Received: 27 May 2023 | Revised: 29 December 2023 | Accepted: 5 January 2024

DOI: 10.1002/pmic.202300144

RESEARCH ARTICLE

Proteomics
Proteomics and Systems Biology

Intensity and retention time prediction improves the rescoring of protein-nucleic acid cross-links

Arslan Siraj[1,2] | Robbin Bouwmeester[3,4] | Arthur Declercq[3,4] | Luisa Welp[5,6] | Aleksandar Chernev[5] | Alexander Wulf[5] | Henning Urlaub[5,6] | Lennart Martens[3,4] | Sven Degroeve[3,4] | Oliver Kohlbacher[1,2] | Timo Sachsenberg[1,2]

[1]Department of Computer Science, Applied Bioinformatics, University of Tübingen, Tübingen, Germany

[2]Institute for Biological and Medical Informatics, University of Tübingen, Tübingen, Germany

[3]Department of Biomolecular Medicine, Ghent University, Gent, Belgium

[4]VIB-UGent Center for Medical Biotechnology, VIB, Gent, Belgium

[5]Bioanalytical Mass Spectrometry, Max Planck Institute for Multidisciplinary Sciences, Göttingen, Germany

[6]Bioanalytics, Institute of Clinical Chemistry, University Medical Center Göttingen, Göttingen, Germany

Correspondence
Timo Sachsenberg, Department of Computer Science, University of Tübingen, Tübingen, Germany.
Email: timo.sachsenberg@uni-tuebingen.de

Funding information
European Union's Horizon 2020 research and innovation program; Marie Skłodowska-Curie, Grant/Award Number: 956148; Ministry of Science, Research and Arts Baden-Württemberg; Deutsche Forschungsgemeinschaft, Grant/Award Numbers: SFB 860, SFB 1565, SPP 1935; Research Foundation Flanders (FWO), Grant/Award Numbers: 12B7123N, G010023N, G028821N, 1SE3722; Vlaams Agentschap Innoveren en Ondernemen, Grant/Award Number: HBC.2020.2205; European Union's Horizon 2020 Programme, Grant/Award Numbers: H2020-INFRAIA-2018-1, 823839; Ghent University Concerted Research Action, Grant/Award Number: BOF21/GOA/033

Abstract

In protein-RNA cross-linking mass spectrometry, UV or chemical cross-linking introduces stable bonds between amino acids and nucleic acids in protein-RNA complexes that are then analyzed and detected in mass spectra. This analytical tool delivers valuable information about RNA-protein interactions and RNA docking sites in proteins, both in vitro and in vivo. The identification of cross-linked peptides with oligonucleotides of different length leads to a combinatorial increase in search space. We demonstrate that the peptide retention time prediction tasks can be transferred to the task of cross-linked peptide retention time prediction using a simple amino acid composition encoding, yielding improved identification rates when the prediction error is included in rescoring. For the more challenging task of including fragment intensity prediction of cross-linked peptides in the rescoring, we obtain, on average, a similar improvement. Further improvement in the encoding and fine-tuning of retention time and intensity prediction models might lead to further gains, and merit further research.

KEYWORDS
fragment peak intensities, protein-RNA cross-linking mass spectrometry, rescoring, retention time, transfer learning

Abbreviations: LC-MS, liquid chromatography mass spectrometry; PARs, photoactivatable ribonucleosides; UV, ultraviolet; 4SU, 4-thiouridine; NM, nitrogen mustard; DEB, diepoxybutane; RBPs, RNA binding proteins; TiO_2, titanium dioxide; PTMs, post-translational modifications; MS/MS, tandem mass spectra; HCD, high-energy collision-induced dissociation; FDR, false discovery rate; PSMs, peptide-spectrum matches; CSMs, crosslink-spectrum matches; FMR, false match rate.

1 | INTRODUCTION

Interactions between proteins and RNA are important for many biological processes, such as gene expression, RNA splicing, and translation [1, 2]. Understanding these interactions is essential for the identification of therapeutic targets and development of novel therapies [3]. Protein-RNA crosslinking has become an increasingly popular technique for studying protein-RNA interactions. With advancements in mass spectrometry (MS) techniques, it is used to identify nucleotide-binding sites in cross-linked complexes with single amino acid precision [4]. UV-induced crosslinking uses irradiation with UVC (254 nm) to induce covalent bonds between predominantly U nucleotides and amino acids in close spacial proximity, while UVA (365 nm) is used to selectively activate RNA-incorporated photoactivatable ribonucleosides (PARs) to crosslink RNA binding proteins (RBPs) [5, 6]. PARs such as 4-thiouridine (4SU) are commonly used in UVA-based crosslinking. In chemical protein-RNA crosslinking, reactive reagents like nitrogen mustard (NM) or diepoxybutane (DEB) act as linkers between nucleotides and amino acids and can bind almost all RNA bases [7]. Typically, chemical crosslinkers harbor two functional groups that are reactive toward amino acids and nucleotides allowing them to form covalent bonds between the molecule species. Both UV and chemical crosslinking create stable bonds between RNA and proteins. During sample preparation, proteases (e.g., Trypsin) and RNAses cleave proteins and RNA into manageable sizes. Typically, the length of the oligonucleotides cross-linked to peptides varies between 1 and 4 nucleotides after digestion and may depend on the protocol [8–10]. Cross-linked RNA-peptide heteroconjugates are enriched using titanium dioxide (TiO_2) solid phase extraction which removes the non-crosslinked peptides. The purified heteroconjugates are separated by liquid chromatography and injected into a mass spectrometer during LC-MS analysis.

However, these cross-links are sensitive to sample processing and several other factors (denaturation, digestion, chemical instability of cross-linked adducts, fragmentation behavior, etc.) which lead to a wide range of neutral losses on precursor ions, as well as fragment ions [11, 12]. In contrast to classic post-translational modifications (PTMs), cross-links may have oligonucleotides of different of oligonucleotides bound to a peptide, unspecific binding to residues, and the presence of numerous neutral losses (e.g., water, ammonia, base, and ribose losses) observed in protein-RNA cross-linking poses additional challenges to their identification and a large number of candidates that need to be considered [5, 13, 14].

Like traditional peptide database search in shotgun proteomics, protein-RNA crosslink peptides can be computationally identified from tandem mass spectra (MS/MS) by comparing them to theoretical spectra derived from protein databases. Specialized protein-RNA crosslink search engines support large numbers of delta and neutral loss masses caused by the crosslinking chemistry, crosslinking reagents, and protocols need to accommodate for the diverse precursor and fragmentation adducts [15, 16]. We recently developed NuXL (publication in progress), a nucleotide cross-link search engine that searches cross-linked peptide candidates while considering their unique fragmentation behavior. NuXL controls the false-discovery rate at the level of cross-links and supports an extendable set of UV and chemical cross-linking protocols.

In addition to the challenges associated with the increased search space, high-energy collision-induced dissociation (HCD) fragmentation of cross-links also yields more complex fragmentation patterns than non-cross-linked peptides. The presence of RNA moieties in cross-linked peptides results in the formation of additional fragment ions typically not observed in non-cross-linked peptides, for example, prefix ions (b-ions), suffix ions (y-ions), immonium ions, or marker ions, all of which may carry RNA derived fragment adducts [15]. Since protein-RNA cross-links often occur in low abundance, cross-linked peptide levels are also relatively low in comparison to unmodified peptides in the sample despite various enrichment strategies [2, 10]. Consequently, the identification of cross-linked peptides is considered challenging.

Rescoring search engine results using a semi-supervised machine learning approach, as implemented by the percolator algorithm, is an established method frequently applied to improve identification rates. Percolator trains a linear SVM on multiple peptide-spectrum matches (PSMs) characteristics (like scores, multiple subscores, or general peptide properties) to improve discrimination between true and false identifications and to assign a more discriminative, combined score [17, 18]. Recently, adding peptide properties in the rescoring that are themselves derived from retention time and intensity prediction methods has become common.

Significance Statement

Protein-RNA cross-linking mass spectrometry is an emerging method used to study protein-RNA interactions that play a pivotal role in a cell. Crosslink information on amino acid level delivers valuable spatial information that can help in deducing functional and structural information about proteins. The cross-linking reaction leads to a large search space of structurally and physicochemical diverse modifications that lead to complex fragmentation patterns. Because of low cross-linking yields and signal intensities, the confident identification of protein-RNA cross-links is challenging and an active area of research. In this study, we investigate the applicability of approaches developed for unmodified peptides or peptides with less complex PTMs to protein-RNA cross-links. We show that the task of peptide retention time prediction can be transferred to the task of protein-RNA retention time prediction and identification rates are improved using a simple encoding with atomic compositions. For the more complex task of intensity prediction, we achieve similar improvements using standard methods and by focusing on more conserved fragmentation patterns.

Methods that incorporate retention time and intensity predictions in the rescoring process have a long history, with a trend toward more complex machine learning models [19, 20]. Common to these methods, the deviation between observation and predictions is used to guide the rescoring process and has been shown to increase the identification sensitivity and specificity [21, 22]. More recently, deep learning methods have been used to train models to predict the retention time and intensity of modified or unmodified peptides [23, 24]. An example of such a deep learning model is DeepLC which can accurately predict peptide retention times of unmodified and modified peptides, even for modifications not seen during training [25]. Intensity prediction, on the other hand, is the process of estimating the relative abundance or intensity of a peptide fragment in an MS experiment. MS2PIP is a computational tool that uses machine learning algorithms to predict peptide fragment intensities based on various peptide and spectral features [26].

Because protein-RNA cross-links lead to numerous RNA modifications (12–127, depending on the protocol in this study) diverse in physicochemical properties and structure, we expected challenges in the direct applicability of existing predictors or models trained only on unmodified peptides or peptides with classical PTMs. We, thus, set out to investigate if existing retention time and intensity prediction methods can be readily (or with minor modifications) applied to the domain of protein-RNA cross-linking to improve the identification rate of cross-linked peptides.

2 | MATERIALS AND METHODS

2.1 | Datasets

We collected 4-thiorudine (4SU) based protein-RNA crosslink data from the pRBSID experiment performed on human samples, as published by Bae et al. [5] (PXD023401). In addition, we collected 4SU protein-RNA crosslink data from the iTRAPP experiment of *Saccharomyces cerevisiae* samples, as published by Shchepachev et al. [14] (PXD011071). We added in-house generated datasets, employing conventional ultraviolet (UV) cross-linking and chemical cross-linking using NM and DEB on *Escherichia coli* samples.

2.2 | Cross-link identification data

We used the protein-RNA crosslinking search engine NuXL (v. 2023-02-01, publication in progress, see the Supplementary Material for a download link to the pre-release of the software; search engine parameters are given in Table S1) to generate cross-link identifications used in this study. To derive accurate retention times of cross-linked peptides, we provided high-confidence cross-link spectrum matches (CSMs, $q < 0.01$) as input to the OpenMS FeatureFinderIdentification algorithm [27]. The retention time at the apex of the extracted elution profiles was subsequently used as retention time used in the training (see Table S2 for details).

2.3 | Rescoring and evaluation

In all experiments, we used the percolator algorithm (version 3.05) for rescoring [28]. In all cases, we evaluated the impact of retention time and fragment intensity features on the number of identified CSMs at the CSM-level q-value. To ensure that additional features do not lead to overfitting (e.g., underestimate the true FDR based on the empirical target-decoy FDR [28, 29]), we used an entrapment experiment in which an equal number of entrapment proteins were added to the protein database (*E. coli* K12 + *Homo sapiens* UniProt sequences) [30]. The main difference to a standard entrapment experiment is that we focused on controlling the cross-link FDR and thus considered only CSMs. CSMs mapping to the entrapment database were considered false matches. The false match rate (FMR) among the set of reported target CSMs was compared to the empirically determined CSM-level FDR. To assess if the empirical FDR is valid and does not overestimate the true FDR, the FMR should be similar and not consistently larger than the corresponding empirical FDR [31].

2.4 | Retention time prediction

Currently, there is comparably little experimental data available that covers different protein-RNA crosslinking protocols, making it infeasible to train a deep learning model from scratch. However, in deep learning, “fine-tuning” is a common transfer learning technique used to improve the performance of a pre-trained model on a specific task or dataset [23, 20, 32]. The process of fine-tuning involves taking a pre-trained model, usually trained on a large and general dataset. The model is then trained further on the more specialized and smaller dataset for a specific task (typically with some slight modification of the training process) [20]. In this study, we investigate fine-tuning the DeepLC base model (originally trained on tryptic peptides) for prediction of retention time for protein-RNA crosslinking. The DeepLC (version 1.2.1) retention time predictor incorporates the atomic composition of the modification to train the model, to learn patterns that generalize to unseen modifications. We encoded protein-RNA cross-link modifications across all protocols (modifications per protocol UV: 12, DEB, NM: 14, 4SU: 127). The site of the modifications was encoded using the site localization as determined by the NuXL search engine. We used DeepLC as the base model for fine-tuning and froze all layers until the concatenate part, which extracts low-level features. Furthermore, we then train the model to learn the high-level specific features and patterns of newly observed cross-link modifications. To train the generic model, we split all datasets into independent train and test datasets (see Table S2 for details). For the protocol specific models, datasets were grouped according to the cross-linking protocol and split into independent train and test datasets. For both specific and generic models, train and test data was filtered to be composed of only non-redundant cross-links. Fine-tuning of generic and specific models was performed with DeepLCRetrainer (version 0.1.13, see Table S3 for hyperparameters and Figure S2 for results on the train/test split). We calculate retention time features (see Table S4 for full list) according

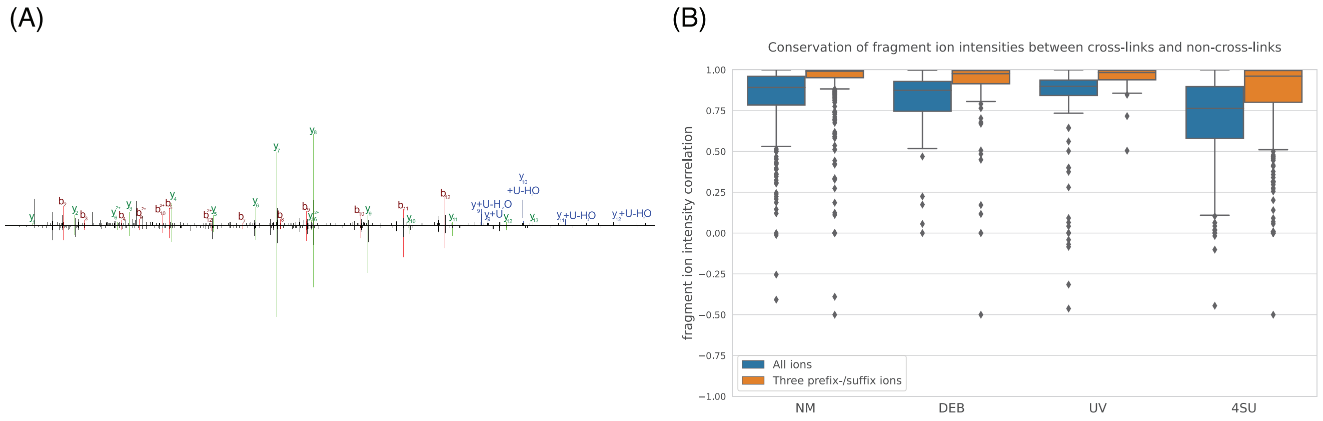

FIGURE 1 (A) Intensities of a crosslinked peptide (uridine cross-linked to VEGTEPTTAFNL**F**VGNLNFNK, cross-linked residue indicated in bold) mirrored on intensities of the same non-crosslinked peptide. (B) Maximum observed intensity correlation between prefix (b-ions) and suffix (y-ions) ions. For all protocols, considering only three prefix/suffix ions resulted in higher intensity correlations between non-cross-linked and cross-linked peptides as opposed to considering all prefix/suffix ion intensities in the correlation calculation.

to the formula used in MS2Rescore [21] using the predictions of our DeepLC models. These features form our retention time features used in rescoring with percolator.

2.5 | Intensity prediction

Currently, transfer learning of peptide intensities between unmodified and modified peptides is an active area of research. Due to the sparsity of training data and large number of modifications, we investigated an approach that leverages the conserved fragmentation behavior of cross-links. It has previously been observed that the intensities of the first few prefix ions (e.g., b-ions close to the N-terminus) and suffix ions (e.g., y-ions close to the C-terminus) of non-cross-linked peptides can be predicted with high accuracy and are discriminative for true and false hits [33]. To investigate if the relative prefix and suffix-ion intensities of cross-linked peptides and their non-cross-linked counterparts are well conserved (and thus lead to transferable intensities, see Figure 1A for one example), we calculated how well fragment ion intensities between peptides and their cross-linked counterpart correlate. We calculated the correlation considering all b- and y-ion intensities and the maximum correlation obtained if only intensities of the first three prefix and suffix ions are correlated. The total number of peptide pairs (peptides and their cross-linked counterpart were 4SU: 502, NM: 502, DEB: 96, UV: 116). The reasoning for choosing only the first three prefix/suffix ions is that the nucleotide is bound to the short prefix or the short suffix—or neither. Considering all prefix and suffix ions lead to lower correlations for the different protocols (see Figure 1B) and also (on average) slightly worse results if used in rescoring (see Figure S6) compared to using the shorter prefix/suffix-ions instead. We argue that these short prefixes or suffixes on average provide better conserved fragment intensities than considering all fragment ions. Based on the correlation analysis, we calculated our final intensity predictions features accordingly: the maximum correlation between observed and predicted first three prefix (b1-, b2-, and b3-ions) or suffix fragment ions (y1-, y2-, and y3-ions). While less frequently observed, b1-ions were included for the sake of consistency (e.g., they are also generated by the MS2PIP peak intensity prediction). The intensities for missing peaks were set to zero. To predict fragment intensities, we used the MS2PIP HCD XGBoost-based machine learning model which was trained on tryptic and non-tryptic (immunopeptides) HCD peptides [21].

3 | RESULTS

First, we investigated whether retention time prediction can be used to improve identification rates. We compared the performance of three different models: the baseline model trained solely on non-cross-linked peptides, a generic model fine-tuned on cross-links from all protocols, and models fine-tuned on specific protocols. In all cases, the generic and specific models performed better than the base model (see Table S5). Specifically, the protocol-specific models exhibited more accurate retention time predictions compared to the generic model in most protein-RNA crosslinking protocols (NM, DEB, and 4SU, see Figure S1 for a summary). Interestingly, the UV cross-linking protocol did not benefit from the UV specific model and performed better with the generic model. One explanation could be that UV cross-linking generates a more diverse set of fragment adducts compared to chemical and nucleotide analog cross-linking, and therefore benefits from the additional training data.

Next, we examined the individual impact of including retention time prediction and intensity prediction on the identification performance. The performance of the protocol-specific model and generic models is shown in Figure S5. For NM, DEB, and 4SU protocols the specific models performed better than the generic model (except in case of UV crosslinking). We selected the best models for final rescoring (see Figure 2 and Figure S3). The number of identifications at 1% CSM-level FDR for each rescoring experiment is provided in the figure's heading.

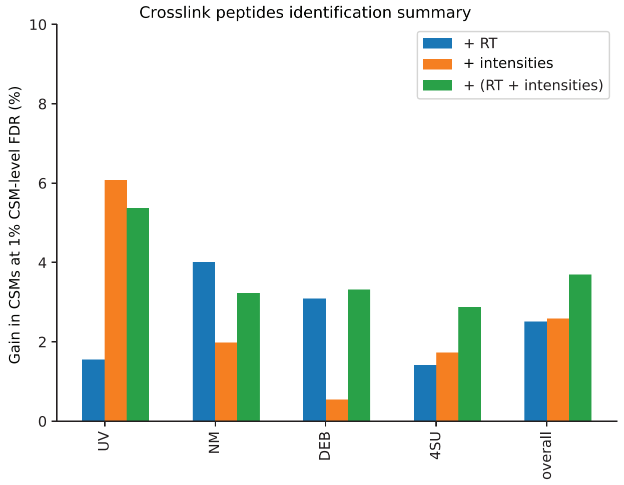

FIGURE 2 Relative improvement in identification performance for the four different protocols (at 1% cross-link spectrum match [CSM]-level FDR) if additional features (from retention time predictions: +RT, intensity predictions: +intensities, or both: RT + intensities) are added to the standard NuXL features prior to rescoring using percolator.

Our findings reveal that the inclusion of both retention time and intensity features is beneficial and increased the number of identified CSMs (avg. gain of +3.70%, in individual runs a max. gain of 10.47% for UV cross-linking, CSM *q*-value < 0.01) compared to running percolator solely on the default NuXL features (see Figure 2 and Figure S4). When comparing the effects of adding retention time predictions versus adding intensity predictions, we observed similar gains: including only retention time predictions resulted in on average 2.51% improvement in identification rates, while including only intensity prediction improved results in the average case by 2.59%. Our entrapment experiments, evaluated on the CSMs, did not indicate overfitting (see Figure 3 for one example). The effects of the additional features on identification rate and entrapment testing across all protocols are provided in Figure S3. Because the MS2Rescore features have been successfully used in the past to boost identification rates, we also investigated how the additional MS2Rescore features influence rescoring results when combined with the NuXL features (see Figure S7). Interestingly, the entrapment plot indicated that some overfitting exists (underestimation of the target-decoy FDR) when all MS2Rescore features are added (as opposed to adding only our retention time and intensity features). This might indicate that additional MS2Rescore features are not suitable for rescoring of protein-RNA crosslinks.

4 | DISCUSSION

While retention time and intensity prediction methods for unmodified peptides are well established, transferring these methods to modified peptides, especially to those carrying large and diverse adducts, is still a largely unsolved problem and an active area of research. A common impediment to these methods is the quality and quantity of training data available. Protein-RNA cross-linked peptides suffers from the same issue. It is difficult to obtain sufficient data to train predictive machine learning models on retention time and intensity features. In this study, we leveraged transfer learning to predict the retention time of cross-linked peptides and achieved a modest increase in the identification rate of protein-RNA cross-links. Further improvement in retention time prediction can be expected by including features that

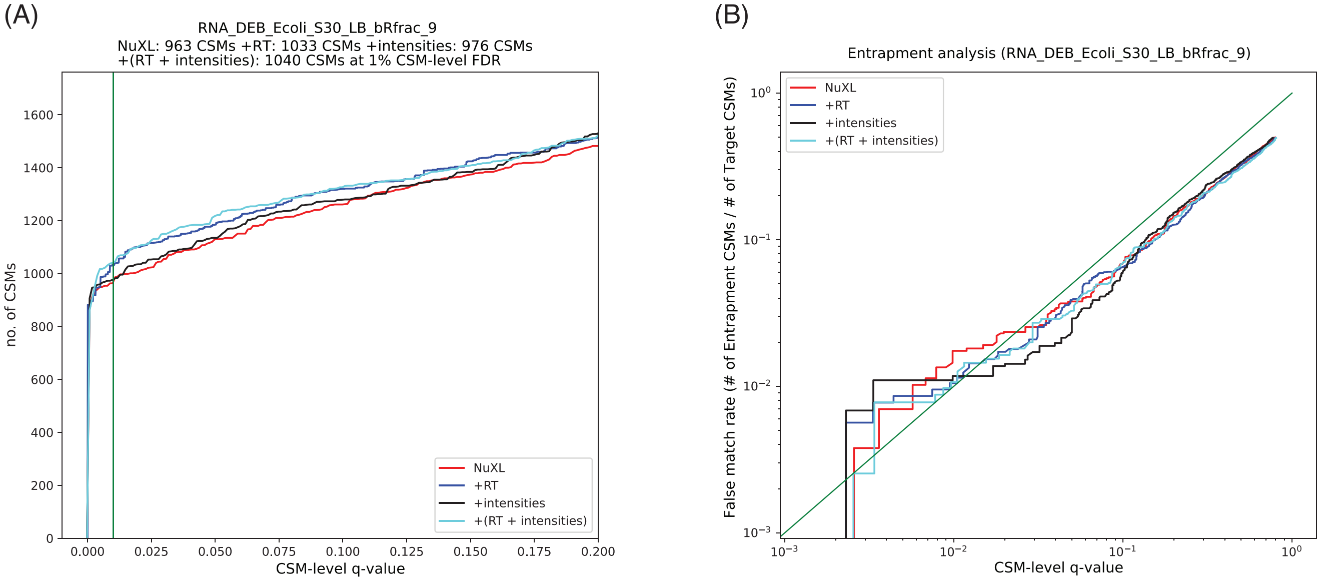

FIGURE 3 (A) Pseudo-ROC curves for the different set of features used in rescoring (here: diepoxybutane [DEB] protocol). The y-axis represents the number of cross-link spectrum matches (CSMs) identified at different CSM-level *q*-value thresholds. The green vertical line indicates the commonly used 0.01 CSM-level *q*-value threshold. (B) Comparison between false match rate (FMR) *q*-value (y-axis) and CSM-level *q*-value (x-axis) for CSM-level entrapment analysis (diagonal reference line indicates matching FMR and empirical CSM-level *q*-value).

expand over the simple encoding of the atomic composition. For the more challenging task of predicting the first prefix and suffix intensities of cross-linked peptides, we observed similar improvements in identification rates. Here, the dependency on the cross-linking protocol was more pronounced. Developing transfer learning methods that can cope with the large number of modifications for intensity predictors and including more cross-link data once they become available is a future research direction to further improve the identification rates of protein-RNA cross-linked peptides. Additional improvements could potentially also be obtained by including cross-link specific fragment ions into the spectrum prediction or fine-tuning process. A final note of caution based on our experiment with including all MS2Rescore features is that blindly including features can easily lead to overfitting—particularly for the large search space of cross-linked peptides. We thus want to underscore the importance of not solely relying on target-decoy-based FDR estimation but also to conduct entrapment experiments whenever new features are incorporated in a rescoring process that involves large search spaces.

ACKNOWLEDGMENTS

Arslan Siraj and Sven Degroeve are part of the MSCA-ITN-2020 PROTrEIN project, which received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No: 956148. Timo Sachsenberg is supported by the Ministry of Science, Research and Arts Baden-Württemberg. Henning Urlaub is supported by grants from the Deutsche Forschungsgemeinschaft (DFG; SFB 860, SFB 1565, SPP 1935). Arthur Declercq and Lennart Martens acknowledge funding from the Research Foundation Flanders (FWO) [12B7123N, G010023N, G028821N, 1SE3722]; Robbin Bouwmeester acknowledges funding from the Vlaams Agentschap Innoveren en Ondernemen [HBC.2020.2205]; Sven Degroeve and Lennart Martens acknowledge funding from the European Union's Horizon 2020 Programme (H2020-INFRAIA-2018-1) [823839]; Lennart Martens acknowledges funding from the Ghent University Concerted Research Action [BOF21/GOA/033].

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Scripts and associated data are available at https://github.com/Arslan-Siraj/NuXL_rescore.git and https://github.com/Arslan-Siraj/NuXL_rescore/releases/tag/v1.0.0. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD048145. Username: reviewer_pxd048145@ebi.ac.uk. Password: x9AvcJH2

ORCID

Arslan Siraj https://orcid.org/0009-0006-5748-0027
Lennart Martens https://orcid.org/0000-0003-4277-658X
Timo Sachsenberg https://orcid.org/0000-0002-2833-6070

REFERENCES

1. Hentze, M. W., Castello, A., Schwarzl, T., & Preiss, T. (2018). A brave new world of RNA-binding proteins. *Nature Reviews Molecular Cell Biology*, *19*(5), 327–341.
2. Ramanathan, M., Porter, D. F., & Khavari, P. A. (2019). Methods to study RNA–protein interactions. *Nature Methods*, *16*(3), 225–234.
3. Kelaini, S., Chan, C., Cornelius, V. A., & Margariti, A. (2021). RNA-binding proteins hold key roles in function, dysfunction, and disease. *Biology*, *10*(5), 366.
4. Götze, M., Sarnowski, C. P., de Vries, T., Knorlein, A., Allain, F. H. T., Hall, J., Aebersold, R., Leitner, A., & Leitner, A. (2021). Single nucleotide resolution RNA–protein cross-linking mass spectrometry: A simple extension of the CLIR-MS workflow. *Analytical Chemistry*, *93*(44), 14626–14634.
5. Bae, J. W., Kim, S., Narry Kim, V., & Kim, J. S. (2021). Photoactivatable ribonucleosides mark base-specific RNA-binding sites. *Nature Structural & Molecular Biology*, *12*(1), 6026.
6. Bae, J. W., Chul Kwon, S., Na, Y., Narry Kim, V., & Kim, J. S. (2020). Chemical RNA digestion enables robust RNA-binding site mapping at single amino acid resolution. *Nat Structural Molecular Biology*, *27*(7), 678–682.
7. Van Ende, R., Balzarini, S., & Geuten, K. (2020). Single and combined methods to specifically or bulk-purify RNA–protein complexes. *Biomolecules*, *10*(8), 1160.
8. Hafner, M., Katsantoni, M., Köster, T., Marks, J., Mukherjee, J., Staiger, D., Ule, J., & Zavolan, M. (2021). CLIP and complementary methods. *Nature Reviews Methods*, *1*(1), 20.
9. Urdaneta, E. C., & Beckmann, B. M. (2020). Fast and unbiased purification of RNA-protein complexes after UV cross-linking. *Methods (San Diego, Calif.)*, *178*, 72–82.
10. Sarnowski, C. P., Knörlein, A., De Vries, T., Götze, M., Beusch, I., Aebersold, R., Allain, F. H. T., Hall, J., & Leitner, A. (2022). Sensitive detection and structural characterisation of UV-induced cross-links in protein-RNA complexes using CLIR-MS. *bioRxiv*, 2022–2023.
11. Vieira-Vieira, C. H., & Selbach, M. (2021). Opportunities and challenges in global quantification of RNA-protein interaction via UV cross-linking. *Frontiers in Molecular Biosciences*, *8*, 669939.
12. McHugh, C. A., Russell, P., & Guttman, M. (2014). Methods for comprehensive experimental identification of RNA-protein interactions. *Genome Biology*, *15*, 1–10.
13. Götze, M., Sarnowski, C. P., de Vries, T., Knorlein, A., Allain, F. H. T., Hall, J., Aebersold, R., & Leitner, A. (2021). Single nucleotide resolution RNA–protein cross-linking mass spectrometry: A simple extension of the CLIR-MS workflow. *Analytical Chemistry*, *93*(44), 14626–14634.
14. Shchepachev, V., Bresson, S., Spanos, C., Petfalski, E., Fischer, L., Rappsilber, J., & Tollervey, D. (2019). Defining the RNA interactome by total RNA-associated protein purification. *Molecular Systems Biology*, *15*(4), e8689.
15. Kramer, K., Sachsenberg, T., Beckmann, B. M., Qamar, S., Boon, K. L., Hentze, M. W., Kohlbacher, O., & Urlaub, H. (2014). Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins. *Nature Methods*, *11*(10), 1064–1070.
16. Trendel, J., Schwarzl, T., Horos, R., Prakash, A., Bateman, A., Hentze, M. W., & Krijgsveld, J. (2019). The human RNA-binding proteome and its dynamics during translational arrest. *Cell*, *176*(1-2), 391–403.e19.
17. Fondrie, W. E., & Noble, W. S. (2021). mokapot: Fast and flexible semisupervised learning for peptide detection. *Journal of Proteome Research*, *20*(4), 1966–1971.
18. Granholm, V., Noble, W. S., & Käll, L. (2012). A cross-validation scheme for machine learning algorithms in shotgun proteomics. *BMC Bioinformatics*, *13*, 1–8.

19. Fedorova, E. S., Matyushin, D. D., Plyushchenko, I. V., Stavrianidi, A. N., & Buryak, A. K. (2022). Deep learning for retention time prediction in reversed-phase liquid chromatography. *Journal of Chromatography A*, *1664*, 462792.
20. Wen, B., Zeng, W. F., Liao, Y., Shi, Z., Savage, S. R., Jiang, W., & Zhang, B. (2020). Deep learning in proteomics. *Proteomics*, *20*(21-22), 1900335.
21. Declercq, A., Bouwmeester, R., Hirschler, A., Carapito, C., Degroeve, S., Martens, L., & Gabriels, R. (2022). MS2Rescore: Data-driven rescoring dramatically boosts immunopeptide identification rates. *Molecular & Cellular Proteomics: MCP*, *21*(8), 100266.
22. Gabriel, W., The, M., Zolg, D. P., Bayer, F. P., Shouman, O., Lautenbacher, L., Schnatbaum, K., Zerweck, J., Knaute, T., Delanghe, B., Huhmer, A., Wenschuh, H., Reimer, U., Médard, G., Kuster, B., Wilhelm, M., & Wilhelm, M. (2022). Prosit-TMT: Deep learning boosts identification of TMT-labeled peptides. *Analytical Chemistry*, *94*(20), 7181–7190.
23. Zeng, W. F., Zhou, X. X., Willems, S., Ammar, C., Wahle, M., Bludau, I., Voytik, E., Strauss, M. T., & Mann, M. (2022). AlphaPeptDeep: A modular deep learning framework to predict peptide properties for proteomics. *Nature Communications*, *13*(1), 7238.
24. Tarn, C., & Zeng, W. F. (2021). pDeep3: Toward more accurate spectrum prediction with fast few-shot learning. *Analytical Chemistry*, *93*(14), 5815–5822.
25. Bouwmeester, R., Gabriels, R., Hulstaert, N., Martens, L., & Degroeve, S. (2021). DeepLC can predict retention times for peptides that carry as-yet unseen modifications. *Nature Methods*, *18*(11), 1363–1369.
26. Degroeve, S., & Martens, L. (2013). MS2PIP: A tool for MS/MS peak intensity prediction. *Bioinformatics*, *29*(24), 3199–3203.
27. Weisser, H., & Choudhary, J. S. (2017). Targeted feature detection for data-dependent shotgun proteomics. *Journal of Proteome Research*, *16*(8), 2964–2974.
28. The, M., MacCoss, M. J., Noble, W. S., & Käll, L. (2016). Fast and accurate protein false discovery rates on large-scale proteomics data sets with percolator 3.0. *Journal of the American Society for Mass Spectrometry*, *27*, 1719–1727.
29. Choi, H., & Nesvizhskii, A. I. (2008). False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *Journal of Proteome Research*, *7*(01), 47–50.
30. Feng, X. D., Li, L. W., Zhang, J. H., Zhu, Y. P., Chang, C., Shu, K. X., & Ma, J. (2017). Using the entrapment sequence method as a standard to evaluate key steps of proteomics data analysis process. *BMC Genomics*, *18*(2), 1–9.
31. Lin, A., Short, T., Noble, W. S., & Keich, U. (2022). Improving peptide-level mass spectrometry analysis via double competition. *Journal of Proteome Research*, *21*(10), 2412–2420.
32. Meyer, J. G. (2021). Deep learning neural network tools for proteomics. *Cell Reports Methods*, *1*(2), 100003.
33. Degroeve, S., Gabriels, R., Velghe, K., Bouwmeester, R., Tichshenko, N., & Martens, L. (2021). ionbot: A novel, innovative and sensitive machine learning approach to LC-MS/MS peptide identification. *bioRxiv*, 2021–2027.

SUPPORTING INFORMATION

Additional supporting information may be found online https://doi.org/10.1002/pmic.202300144 in the Supporting Information section at the end of the article.

How to cite this article: Siraj, A., Bouwmeester, R., Declercq, A., Welp, L., Chernev, A., Wulf, A., Urlaub, H., Martens, L., Degroeve, S., Kohlbacher, O., & Sachsenberg, T. (2024). Intensity and retention time prediction improves the rescoring of protein-nucleic acid cross-links. *Proteomics*, *24*, e2300144. https://doi.org/10.1002/pmic.202300144