] (Fig. 1). Unlike conventional tryptic peptides, the generation of immunopeptides lacks well-defined protease specificity rules, necessitating that every possible protein subsequence within human leukocyte antigen (HLA) class-specific length constraints has to be considered during spectrum annotation. Consequently, the search space significantly expands, resulting in increased false positive spectrum annotation rates and reduced peptide identification sensitivity [7]. A compounding factor hereof is that the search space in immunopeptidomics is often even further extended beyond the canonical human proteome, by incorporating somatic mutations [8], pathogen genomes [4], novel or unannotated open reading frames (nuORFs) [9,10], and post-translational modifications (PTMs) [11]. Additionally, the non-tryptic nature of immunopeptides results in unique spectral characteristics, such as strong internal ion series and neutral losses [12]. Because many immunopeptides do not contain basic lysine or arginine residues at the C-terminus, they are often singly charged [13], in contrast to multiply charged peptides encountered during bottom-up proteomics, making their MS analysis more intricate due to poor ionization and incomplete fragmentation coverage [14]. Moreover, immunopeptides are often present at low abundances, which can also cause changes in fragment ion intensities [15].

The most common approach to deduce the originating peptide behind an MS/MS spectrum is sequence database searching [16]. It involves comparing MS/MS spectra against all theoretically possible peptides derived from the protein sequences present in the database, in combination with decoy sequences that are used for statistical confidence estimation [17]. However, large databases, such as in immunopeptidomics, introduce challenges, such as an increase in random high-scoring matches, including high scoring decoys, resulting in fewer identifications at a given FDR (usually 1%) [18]. Consequently, numerous informative spectra are discarded, leading to the oversight of many potentially therapeutically relevant targets.

PSM rescoring has emerged as a promising solution to challenges in immunopeptidomics data analysis, reducing false positives and improving identification rates. This typically involves using (semi-)supervised machine learning (ML) algorithms, such as Percolator [19], to generate a new score, incorporating additional PSM features, to better distinguish between correct and incorrect PSMs, compared to the search engine score. Recent advancements in peptidoform property prediction have led to the incorporation of features for PSM rescoring based on machine learning predictions. For example, fragment ion intensity predictions can

now be used as a realistic proxy for ground truth MS/MS data from which spectral similarity features can be calculated for PSM rescoring. In this review, we will explore the tools currently available for PSM rescoring and discuss how they can be used to obtain more information from immunopeptidomics data.

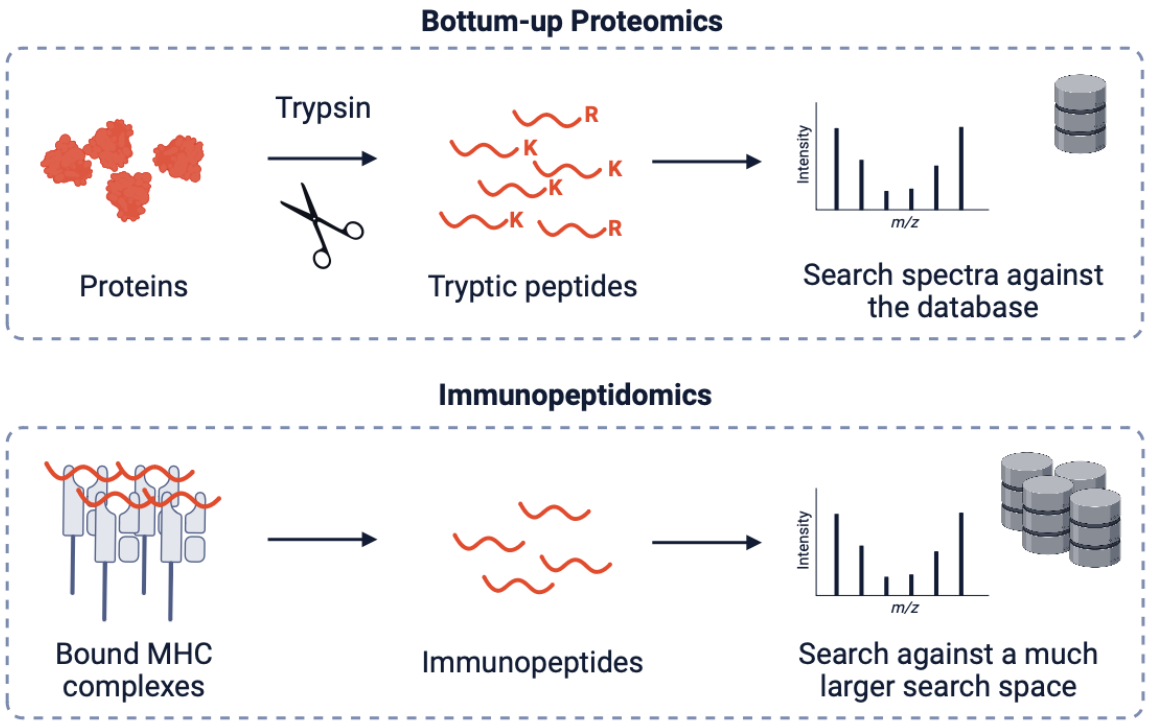

Fig. 1 Differences between bottom-up proteomics and immunopeptidomics. The immunopeptidomics experimental workflow differs from the bottom-up proteomics workflow based on the way peptides are obtained prior to MS analysis. In immunopeptidomics, peptides are processed *in vivo* or *in vitro* inside the cell. The peptide-bound MHC molecules are extracted and the immunopeptides are eluted. In bottom-up proteomics, the peptides are produced during an experimental step where a protease, usually trypsin, is used to produce the peptides. This allows a reduction of the search space based on known digestion rules, which is not possible in immunopeptidomics.

PSM Rescoring: Leveraging Machine Learning for Enhanced Immunopeptidomics Data Analysis

To address challenges in MS-based immunopeptidomics data analysis, PSM rescoring has emerged as a machine learning-based solution that can significantly reduce false positive spectrum assignments and improve immunopeptide identification rates. During PSM rescoring, machine learning algorithms such as Percolator [19], PeptideProphet [20], and mokapot [19], are used to post-process unfiltered database search results and learn an optimized score that effectively distinguishes between correct and incorrect PSMs based on various informative PSM features.

PSM rescoring is most often done using Percolator [19], which uses semi-supervised machine learning (ML) to build a linear support vector machine (SVM)-based classifier that can distinguish between correct and incorrect PSMs. It begins by (1) parsing the initial unfiltered database search results into two groups: PSMs with high scores from target peptides (positives) and PSMs from decoy peptides (negatives). Then (2) an SVM-based classifier is trained to discriminate between the positives and the negatives, based on features describing the PSMs. Finally, (3) a new score is calculated for all PSMs using the trained classifier. These three steps are repeated several times until convergence is reached. The newly computed scores should result in a greater number of confidently identified peptides, improving the peptide identification performance (Fig. 2).

To avoid overfitting, Percolator uses three-fold cross-validation during training of the SVM. The inputs are divided into three equally sized subsets and the classifier is trained three times. Each time all but one of the subsets is used for training and the excluded subset is used for validation. This process is repeated three times, so all combinations of training and validation sets are used. In addition, Percolator employs nested three-fold cross-validation within each training set to select the most suitable hyperparameters for the SVM training. The two training subsets are again divided into three parts, of which two serve as training and one as validation [22].

Notably, Percolator is extremely flexible, allowing its user to consider any set of features to describe a dataset of PSMs. Initially, features based on PSM metadata were used, such as the search engine score, the peptide length, and the charge state. However, recently several researchers have taken advantage of this flexibility by integrating features from newly developed ML-based predictors to boost the number of high-confidence identifications. The basic assumption is that correct peptides should closely resemble ground truth measurements, while features for incorrect peptides are more likely to follow a random distribution. Because it is infeasible to have ground truth measurements for all possible peptides and predictors have drastically improved in accuracy, predictions are used as a proxy for true ground truth data (Fig. 2).

It is important to note that for rescoring, features need to be carefully evaluated to ensure that they are unbiased. Features considered by the classifier should be very similar between decoys and incorrect target PSMs. Else, rather than resulting in a better separation between correct and incorrect PSMs, the classifier would systematically assign higher scores to target PSMs compared to the decoys and produce biased results. For example, a feature included in the initial version of Percolator [19] describing the number of PSMs that match to the same protein (numProt) was identified as a biased feature [23], probably because it is more likely that proteins from which incorrect target peptides are derived have a higher PSM count compared to decoy proteins.

To evaluate feature biases, entrapment searches can be used, in which MS/MS spectra are searched against the sample database and against an entrapment database consisting of sequences from evolutionary distant organisms. When the entrapment partition is sufficiently large, the probability that a random match hits the sample database should be negligible.

Entrapment PSMs are treated as a representation of incorrect matches and can be used to detect whether there is a correct estimation of the FDR [23–25].

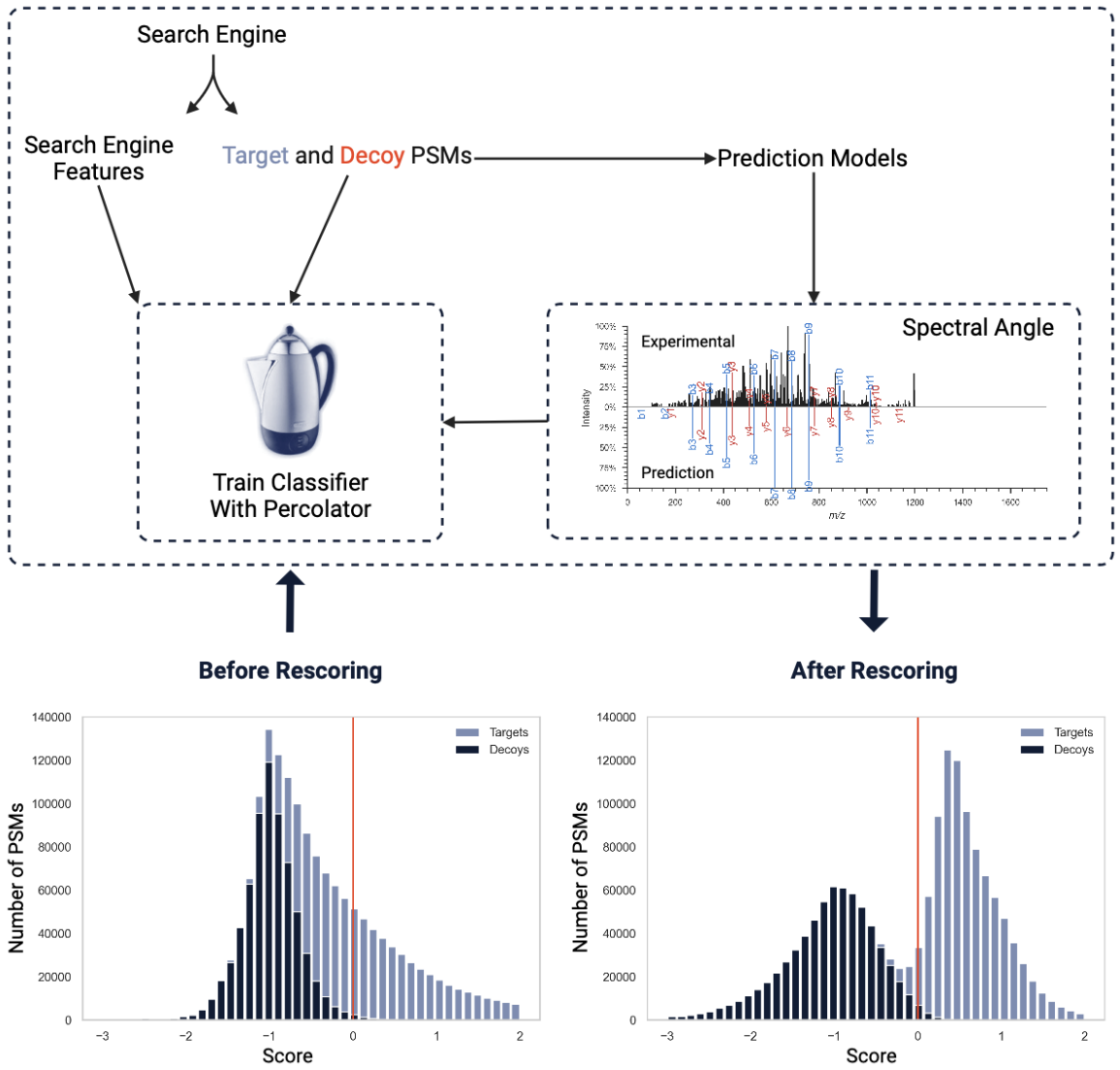

Fig. 2 Overview of the PSM rescoring pipeline. PSM rescoring is a post-processing step of unfiltered database search results. Features from the search engines, such as the search engine score, the peptide length, and the charge state are combined with machine learning-based features, such as spectrum similarity. PSM rescoring should result in a better separation between correct and incorrect PSMs, with more annotated spectra at a fixed FDR (1% shown as a red line).

Common feature types used during immunopeptide PSM rescoring

Existing rescoring tools mainly differ from each other by their use of distinct feature sets and prediction models. Some PSM rescoring tools use only a few features by default [26–28], while others use dozens [29,30] to 100 features (https://github.com/wilhelm-lab/oktoberfest)

[31]. Additionally, some tools allow adjusting the number of features used. For example, MSBooster [26] has an option to use correlated features or not, and inSPIRE [29] allows manual feature inclusion and exclusion. Although a minor variation in feature sets that are used will likely have limited effect on the performance, in light of the diversity in the number and type of features used by various PSM rescoring tools, it is important to strategically select relevant and informative features.

In addition to general information on the PSM, features can be added based on predictions, representing the similarity between the observed and predicted peptidoform properties, such as spectral similarity and retention time differences. Rescoring tools with only a few features use a single representative feature for each type, whereas more extensive feature sets include multiple related values as distinct features. The concept behind using such related features is that even when they exhibit some statistical correlation, they may not be completely redundant due to subtle variations in their information content and thus may still provide complementary signals to improve the predictive performance [32]. For example, the more extensive feature set of MS^{2}Rescore has been suggested as a possible driver of its performance [31]. It is only when features are perfectly correlated that they are truly redundant, in which case adding them yields no additional information [32].

However, employing an extensive feature set needs careful consideration. It increases the risk of overfitting, leading to a reduced number of PSMs identified when cross-validation is applied within Percolator. Because overfitting is especially an issue when a small dataset is used, which is common in immunopeptidomics because fewer PSMs are identified compared to standard proteomics, carefully considering the features used in the PSM rescoring pipeline is especially relevant.

Spectral similarity features

Intensity-based scores representing the similarity between an observed MS/MS spectrum and the predicted MS/MS spectrum for specific peptides are valuable features in rescoring [33] that are used by the majority of currently available rescoring tools (Table 1).

While the *m/z* values of the peptide fragments can easily be calculated from the corresponding amino acid sequence, the exact intensities of those fragments depend on the unique chemistry of each peptide sequence [34,35]. By predicting the fragment ion intensities we can thus generate an MS/MS spectrum. Accurate prediction of these fragment ion intensities has been an active area of research for a number of years [12,36–40]. It is important to note that because of the distinct spectral characteristics of immunopeptides, fragment ion intensity prediction tools should be retrained using immunopeptidomics data to drastically improve prediction accuracy [12,31].

Various spectral similarity features can be used. Basic features indicate the presence/absence of expected fragment ions and their *m/z* deviation, which does not require predicted fragment ion intensities. Additionally, various types of spectral correlation can be used based on predicted fragment ion intensities, such as pearson correlation, spearman correlation, mean square error, cosine similarity [41], normalized spectral contrast angle (SA) [34], spectral entropy [43], and others. Besides calculating such scores for the full spectrum, some PSM

rescoring tools also include similar features separately for b- and y-ions or when only considering the top-*k* peaks.

Another possible spectral feature is the Prosit-delta score. This feature, used by inSPIRE [29], represents the stability of the SA after two amino acid positions are switched, and is conceptually related to the ΔXCorr score originally used by Percolator [19]. Sequences with a low Prosit-delta are more likely to be misassigned to similar, though incorrect peptide sequences.

Retention time features

Retention time features capture the behavior of the analytes on the liquid chromatography (LC) column. Based on theoretical retention times predicted for the candidate peptides, the difference with the experimentally observed retention time (ΔRT) can be calculated, typically in terms of the absolute differences. Most rescoring pipelines incorporate retention time features (Table 1).

DeepRescore [28] reported that while dropping ΔRT and SA led to reduced peptide identifications, only minor reductions in the number of identified peptides were observed when only ΔRT and SA were considered. This indicates the dominant contributions of these features. The interplay between ΔRT and SA was further highlighted by the fact that eliminating either of these features resulted in a loss of peptide identifications. This observation implies a complementary contribution from both features, which is expected as the LC behavior of a peptide is independent of its MS behavior.

Ion mobility features

Ion mobility spectrometry provides an additional dimension that separates ions based on their size and shape in the gas phase. The measured ion mobility can be used to derive a collisional cross section (CCS), which intrinsically depends on the ion structure. Interestingly, isobaric and isomeric peptide sequences can potentially be distinguished by their different CCS [44]. The difference between the observed and the predicted CCS [45,46] can be used similarly as ΔRT. A few pipelines support the use of ion mobility features [26,30]. As ion mobility becomes more prevalent on modern instruments, ion mobility-based features will become more common.

In MSBooster [26] ion mobility features are not included in the default set of features, as they only marginally improved identifications. The limited benefit of ML-based ion mobility features may be explained by the high dependence of the CCS on the precursor mass and charge. Because decoy PSMs have the same charge and a similar mass as the unknown true target peptide, their CCS predictions are highly correlated with the experimental values. Potentially more relevant features are simply the peptide mass, length, charge state, and experimental ion mobility value.

Binding affinity features

When prior knowledge is available on the HLA alleles that are present in the sample, tools like NetMHCpan [47] can be used to predict the binding affinity of a certain peptide to a specific MHC molecule. Features based on these predictions incorporate biological knowledge, in contrast to the previously mentioned features that are purely chemistry based. It is important to note the biases that can be introduced by imposing expected binding motifs based on prior knowledge. Moreover, because some HLA alleles are better characterized than others, the strength of the predictors varies depending on the HLA allele, potentially introducing a bias towards well-known HLA alleles. This conflicts with part of the goal of using untargeted mass spectrometry-based immunopeptidomics for the unbiased discovery of immunopeptides. Predicted binding affinity based features are supported by only two PSM rescoring tools, MHCquant [27] and inSPIRE [29].

inSPIRE [29] uses MHC binding affinity predictions in combination with RT and fragment ion intensity predictions and showed that not only peptides with higher binding affinities were identified, but also those with overall better spectral features. This suggests that the features work in concert rather than one being prioritized over the others. In contrast, MHCquant [27] only uses binding affinity as a feature and thus has an increased risk of introducing a bias, resulting in the identification of mainly already known binders. Other features, such as spectral similarity features and retention time features, provide a quality check. Without these features there is no filter for low-quality spectra, which can lead to an increased risk of introducing false positive peptide identifications primarily based on a couple of amino acids on specific anchor positions.

Considerations for the prediction models

Prediction models are used instead of ground truth values, as it would be infeasible to obtain reference measurements for all potential candidate peptides, especially in an immunopeptidomics setting. Evidently, the closer the predictions are to the ground truth, the more relevant the features will be. For example, MHCquant [27] tested ΔRT and concluded that it was not an effective feature. In contrast, for DeepRescore [28] ΔRT was consistently one of the features with the highest weight. A possible explanation for this apparent contradictory conclusion is that the RTPredict tool [48,49] MHCquant used was published more than 15 years ago and might provide less accurate predictions compared to more recent alternative retention time prediction models. Hence, this illustrates that highly accurate predicted peptide properties are needed to provide informative features for PSM rescoring.

Improvements in prediction models could be implemented in the future to obtain more effective features. In the last couple of years, transformers [50] have rapidly become the dominant neural network architecture, including several applications in bioinformatics. For example, a transformer, in combination with a pretraining strategy, outperformed Prosit's recurrent neural network in predicting fragment ion intensities [51]. Another way to improve the model performance is to use transfer learning. This is a process by which a neural network is first trained on a large dataset, after which it is adapted to another related task or dataset by finetuning it on a smaller training dataset. In this way, model performance can be optimized

for specific datasets, even when only a limited amount of training data is available. Transfer learning is used by several deep learning tools [30,39,52] and may help to create models better suited for different scenarios, such as finetuning them for different fragmentation mechanisms, instrument platforms, and even lab-specific data properties [53].

Future perspectives

Further expand the feature set

Originally the Percolator feature set included basic PSM features such as the search engine score, the peptide length, and the charge state, which has currently been extended with features based on ML predictions. As PSM rescoring continues to evolve, we expect that features will be updated employing improved prediction models and that the currently used feature sets will be extended with novel features containing orthogonal information. For example, based on the isotope distribution it is possible to determine the number of Sulphur atoms present in the sample, indicating the number of cysteine residues [54].

Currently the spectral features are typically only based on canonical b- and y-ions, optionally considering common neutral losses as well. With the advent of full spectrum prediction models [55] this might be expanded to the entire MS/MS spectrum, using the intensities for every *m/z* bin. This may include internal fragment ion intensities, which are common for immunopeptides that lack basic C-terminal residues.

Identification of post-translational modifications

Post-translational modifications (PTMs) can regulate key processes by altering protein structure, location, or function [56]. These modified proteins, along with their unmodified counterparts, enter antigen processing pathways resulting in both unmodified and modified immunopeptides being presented by MHC molecules [57]. PTMs on immunopeptides have been reported to modulate antigen presentation and recognition [58].

However, because modified peptides can have different properties compared to their unmodified counterparts, predictions from deep learning tools trained on modified peptides are expected to perform better than simply shifting fragment ions and using the same retention time as for the unmodified peptide. With the use of transfer learning, the fragment ion intensity prediction tool pDeep2 [59] was extended to include 22 PTMs. Notably, it is difficult to train good models for low-abundant PTMs, as unlike for common PTMs, it is more challenging to collect sufficient MS/MS spectra.

When there are a couple of PTMs of interest, it is recommended to choose up to about 5 PTMs in a standard sequence database search. When a more explorative approach is preferred, one could consider using open modification searching, an unbiased approach for investigating PTMs [60]. In standard database searching an observed spectrum is not scored against the complete database, instead, only a subset of the peptides are scored, namely those whose mass is within a narrow tolerance window of the observed precursor mass. When this is extended to a larger window, a so-called open search is performed in which a

modified peptide can be matched to its unmodified counterpart, enabling the identification of any PTM within the mass window.

There are some technical challenges when performing open modification searching that are exacerbated by the large search space in immunopeptidomics. Efficient new tools tackle these technical challenges and have been starting to integrate rescoring [30,61,62]. However, it is important to note that currently most PTMs found with open modification searching might not be supported by the underlying predictors. In this case, transfer learning can be used [52], as well as emerging algorithmic solutions that can predict peptide properties for unseen PTMs [30,63]. Additionally, even though several approaches have been proposed to control the FDR in open modification searching [64], further research is necessary to investigate how the FDR behaves and to ensure that PSM rescoring does not introduce biases.

De novo peptide sequencing

An alternative to standard sequence database searching and open modification searching that allows the identification of novel or unexpected peptides is *de novo* peptide sequencing. Rather than searching for peptides in a database, the peptide sequence is directly determined from the MS/MS spectrum [65]. This provides some advantages in immunopeptidomics, where it is challenging to construct a complete database accounting for genetic variation, unannotated open reading frames, and the presence of pathogens. However, because spectra might not contain a complete fragmentation pattern, as well as the presence of noise peaks, *de novo* peptide sequencing is a challenging task [66].

Deep learning has led to a new generation of *de novo* sequencing tools, with promising applications in immunopeptidomics [67–70]. An important consideration is that due to the different peptide and spectral characteristics of immunopeptidomics data, depending on the initial training strategy, such deep learning-based *de novo* sequencing tools typically need to be optimized to handle immunopeptides, for example to avoid a tryptic bias in their results [66].

Adjustment to new experimental settings

When new experimental setups are being implemented, prediction models and feature sets will need to be updated and developed. For example, while the initial Prosit fragment ion intensity prediction model was trained on data from Thermo Scientific Orbitrap instruments [38], it was recently optimized to accurately predict fragment ion intensities for timsTOF data as well by finetuning the original model [53]. Similarly, as new instrument platforms are introduced, such as electron activated dissociation provided by Sciex ZenoTOF instruments or data from Thermo Scientific Astral instruments, new prediction models need to be developed. A challenge with developing prediction models for new experimental settings is that initially only a limited amount of training data might be available. Therefore, if there is still some similarity between the new experimental set-up and previous models, such as similar fragmentation methods on related instrument platforms, finetuning can be used instead [53]. In contrast, when the two settings are sufficiently different, retraining new prediction models from scratch might be necessary. When using PSM rescoring there needs to be a careful

consideration of the settings in which the machine learning models were developed, as using suboptimal prediction models can lead to unexpected gaps in performance, as demonstrated by the RT example previously described.

Conclusion

The integration of machine learning into immunopeptidomics data analysis through PSM rescoring holds immense potential. By increasing the number of MS/MS spectra that can be accurately annotated, this approach addresses the important challenge of immunopeptide identification, which can ultimately lead to advances in our understanding of immune responses. While PSM rescoring has gained in popularity over the past few years and has grown into a common analysis strategy, we expect that it will become even more ubiquitous in the near future by shifting from an external post-processing step to a core component integrated in search engines directly.

Conflict of Interest

K.L. holds shares in ImmuneWatch BV: an immunoinformatics company. K.B. is a co-founder and shareholder of ImmuneSpec BV.

References

[1] Vaughan, K., Xu, X., Caron, E., Peters, B., Sette, A., Deciphering the MHC-associated peptidome: a review of naturally processed ligand data. *Expert Rev. Proteomics* 2017, 14, 729–736.

[2] Peltonen, K., Feola, S., Umer, H.M., Chiaro, J., et al., Therapeutic Cancer Vaccination with Immunopeptidomics-Discovered Antigens Confers Protective Antitumor Efficacy. *Cancers* 2021, 13, 3408.

[3] Prinz, J.C., Immunogenic self-peptides - the great unknowns in autoimmunity: Identifying T-cell epitopes driving the autoimmune response in autoimmune diseases. *Front. Immunol.* 2023, 13.

[4] Leddy, O.K., White, F.M., Bryson, B.D., Leveraging Immunopeptidomics To Study and Combat Infectious Disease. *mSystems* 2021, 6, e00310-21.

[5] Mayer, R.L., Impens, F., Immunopeptidomics for next-generation bacterial vaccine development. *Trends Microbiol.* 2021, 29, 1034–1045.

[6] Purcell, A.W., Ramarathinam, S.H., Ternette, N., Mass spectrometry–based identification of MHC-bound peptides for immunopeptidomics. *Nat. Protoc.* 2019, 14, 1687–1707.

[7] Nesvizhskii, A.I., A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics* 2010, 73, 2092–2123.

[8] Tretter, C., de Andrade Krätzig, N., Pecoraro, M., Lange, S., et al., Proteogenomic analysis reveals RNA as a source for tumor-agnostic neoantigen identification. *Nat. Commun.* 2023, 14, 4632.

[9] Guilloy, N., Brunet, M.A., Leblanc, S., Jacques, J.-F., et al., OpenCustomDB: Integration of Unannotated Open Reading Frames and Genetic Variants to Generate More Comprehensive Customized Protein Databases. *J. Proteome Res.* 2023, 22, 1492–1500.

[10] Ouspenskaia, T., Law, T., Clauser, K.R., Klaeger, S., et al., Unannotated proteins expand the MHC-I-restricted immunopeptidome in cancer. *Nat. Biotechnol.* 2022, 40, 209–217.

[11] Kacen, A., Javitt, A., Kramer, M.P., Morgenstern, D., et al., Uncovering the modified immunopeptidome reveals insights into principles of PTM-driven antigenicity 2021,

2021.04.10.438991.
[12] Wilhelm, M., Zolg, D.P., Graber, M., Gessulat, S., et al., Deep learning boosts sensitivity of mass spectrometry-based immunopeptidomics. *Nat. Commun.* 2021, 12, 3346.
[13] Chen, R., Fauteux, F., Foote, S., Stupak, J., et al., Chemical Derivatization Strategy for Extending the Identification of MHC Class I Immunopeptides. *Anal. Chem.* 2018, 90, 11409–11416.
[14] Pfammatter, S., Bonneil, E., Lanoix, J., Vincent, K., et al., Extending the Comprehensiveness of Immunopeptidome Analyses Using Isobaric Peptide Labeling. *Anal. Chem.* 2020, 92, 9194–9204.
[15] Orsburn, B.C., Time-of-Flight Fragmentation Spectra Generated by the Proteomic Analysis of Single Human Cells Do Not Exhibit Atypical Fragmentation Patterns. *J. Proteome Res.* 2023.
[16] Eng, J.K., Searle, B.C., Clauser, K.R., Tabb, D.L., A Face in the Crowd: Recognizing Peptides Through Database Search. *Mol. Cell. Proteomics MCP* 2011, 10, R111.009522.
[17] Elias, J.E., Gygi, S.P., Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* 2007, 4, 207–214.
[18] Mann, M., Wilm, M., Error-Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags. *Anal. Chem.* 1994, 66, 4390–4399.
[19] Käll, L., Canterbury, J.D., Weston, J., Noble, W.S., MacCoss, M.J., Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* 2007, 4, 923–925.
[20] Keller, A., Nesvizhskii, A.I., Kolker, E., Aebersold, R., Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. *Anal. Chem.* 2002, 74, 5383–5392.
[21] Fondrie, W.E., Noble, W.S., mokapot: Fast and Flexible Semisupervised Learning for Peptide Detection. *J. Proteome Res.* 2021, 20, 1966–1971.
[22] Granholm, V., Noble, W.S., Käll, L., A cross-validation scheme for machine learning algorithms in shotgun proteomics. *BMC Bioinformatics* 2012, 13, S3.
[23] Granholm, V., Noble, W.S., Käll, L., On Using Samples of Known Protein Content to Assess the Statistical Calibration of Scores Assigned to Peptide-Spectrum Matches in Shotgun Proteomics. *J. Proteome Res.* 2011, 10, 2671–2678.
[24] Feng, X., Li, L., Zhang, J., Zhu, Y., et al., Using the entrapment sequence method as a standard to evaluate key steps of proteomics data analysis process. *BMC Genomics* 2017, 18, 143.
[25] Vaudel, M., Burkhart, J.M., Breiter, D., Zahedi, R.P., et al., A Complex Standard for Protein Identification, Designed by Evolution. *J. Proteome Res.* 2012, 11, 5065–5071.
[26] Yang, K.L., Yu, F., Teo, G.C., Li, K., et al., MSBooster: improving peptide identification rates using deep learning-based features. *Nat. Commun.* 2023, 14, 4539.
[27] Bichmann, L., Nelde, A., Ghosh, M., Heumos, L., et al., MHCquant: Automated and Reproducible Data Analysis for Immunopeptidomics. *J. Proteome Res.* 2019, 18, 3876–3884.
[28] Li, K., Jain, A., Malovannaya, A., Wen, B., Zhang, B., DeepRescore: Leveraging Deep Learning to Improve Peptide Identification in Immunopeptidomics. *PROTEOMICS* 2020, 20, 1900334.
[29] Cormican, J.A., Horokhovskyi, Y., Soh, W.T., Mishto, M., Liepe, J., inSPIRE: An Open-Source Tool for Increased Mass Spectrometry Identification Rates Using Prosit Spectral Prediction. *Mol. Cell. Proteomics* 2022, 21.
[30] Zeng, W.-F., Zhou, X.-X., Willems, S., Ammar, C., et al., AlphaPeptDeep: a modular deep learning framework to predict peptide properties for proteomics. *Nat. Commun.* 2022, 13, 7238.
[31] Declercq, A., Bouwmeester, R., Hirschler, A., Carapito, C., et al., MS2Rescore: Data-Driven Rescoring Dramatically Boosts Immunopeptide Identification Rates. *Mol. Cell.*

Proteomics 2022, 21, 100266.
[32] Guyon, I., Elisseeff, A., An Introduction to Variable and Feature Selection n.d.
[33] C Silva, A.S., Bouwmeester, R., Martens, L., Degroeve, S., Accurate peptide fragmentation predictions allow data driven approaches to replace and improve upon proteomics search engine scoring functions. *Bioinforma. Oxf. Engl.* 2019, 35, 5243–5248.
[34] Huang, Y., Triscari, J.M., Tseng, G.C., Pasa-Tolic, L., et al., Statistical Characterization of the Charge State and Residue Dependence of Low-Energy CID Peptide Dissociation Patterns. *Anal. Chem.* 2005, 77, 5800–5813.
[35] Tabb, D.L., Smith, L.L., Breci, L.A., Wysocki, V.H., et al., Statistical Characterization of Ion Trap Tandem Mass Spectra from Doubly Charged Tryptic Peptides. *Anal. Chem.* 2003, 75, 1155–1163.
[36] Elias, J.E., Gibbons, F.D., King, O.D., Roth, F.P., Gygi, S.P., Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nat. Biotechnol.* 2004, 22, 214–219.
[37] Degroeve, S., Maddelein, D., Martens, L., MS2PIP prediction server: compute and visualize MS2 peak intensity predictions for CID and HCD fragmentation. *Nucleic Acids Res.* 2015, 43, W326-330.
[38] Gessulat, S., Schmidt, T., Zolg, D.P., Samaras, P., et al., Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* 2019, 16, 509–518.
[39] Zhou, X.-X., Zeng, W.-F., Chi, H., Luo, C., et al., pDeep: Predicting MS/MS Spectra of Peptides with Deep Learning. *Anal. Chem.* 2017, 89, 12690–12697.
[40] Meyer, J.G., Deep learning neural network tools for proteomics. *Cell Rep. Methods* 2021, 1, 100003.
[41] Stein, S.E., Scott, D.R., Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass Spectrom.* 1994, 5, 859–866.
[42] Toprak, U.H., Gillet, L.C., Maiolica, A., Navarro, P., et al., Conserved Peptide Fragmentation as a Benchmarking Tool for Mass Spectrometers and a Discriminating Feature for Targeted Proteomics*. *Mol. Cell. Proteomics* 2014, 13, 2056–2071.
[43] Li, Y., Kind, T., Folz, J., Vaniya, A., et al., Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification. *Nat. Methods* 2021, 18, 1524–1531.
[44] Wu, C., Siems, W.F., Klasmeier, J., Hill, H.H., Separation of isomeric peptides using electrospray ionization/high-resolution ion mobility spectrometry. *Anal. Chem.* 2000, 72, 391–395.
[45] Meier, F., Köhler, N.D., Brunner, A.-D., Wanka, J.-M.H., et al., Deep learning the collisional cross sections of the peptide universe from a million experimental values. *Nat. Commun.* 2021, 12, 1185.
[46] Demichev, V., Szyrwiel, L., Yu, F., Teo, G.C., et al., dia-PASEF data analysis using FragPipe and DIA-NN for deep proteomics of low sample amounts. *Nat. Commun.* 2022, 13, 3944.
[47] Reynisson, B., Alvarez, B., Paul, S., Peters, B., Nielsen, M., NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.* 2020, 48, W449–W454.
[48] Pfeifer, N., Leinenbach, A., Huber, C.G., Kohlbacher, O., Statistical learning of peptide retention behavior in chromatographic separations: a new kernel-based approach for computational proteomics. *BMC Bioinformatics* 2007, 8, 468.
[49] Pfeifer, N., Leinenbach, A., Huber, C.G., Kohlbacher, O., Improving peptide identification in proteome analysis by a two-dimensional retention time filtering approach. *J. Proteome Res.* 2009, 8, 4109–4115.
[50] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., et al., in:, *Adv. Neural Inf. Process. Syst.*, vol. 30, Curran Associates, Inc., 2017.
[51] Ekvall, M., Truong, P., Gabriel, W., Wilhelm, M., Käll, L., Prosit Transformer: A transformer

for Prediction of MS2 Spectrum Intensities. *J. Proteome Res.* 2022, 21, 1359–1364.
[52] Wilburn, D.B., Shannon, A.E., Spicer, V., Richards, A.L., et al., Deep learning from harmonized peptide libraries enables retention time prediction of diverse post translational modifications 2023, 2023.05.30.542978.
[53] Adams, C., Gabriel, W., Laukens, K., Wilhelm, M., et al., Fragment ion intensity prediction improves the identification rate of non-tryptic peptides in TimsTOF 2023, 2023.07.17.549401.
[54] Agten, A., Vilenne, F., Valkenborg, D., *A compositional data model to predict the isotope distribution for average peptides using a compositional spline model.*, Preprints, 2023.
[55] Liu, K., Li, S., Wang, L., Ye, Y., Tang, H., Full-Spectrum Prediction of Peptides Tandem Mass Spectra using Deep Neural Network. *Anal. Chem.* 2020, 92, 4275–4283.
[56] Uy, R., Wold, F., Posttranslational covalent modification of proteins. *Science* 1977, 198, 890–896.
[57] Engelhard, V.H., Altrich-Vanlith, M., Ostankovitch, M., Zarling, A.L., Post-translational modifications of naturally processed MHC-binding epitopes. *Curr. Opin. Immunol.* 2006, 18, 92–97.
[58] Ramarathinam, S.H., Croft, N.P., Illing, P.T., Faridi, P., Purcell, A.W., Employing proteomics in the study of antigen presentation: an update. *Expert Rev. Proteomics* 2018, 15, 637–645.
[59] Zeng, W.-F., Zhou, X.-X., Zhou, W.-J., Chi, H., et al., MS/MS Spectrum Prediction for Modified Peptides Using pDeep2 Trained by Transfer Learning. *Anal. Chem.* 2019, 91, 9724–9731.
[60] Chick, J.M., Kolippakkam, D., Nusinow, D.P., Zhai, B., et al., A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nat. Biotechnol.* 2015, 33, 743–749.
[61] Arab, I., Fondrie, W.E., Laukens, K., Bittremieux, W., Semisupervised Machine Learning for Sensitive Open Modification Spectral Library Searching. *J. Proteome Res.* 2023, 22, 585–593.
[62] Freestone, J., Noble, W.S., Keich, U., Analysis of tandem mass spectrometry data with CONGA: Combining Open and Narrow searches with Group-wise Analysis 2023, 2023.05.02.539167.
[63] Bouwmeester, R., Gabriels, R., Hulstaert, N., Martens, L., Degroeve, S., DeepLC can predict retention times for peptides that carry as-yet unseen modifications. *Nat. Methods* 2021, 18, 1363–1369.
[64] Freestone, J., Short, T., Noble, W.S., Keich, U., Group-walk: a rigorous approach to group-wise false discovery rate analysis by target-decoy competition. *Bioinformatics* 2022, 38, ii82–ii88.
[65] Muth, T., Hartkopf, F., Vaudel, M., Renard, B.Y., A Potential Golden Age to Come—Current Tools, Recent Use Cases, and Future Avenues for De Novo Sequencing in Proteomics. *PROTEOMICS* 2018, 18, 1700150.
[66] Yilmaz, M., Fondrie, W., Bittremieux, W., Oh, S., Noble, W.S., in:, *Proc. 39th Int. Conf. Mach. Learn.*, PMLR, 2022, pp. 25514–25522.
[67] Yilmaz, M., Fondrie, W.E., Bittremieux, W., Nelson, R., et al., Sequence-to-sequence translation from mass spectra to peptides with a transformer model 2023, 2023.01.03.522621.
[68] Tran, N.H., Qiao, R., Xin, L., Chen, X., et al., Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nat. Methods* 2019, 16, 63–66.
[69] Qiao, R., Tran, N.H., Xin, L., Chen, X., et al., Computationally instrument-resolution-independent de novo peptide sequencing for high-resolution devices. *Nat. Mach. Intell.* 2021, 3, 420–425.
[70] Klaproth-Andrade, D., Hingerl, J., Smith, N.H., Träuble, J., et al., Deep learning-driven

fragment ion series classification enables highly precise and sensitive de novo peptide sequencing 2023, 2023.01.05.522752.
[71] Strauss, M.T., Bludau, I., Zeng, W.-F., Voytik, E., et al., AlphaPept, a modern and open framework for MS-based proteomics 2021, 2021.07.23.453379.
[72] Cox, J., Mann, M., MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* 2008, 26, 1367–1372.
[73] Demichev, V., Messner, C.B., Vernardis, S.I., Lilley, K.S., Ralser, M., DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods* 2020, 17, 41–44.
[74] Wen, B., Li, K., Zhang, Y., Zhang, B., Cancer neoantigen prioritization through sensitive and reliable proteogenomics analysis. *Nat. Commun.* 2020, 11, 1759.
[75] Eng, J.K., Jahan, T.A., Hoopmann, M.R., Comet: an open-source MS/MS sequence database search tool. *Proteomics* 2013, 13, 22–24.
[76] Kim, S., Pevzner, P.A., MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* 2014, 5, 5277.
[77] Craig, R., Beavis, R.C., TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 2004, 20, 1466–1467.
[78] Zolg, D.P., Gessulat, S., Paschke, C., Graber, M., et al., INFERYS rescoring: Boosting peptide identifications and scoring confidence of database search results. *Rapid Commun. Mass Spectrom.* n.d., n/a, e9128.
[79] Eng, J.K., McCormack, A.L., Yates, J.R., An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* 1994, 5, 976–989.
[80] Levitsky, L.I., Klein, J.A., Ivanov, M.V., Gorshkov, M.V., Pyteomics 4.0: Five Years of Development of a Python Proteomics Framework. *J. Proteome Res.* 2019, 18, 709–714.
[81] Perkins, D.N., Pappin, D.J., Creasy, D.M., Cottrell, J.S., Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999, 20, 3551–3567.
[82] Zhang, J., Xin, L., Shan, B., Chen, W., et al., PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol. Cell. Proteomics MCP* 2012, 11, M111.010587.
[83] O'Donnell, T.J., Rubinsteyn, A., Laserson, U., MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing. *Cell Syst.* 2020, 11, 42-48.e7.
[84] Vaudel, M., Burkhart, J.M., Zahedi, R.P., Oveland, E., et al., PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat. Biotechnol.* 2015, 33, 22–24.
[85] Kong, A.T., Leprevost, F.V., Avtonomov, D.M., Mellacheruvu, D., Nesvizhskii, A.I., MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nat. Methods* 2017, 14, 513–520.
[86] Yu, F., Teo, G.C., Kong, A.T., Fröhlich, K., et al., Analysis of DIA proteomics data using MSFragger-DIA and FragPipe computational platform. *Nat. Commun.* 2023, 14, 4154.

Table

Rescoring pipeline	RT	MS/MS	CCS	BA	Availability?	Open source?	Supported platforms	Usage restrictions	Search engine
AlphaPeptDeep [30]	integrated	integrated	integrated	N/A	PyPI, https://github.com/MannLabs/alphapeptdeep	Yes	DDA, DIA, DDA-PASEF	Apache-2.0 license	AlphaPept [71], MaxQuant [72], DIA-NN [73]
DeepRescore [28]	AutoRT [74]	pDeep2 [39]	N/A	N/A	Nextflow, https://github.com/bzhanglab/DeepRescore	Yes	DDA	Unspecified	Comet [75], MaxQuant [72], MS-GF+ [76], and X!Tandem [77]
INFERYS [78]	N/A	integrated	N/A	N/A	integrated in Thermo Scientific™ Proteome Discoverer™ software	No	DDA	Commercial	Sequest HT [79]
inSPIRE [29]	Prosit [12] / pyteomics [80]	Prosit [12] / MS^{2}PIP [37]	N/A	netMHCpan [47]	https://github.com/QuantSysBio/inSPIRE	Yes	DDA	GPL-2.0 license	Mascot [81], MaxQuant [72], PEAKS [82]
MHCquant [27]	N/A	N/A	N/A	MHCFlurry [83]	https://github.com/nf-core/mhcquant	Yes	DDA	MIT license	Comet [75] (Integrated)
MS^{2}Rescore [31]	DeepLC [63]	MS^{2}PIP [37]	N/A	N/A	https://github.com/compomics/ms2rescore	Yes	DDA	Apache-2.0 license	MaxQuant [72], MS-GF+ [76], PEAKS [82], PeptideShaker [84], X!Tandem [77]

MSBooster [26]	DIA-NN [46,73]	DIA-NN [46,73]	DIA-NN [46,73]	N/A	Integrated into FragPipe https://fragpipe.nesvilab.org/ https://github.com/Nesvilab/MSBooster	Yes	DDA, DIA, DDA-PASEF	LGPL-3.0 license	MSFragger [85], MSFragger-DIA [86]
Oktoberfest	Prosit [12]	Prosit [12]	N/A	N/A	https://github.com/wilhelm-lab/oktoberfest	Yes	DDA	MIT license	MaxQuant [72], MSFragger [85], custom

Table 1 Overview of currently available PSM rescoring tools used in immunopeptidomics data analysis.