





XL-MSDigger: a deep learning-based, versatile solution for cross-linking mass spectrometry

Received: 11 April 2025

Accepted: 3 February 2026

Published online: 10 February 2026

 Check for updatesMoran Chen ^{1,2,3,4}, Yanhong Hao^{1,4}, Xiao Huang¹, Pengfei Wu ¹, Jian Sun¹, Bing Zhang ^{2,3} & Suming Chen ¹✉

Cross-linking mass spectrometry (XL-MS) technology plays an important role in protein structure and protein-protein interaction (PPI) studies. However, there are serious shortcomings in the data processing and analysis methods of XL-MS that limit its ability to perform in-depth analysis. Here, we built a deep learning-based XL-MS versatile platform, XL-MSDigger, to address the bottlenecks of XL-MS technology. The foundation of this platform is a deep learning-based multidimensional information prediction tool, Deep4D-XL, for cross-linked peptides. This tool is capable of accurately predicting retention time, collisional cross-section, and fragment ion intensity information of cross-linked peptides. Through the multidimensional information prediction, we develop rescoring algorithms and workflows for DDA and DIA analysis data, which improve the coverage of DDA-based XL-MS identification, and realize the evaluation of false discovery rate and high-reliability identification of DIA-based XL-MS analysis. In addition, XL-MSDigger enables DIA-based XL-MS analysis using predicted spectral libraries, allowing improved detection of protein-protein interactions in targeted or moderate-scale interaction analyses. We believe that the XL-MSDigger platform can be widely used as a general solution to enhance the performance of XL-MS analysis.

The structure of proteins and their interactions influence their properties and functions^{1,2}. In recent years, cross-linking mass spectrometry (XL-MS) has evolved into a crucial tool for studying protein structures and protein-protein interactions³⁻⁵. XL-MS employs chemical cross-linkers to covalently bond amino acid residues within a certain spatial range, thereby capturing the conformation and interaction interfaces of proteins in their native environments^{6,7}.

The primary targets of XL-MS experiments are cross-linked peptides, composed of two peptides linked by a cross-linker. These peptides exhibit more complex structures and mass spectrometric fragmentation behaviors compared to linear peptides in typical proteomics experiments. The resolving of MS data for these peptides requires consideration of a greater variety of fragment ion types. Due

to the unique structure of cross-linked peptides, database searches used in proteomics face higher computational complexity, often leading to vast search spaces and an elevated risk of false positives. As research progresses, various specialized cross-linked peptide analysis software have been developed, including StavroX⁸, Kojak⁹, Xilmass¹⁰, Merox¹¹, Xolik¹², pLink2¹³ and Scout¹⁴, among others. For instance, pLink2 utilizes a fragment index and a two-step open search strategy to perform proteomic-level cross-linked MS analysis within a shorter timeframe. However, almost all such software relies only on the m/z information of precursor and fragment ions, which may increase the likelihood of random false matches in peptide-spectrum matches (PSMs) when faced with complex XL-MS data and enormous search spaces.

¹The Institute for Advanced Studies, Wuhan University, Wuhan, Hubei, China. ²Lester and Sue Smith Breast Center, Baylor College of Medicine, Houston, Texas, USA. ³Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA. ⁴These authors contributed equally: Moran Chen, Yanhong Hao. ✉e-mail: sm.chen@whu.edu.cn

Recently, post-processing tools like MSBooster¹⁵ and DeepRescore¹⁶ have enhanced the sensitivity of PSM identification in regular proteomics by using deep learning-based prediction of fragment ion intensity, retention time (RT), and collisional cross-section (CCS) in PSM rescoring. The core of these tools is the ability to accurately predict multidimensional peptide information. Since the development of the pDeep¹⁷, researchers have successively developed various models capable of predicting properties for regular peptides, such as Prosit¹⁸, DeepMass¹⁹, DeepDIA²⁰ and Deep4D²¹ for fragment ion intensity prediction, and DeepRT²², DeepLC²³, and AutoRT²⁴ for RT prediction. Mann and colleagues have also developed a model for predicting the CCS of peptides²⁵. However, the deep learning-based prediction model for cross-linked peptides is very rare due to their structural complexity. Currently, only the xiRT²⁶ for predicting the RT and pDeepXL²⁷ for predicting fragment ion intensity are available. To date, a deep learning-based model capable of predicting multidimensional information for cross-linked peptides is lacking.

On the other hand, most XL-MS analyses are currently performed in data-dependent acquisition (DDA) mode. Compared to DDA, data-independent acquisition (DIA) techniques, which fragment precursors in an unbiased manner across predefined windows, can enhance precursor coverage and repeatability²⁸. DIA has become a mainstream method in proteomics and has demonstrated high identification performance, but DIA-based XL-MS methods are just beginning to emerge. We developed 4D-diaXLMS²⁹, a DIA-based workflow that supports proteome-wide cross-linking analysis. However, current methods still rely on sample-specific experimental spectral libraries derived from DDA experiments, primarily leveraging DIA's quantitative capabilities rather than maximizing its advantages in identification coverage^{29,30}. DIA-based regular proteomics has demonstrated that the use of predicted peptide libraries derived from the species' proteome can effectively eliminate the need for time-consuming experimental library construction and expand the identification coverage^{20,21}. Nevertheless, the DIA analysis for cross-linked proteomics based on predicted spectral libraries is still restricted by the tools capable of predicting multidimensional information specific to cross-linked peptides.

In this study, we aim to develop a comprehensive deep learning-based solution to address the key issues currently facing XL-MS analysis, including deep mining of DDA data, FDR control for DIA analysis, and DIA analysis based on predictive spectral libraries. We develop a deep learning-based model, Deep4D-XL, which can accurately predict multidimensional information, including RT, CCS, and fragment ion intensity of cross-linked peptides. A versatile workflow XL-MSDigger is further developed for in-depth XL-MS analysis, supporting both DDA- and DIA-based methods. Compared to existing XL-MS technologies, XL-MSDigger offers several key advancements: 1) by utilizing multidimensional information predicted by Deep4D-XL and a deep neural network-based rescoring model, it can comprehensively extract more information on inter-protein cross-linked peptides and protein-protein interactions (PPI) from DDA-based XL-MS data; 2) it establishes a systematic framework for FDR evaluation in DIA XL-MS analysis and reduces the elevated FDR observed in current DIA XL-MS workflows through optimized rescoring; and 3) it enables DIA-based XL-MS analysis using predicted spectral libraries, thereby enhancing the identification depth of both intra- and inter-protein cross-linked peptides.

Results

Construction of Deep4D-XL model

Deep4D is a deep learning model with robust predictive performance for the multidimensional information (RT, CCS, and fragment ion intensity) of linear peptides and phosphorylated peptides. However, due to significant structural differences between cross-linked peptides and linear peptides, the Deep4D model could not be directly applied to cross-linked peptides. Therefore, we have specifically developed a dedicated model for cross-linked peptides based on the foundational framework of Deep4D, named Deep4D-XL. In this model (Fig. 1a), we employed a Siamese network architecture³¹ in the encoder module to characterize the information of the two linear peptides within the cross-linked peptides, followed by a cross-attention module³² that merges the feature representations of both peptides. Finally, the decoder module, consisting of a convolutional neural network (CNN) and fully connected layers, converts the tensor output from the cross-attention module into RT, CCS and fragment ion intensity.

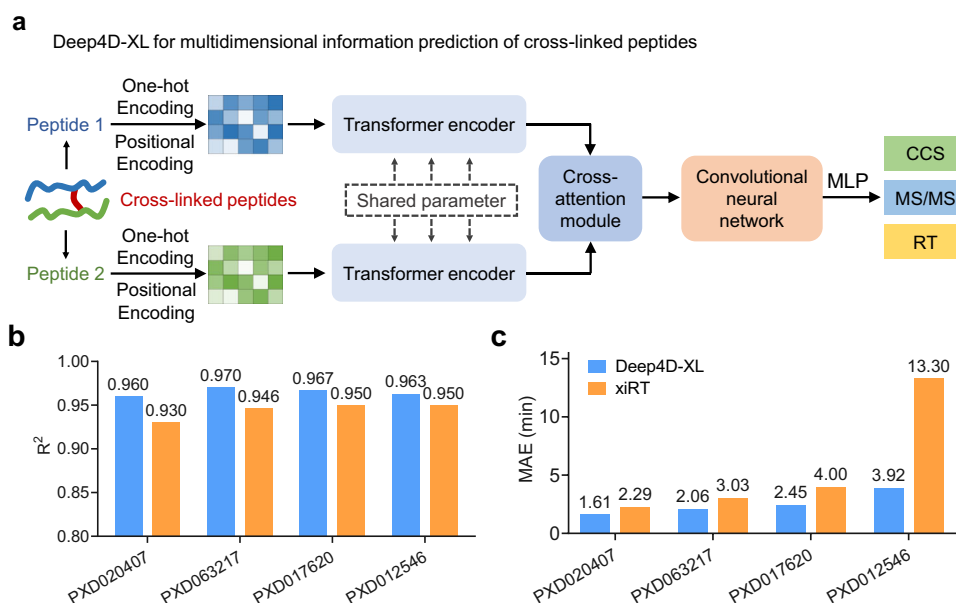


Fig. 1 | Framework and performance comparison of the Deep4D-XL model. **a** Schematic diagram of the Deep4D-XL model framework. **b, c** Comparison of (b) mean absolute error (MAE) and (c) coefficient of determination (R^2) values between

the Deep4D-XL and xiRT models on four datasets. Source data are provided as a Source Data file.

To evaluate the effectiveness of the Deep4D-XL model, we first compared its performance with xiRT in predicting the RT of cross-linked peptides. Specifically, we trained Deep4D-XL on the same published dataset used in the xiRT²⁶ and evaluated its performance against the results reported for xiRT (Supplementary Data 1). Deep4D-XL achieved an R^2 of 0.96 and a median absolute error (MAE) of 1.61 min (Supplementary Fig. 1a), outperforming xiRT, which yielded a lower R^2 of 0.93 and a higher MAE of 2.29 min on the same dataset (Fig. 1b, c). In addition, we trained and tested both xiRT and Deep4D-XL on three other independent retention time datasets. Across all datasets, Deep4D-XL consistently exhibited lower prediction errors and higher coefficients of determination than xiRT. Specifically, Deep4D-XL achieved MAE of 2.06, 2.45, and 3.92 min with corresponding R^2 of 0.970, 0.967, and 0.963, respectively (Supplementary Fig. 1b–d), whereas xiRT showed higher MAE (3.03, 4.00, and 13.30 min) and lower R^2 (0.946, 0.950 and 0.950) on the same datasets (Fig. 1b, c). These results demonstrate that Deep4D-XL provides more accurate and stable RT predictions across diverse experimental settings, motivating us to further explore the potential of the Deep4D-XL model in predicting additional properties of cross-linked peptides.

Dataset construction and multidimensional prediction performance for cross-linked peptides

To enable the Deep4D-XL model to predict additional properties of cross-linked peptides, we first constructed a large-scale, high-quality dataset for training and testing, as no such dataset currently exists. We conducted a deep fractionation of cross-linked peptide samples (40 fractions) from the HeLa cell proteome. These samples were then analyzed using ddaPASEF³³ mode on a timsTOF pro mass spectrometer, and the resulting data were processed with pLink2. In total, 90,116 cross-linked PSMs were identified across all sample fractions (Supplementary Fig. 2). For each identified peptide, we retained the highest-scoring PSM in each observed charge state to extract complete RT, CCS, and MS/MS spectral information. This resulted in 34,115 charge-specific unique cross-linked peptides, thereby generating a large-scale four-dimensional (4D) dataset of cross-linked peptides (Fig. 2a). The theoretical spatial distances of intra-protein cross-linked peptides in the steric structures of proteins were calculated within our dataset. The results showed that, of the 12,580 intra-protein cross-linked peptides that could be matched to protein structures, 9841 peptides (78.2%) had amino acid residue distances within the range constrained by the disuccinimidyl suberate (DSS) cross-linker (Supplementary Fig. 3), indicating the high reliability of this XL-MS dataset.

Next, we randomly divided the dataset into training, validation, and test sets in a ratio of 8.1:0.9:1, and then trained and tested the corresponding models for predicting RT, CCS, and fragment ion intensity. For RT prediction, the trained model demonstrated accurate prediction on the test set (Supplementary Data 1), with a mean absolute error of 2.06 min (Fig. 2b), approaching the performance of the Deep4D model for linear peptides²¹. For CCS prediction, the trained model achieved a median relative error (MRE) of 1.52% on the test set (Fig. 2c), with accuracy nearly comparable to that for linear peptides. This result demonstrates that Deep4D-XL enables accurate CCS prediction for cross-linked peptides using a deep learning framework (Supplementary Data 1).

For the prediction of fragment ion intensity, the complex and varied fragment ion types of cross-linked peptides pose significant challenges. In addition to considering the b and y ions of each linear peptide, it is necessary to consider fragments containing the cross-linked site (see “Methods” section for details). Despite this complexity, the trained model performed well on the test set, achieving a median dot product of 0.88 between predicted and experimental spectra (Supplementary Data 1). In addition, we separately calculated the dot products between the predicted and experimental spectra for fragment ions originating from each of the two peptides, as well as for

fragment ions containing the cross-linking site (Fig. 2d). Figure 2e, f show the prediction performance using two cross-linked peptides as examples. The results show that the predicted spectra for both cross-linked peptides are very similar to their corresponding experimental spectra, again demonstrating the reliability of the prediction.

To demonstrate the broad applicability of the Deep4D-XL model, we further evaluated its generalization across different instruments and cross-linker types. From an Orbitrap-based DSS-crosslinked dataset³⁴, comprising 14,300 cross-linked PSMs, the model was trained and tested, achieving a median dot product of 0.844 (Supplementary Fig. 1e) between predicted and experimental spectra, indicating broad adaptability to different instrument platforms. We also evaluated its performance on other non-cleavable cross-linkers. From an Orbitrap-based PhoX-cross-linked dataset³⁵ comprising 22,306 cross-linked PSMs, the model achieved a median dot product of 0.917 (Supplementary Fig. 1e) between predicted and experimental spectra. To further enhance versatility, we also analyzed an Orbitrap-based cleavable DSBU-crosslinked dataset³⁶, which yielded 20,446 cross-linked PSMs. To accommodate the unique fragmentation characteristics of cleavable cross-linkers, the output dimension of the Deep4D-XL model was adjusted accordingly. The trained model achieved a median dot product of 0.853 on the test set (Supplementary Fig. 1e).

Overview of XL-MSDigger

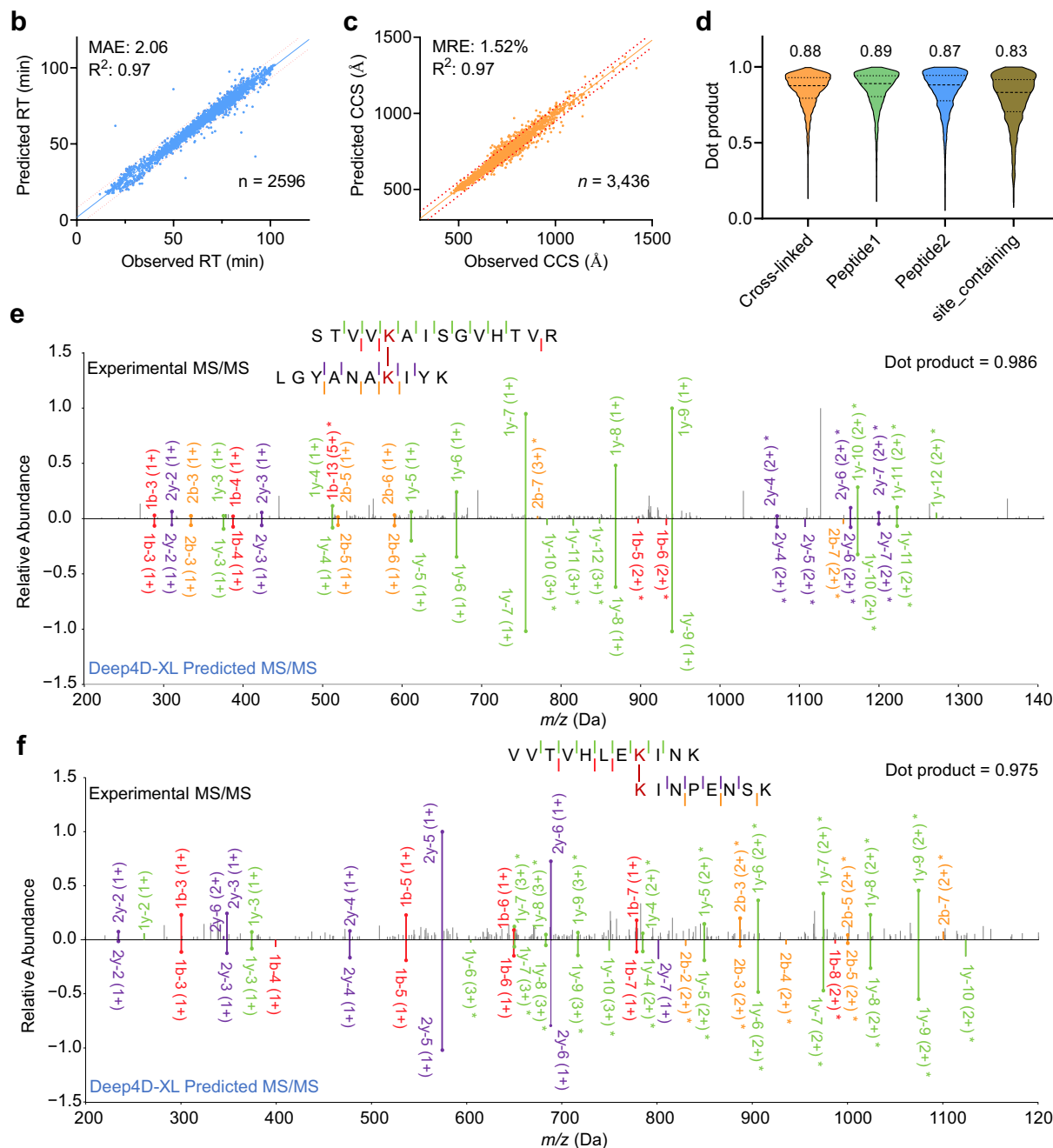
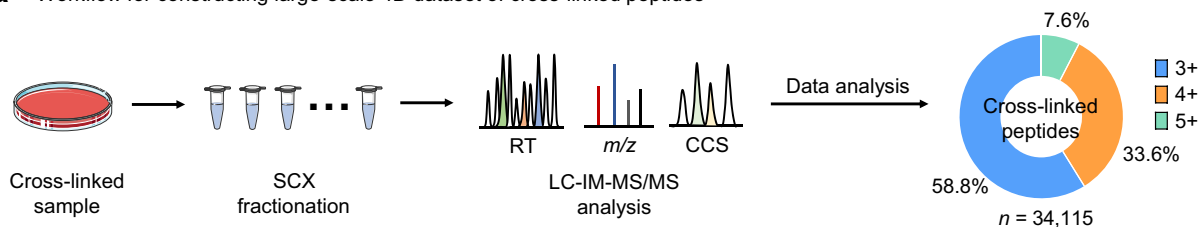
After developing and validating the Deep4D-XL model, we built a versatile computational pipeline for in-depth analysis and mining of XL-MS data, called XL-MSDigger, centered on that model. The main functions of XL-MSDigger are shown in Fig. 3, which includes two primary functional modules for analyzing both DDA and DIA XL-MS data. In DDA data analysis, the workflow begins with the use of the pLink2 software to perform a database search of DDA data to obtain preliminary results. Subsequently, the Deep4D-XL model performs multidimensional predictions for all identified cross-linked peptides, which were used to compute multiple features for each PSM, including Δ RT, Δ CCS, MS/MS similarity and other information. These features are then utilized by a deep neural network to rescore cross-linked peptide results, enhancing the identification of cross-linked peptides. The experimental spectral library generated from DDA results can also be used for subsequent DIA analysis.

In DIA data analysis, a decoy spectral library is first generated based on the target spectral library, and the two libraries are merged. The merged library is then analyzed using DIA-NN³⁷ software to perform initial DIA analysis. Subsequent rescoring of the DIA-NN search results is conducted using a deep neural network, followed by FDR evaluation to obtain the final identification results (see “Methods” for details). For spectral library construction, predicted cross-linked spectral libraries can also be generated based on the Deep4D-XL model, which enhances the identification scope of DIA analysis. The specific performance of each function of XL-MSDigger are detailed in the following results.

Rescoring of DDA-based XL-MS data to improve the coverage of inter-protein crosslinks

The rescoring method for DDA-based XL-MS data in XL-MSDigger includes two main steps: 1) feature extraction for cross-linked peptides based on the Deep4D-XL model; 2) rescoring based on a deep neural network.

Cross-linked peptides are categorized into two types: intra-protein and inter-protein cross-links. In the target-decoy approach, due to the larger search space for inter-protein cross-linked peptides, most of the matched decoy PSMs are inter-protein cross-links, which can negatively impact the identification of target inter-protein cross-links during FDR evaluation. Additionally, inter-protein cross-links are usually less abundant and more challenging to identify, often resulting

a Workflow for constructing large-scale 4D dataset of cross-linked peptides

in very low identification rates. Therefore, we aimed to explore whether the rescoring method based on Deep4D-XL could help improve the identification of inter-protein cross-linked peptides. Here, we used Deep4D-XL to predict RT, CCS, and fragment ion intensity information for all candidate peptides (including both target and decoy cross-linked peptides) found in the search results, extracting 12 features related to them. Together with the four original search result features,

a total of 16 features were used as input for the rescoring model (Supplementary Table 1). In choosing the rescoring model, we evaluated the performance of support vector machines (SVM)³⁸ and deep neural networks (DNN)³⁹. pLink2 continued to be used as the search software for DDA-based XL-MS analysis, and DDA data from seven yeast cross-link samples were analyzed. The results show that pLink2 identified 461 inter-protein cross-linked PSMs across the seven

Fig. 2 | Construction of the large-scale cross-linked peptide 4D information dataset and the performance of the Deep4D-XL model. **a** Workflow for constructing a large-scale cross-linked peptide dataset, including cross-linking and enzymatic digestion of protein samples, fractionation of peptide samples, ion mobility MS analysis, and MS data processing. **b** Predictive performance of the Deep4D-XL model on the RT test set for cross-linked peptides. The red dashed lines represent the 95% prediction bands. Model performance is summarized by the mean absolute error (MAE) and the coefficient of determination (R^2). **c** Predictive performance of the Deep4D-XL model on the CCS test set for cross-linked peptides. The red dashed lines represent the 95% prediction bands. Model performance is summarized by the median relative error (MRE) and the coefficient of

determination (R^2). **d** Dot products between the predicted and experimental spectra for fragment ions originating from cross-linked peptides, each of the two peptides, as well as for fragment ions containing the cross-linking site. The dotted lines within each violin plot represent the median and the first and third quartiles of the data distribution. **e, f** Comparison of experimental MS/MS with predicted MS/MS for cross-linked peptides, (**e**) lysine with the cross-linking site located in the middle of two peptides, and (**f**) lysine with one of the sites located at the end of one peptide. The upper numbers 1 and 2 in the fragment ion labels indicate the two peptides that were cross-linked. An asterisk (*) indicates that the fragment contains the cross-linking site. Source data are provided as a Source Data file.

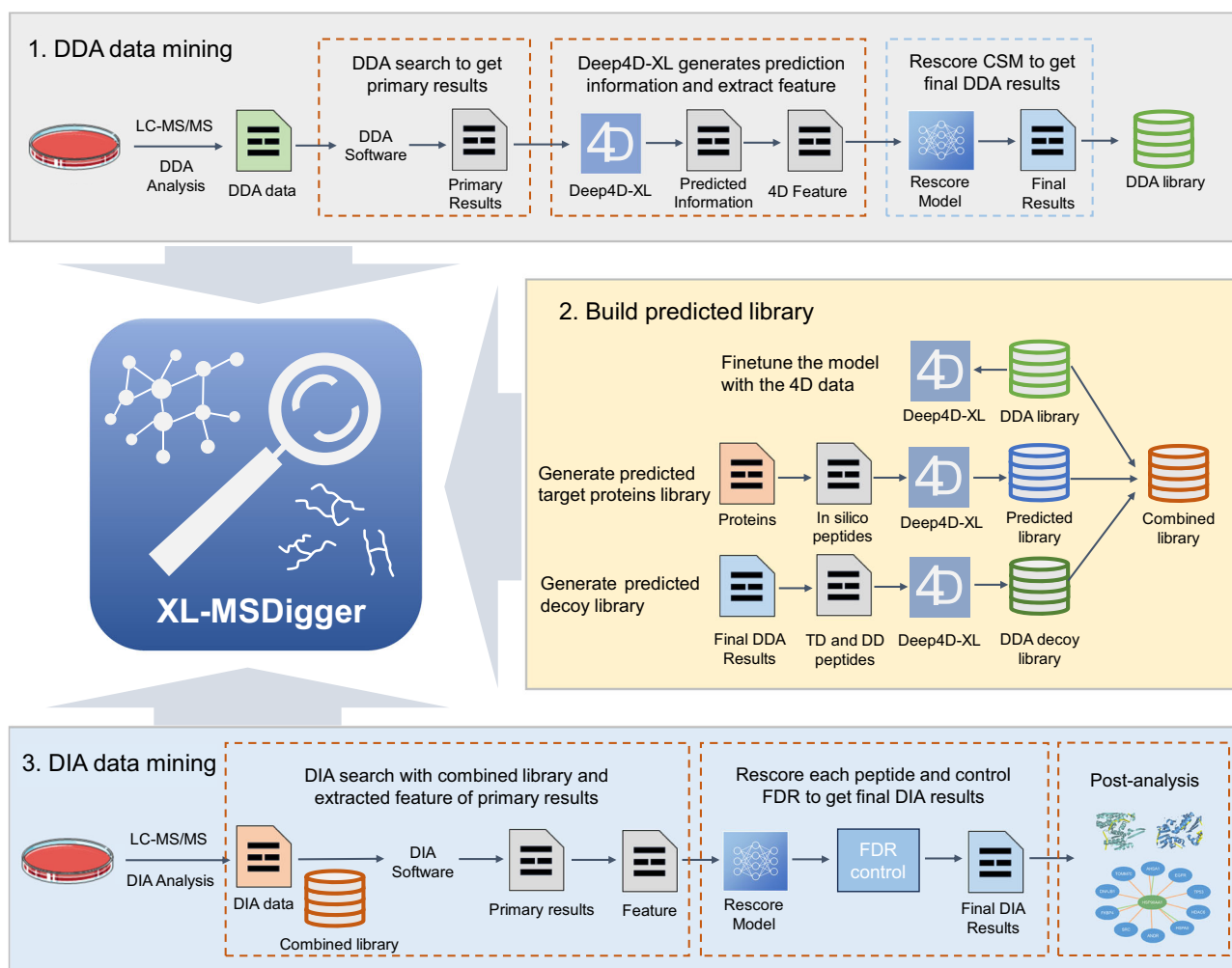


Fig. 3 | Functional modules and workflow of XL-MSDigger. This includes several functional modules such as rescoring of DDA-based XL-MS analysis, FDR assessment of DIA-based XL-MS analysis results, rescoring of DIA-based XL-MS analysis, and 4D DIA-based XL-MS analysis based on a predicted cross-linked spectral library.

samples (Supplementary Fig. 4), while the proposed SVM and DNN rescoring results included 852 and 956 inter-protein cross-linked PSMs, respectively, representing an increase of 85% and 107% in identification quantity compared to the original pLink2 results. These preliminary results demonstrate that the rescoring method based on predictive information from the Deep4D-XL model can enhance the identification of target cross-linked PSMs.

To further validate the effectiveness of the rescoring method in DDA-based XL-MS analysis, we applied it to the aforementioned 40 HeLa cross-linked proteome fraction datasets. The results show that after rescoring, the identification of inter-protein cross-linked PSMs in 37 out of the 40 datasets exceeded the number originally identified by the pLink2 software, with only three datasets showing a slight decrease in the number of inter-protein cross-linked PSMs identified after

rescoring (Fig. 4a). The total number of inter-protein cross-linked PSMs identified by the rescoring was 106.9% higher (2326 *vs.* 1224) compared to the original results from pLink2 software results (Fig. 4b), and the corresponding protein-protein interactions (PPIs) also increased by 76.3% (610 *vs.* 346) (Supplementary Data 2).

Furthermore, under the assumption that the target PSMs meeting the FDR threshold in the original search results were authentic, the objective of rescoring was twofold: to mine as many target inter-protein cross-linked PSMs as possible from those not meeting the original FDR and to retain as many PSMs as possible that met the FDR restrictions initially. Therefore, we defined sensitivity to indicate the proportion of original FDR-compliant PSMs included in the rescoring results relative to all originally FDR-compliant PSMs. Our findings revealed that 1140 of the 1224 inter-protein cross-linked PSMs

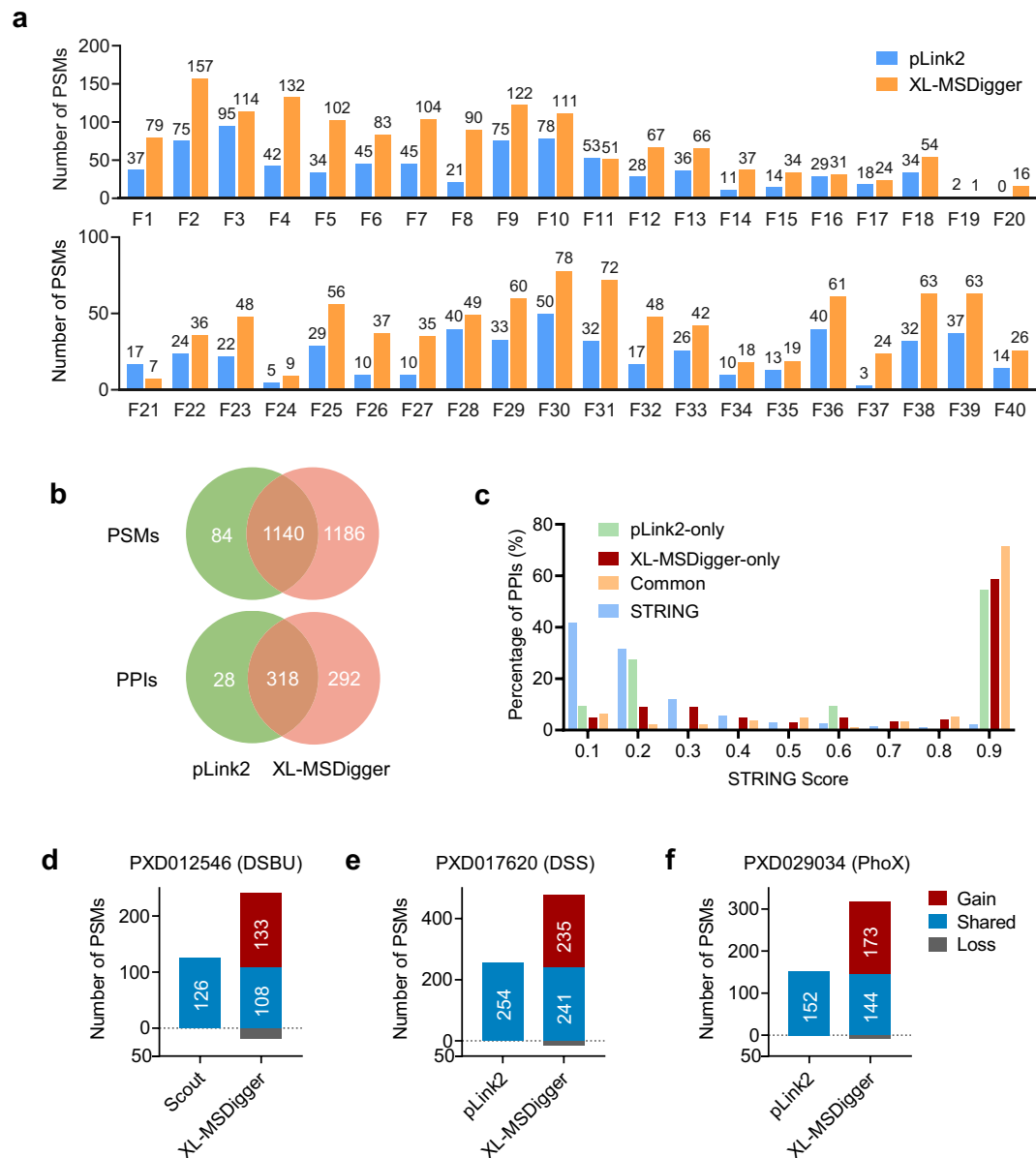


Fig. 4 | Performance of the rescoring method for DDA-based XL-MS analysis in XL-MSDigger. **a** Comparison of inter-protein cross-linked PSM identification counts between pLink2 and XL-MSDigger for 40 HeLa cross-linked proteome fraction samples. **b** Venn diagram of inter-protein cross-linked PSMs and PPIs identified in the results of pLink2 and XL-MSDigger. **c** Distribution of confidence

scores for PPIs uniquely or jointly identified by pLink2 and XL-MSDigger in the STRING database. **d–f** The number of PSMs identified (**d**) by Scout and XL-MSDigger on the PXD012546 (DSBU) dataset, (**e**) by pLink2 and XL-MSDigger on the PXD017620 (DSS) dataset, and (**f**) by pLink2 and XL-MSDigger on the PXD029034 (PhoX) dataset. Source data are provided as a Source Data file.

identified in the original pLink2 results appeared in the rescoring results (Fig. 4b, Supplementary Fig. 5), achieving a sensitivity of 93.1%. Additionally, the sensitivity of the identified PPIs before and after rescoring was similarly high at 88.8% (Fig. 4b). These results demonstrate the reliability of our rescoring algorithm.

We further incorporated the *E. coli* proteome as an entrapment database to provide an additional layer of quality assessment for the identification results. The proportion of inter-protein cross-linked PSMs involving *E. coli* proteins was 1.0% before rescoring and 0.97% after rescoring (Supplementary Data 2), indicating that the quality control of our rescoring approach is comparable to that of pLink2.

Subsequently, we analyzed the confidence scores of all protein–protein interactions (PPIs) in the initial and rescoring results using the STRING⁴⁰ database (Supplementary Data 2, Supplementary Fig. 6). The results show that XL-MSDigger not only outperforms in the

number of identified interactions but also improves result credibility. Among PPIs uniquely identified by XL-MSDigger, 49% lack STRING scores, which is lower than the 61% observed for PPIs uniquely identified by pLink2 (Supplementary Fig. 6a). This indicates our method prioritizes interactions supported by known databases. These PPIs, currently unsupported by STRING data, may include previously unreported discoveries with potential biological significance or interactions within database coverage blind spots. In addition, among uniquely identified PPIs with available STRING annotations, 78% of those recovered by XL-MSDigger have STRING scores above 0.4, compared to only 64% for pLink2 (Fig. 4c, Supplementary Fig. 6b, c). This further demonstrates the ability of the proposed method to uncover more high-confidence results among well-documented interactions.

To evaluate the contribution of individual subscores to the overall rescoring performance, we conducted ablation experiments on 40 DDA datasets by sequentially removing the RT, spectral similarity, and CCS features. The results showed that excluding RT, spectral similarity, and CCS reduced the number of identified PSMs to 74%, 86%, and 99.6% of the full rescoring results, respectively (Supplementary Fig. 7). These findings indicate that RT contributes the most to model performance, followed by spectral similarity, while the CCS feature has only a minimal effect. This observation is consistent with the results reported in the MSBooster study, which demonstrated that CCS values are highly correlated with precursor m/z and charge, leading to nearly identical CCS distributions between target and decoy peptides at the current ion mobility resolution.

Furthermore, to assess the applicability of XL-MSDigger to different mass spectrometry types and cross-linkers, we applied our rescoring method to the previously mentioned Orbitrap-based DSS-, PhoX-, and DSBU-cross-linked datasets (Supplementary Data 5). Specifically, we analyzed subsets of data files from these datasets using pLink2 (for DSS and PhoX) and Scout (for DSBU). For the DSBU-cross-linked dataset, the number increased from 126 to 241 (Fig. 4d); for the DSS-cross-linked dataset, the number of inter-protein cross-linked PSMs increased from 254 to 476 (Fig. 4e); and for the PhoX-cross-linked dataset, the number increased from 152 to 317 (Fig. 4f). These results demonstrate the good versatility and applicability of XL-MSDigger across different experimental settings.

FDR control for DIA-based XL-MS analysis by the rescoring algorithm improves identification reliability

For low-abundance cross-linked peptides, the semi-stochastic selection of the precursor ions in the DDA method may result in the loss of MS/MS information, leading to poor reproducibility in identification. In contrast, the DIA method with unbiased fragmentation of precursors can effectively address this issue. However, a primary challenge facing current DIA-based XL-MS analysis is the lack of a specific FDR assessment method for cross-linked peptides⁴¹. This is because current DIA-based proteomics methods typically rely on decoy spectral libraries for FDR evaluation, yet there are no tools available for constructing decoy cross-linked peptide spectral libraries. Both our previously published 4D-diaXIMS method and the method⁴² by Rappsilber et al. are DIA-based approaches based on sample-specific DDA experimental spectral libraries, which primarily leverage the quantitative capabilities of the DIA-based analysis. In addition, due to the absence of an FDR evaluation method for cross-linked peptides, the authenticity of DIA results cannot be guaranteed once nonspecific cross-linked peptide information is introduced into the spectral library.

Based on the Deep4D-XL model, an FDR evaluation module was added in XL-MSDigger. The Deep4D-XL model was used to predict the multidimensional information of all decoy cross-linked peptides to construct a decoy spectral library, which is then merged with the target experimental spectral library to form an integrated spectral library (Supplementary Fig. 8). A library search is then performed based on the library, and the FDR is calculated based on the number of target and decoy peptides in the search results. This approach was further validated by evaluating diaPASEF⁴³ data from 20 cross-linked samples using the 4D-diaXIMS method. First, all identification results were ranked based on the score output by DIA-NN, and the false discovery rate (FDR) was calculated. Using an FDR threshold of 5%, a total of 5379 target cross-linked precursors were identified across 20 DIA datasets. However, analysis of the DIA-NN score distributions for target and decoy precursors revealed poor separation between the two groups (Fig. 5b). Most of the decoys were of the TD type, which contain one target linear peptide. Because Deep4D-XL can generate fragment ions identical to those of true cross-linked peptides, these ions may coincidentally match experimental spectra, leading existing scoring algorithms to assign spuriously high scores to decoy cross-linked peptides

and thereby inflate the FDR. Addressing this issue will require the development of scoring algorithms specifically optimized for cross-linked peptides to achieve more accurate FDR control in DIA-based XL-MS analysis.

To address the limitations of existing scoring algorithms for DIA-based XL-MS method, we developed a rescoring method inspired by the rescoring approach in DDA analysis. This rescoring algorithm incorporates 20 features (Supplementary Table 2), including those extracted from the theoretical and experimental relationships of multidimensional information of cross-linked peptides, along with multiple outputs from DIA-NN, and employs a deep neural network for rescoring. The effectiveness of the DIA rescoring method of XL-MSDigger was validated using 20 previously analyzed DIA datasets. We validated the effectiveness of this DIA rescoring method in XL-MSDigger using 20 previously analyzed DIA datasets. As shown in Fig. 5a, applying an FDR threshold of 5% resulted in the identification of 25,269 target cross-linked precursors, representing approximately a five-fold increase compared with the results prior to rescoring. Moreover, the score distributions of target and decoy precursors (Fig. 5b) exhibited clear separation after rescoring, indicating that our algorithm effectively distinguishes true targets from decoys. We further plotted the relationship between FDR thresholds and the number of identified target precursors, revealing that XL-MSDigger consistently yields higher identifications across the entire FDR range (Fig. 5c), demonstrating the robustness and superior sensitivity of the proposed rescoring approach.

In-depth DIA-based XL-MS analysis with predicted spectral library

Current DIA-based XL-MS methods rely on a spectral library obtained based on sample-specific DDA analysis, which can limit the overall coverage of DIA analysis because mass spectral information for some cross-linked peptides is not included in the library. In principle, our Deep4D-XL model, together with the DIA rescoring algorithm and FDR control of XL-MSDigger, enables the DIA-based XL-MS analysis using predicted spectral libraries. However, generating a proteome-wide theoretical cross-linked spectral library is computationally impractical due to its massive scale, making such an approach unfeasible at this stage. Therefore, we adopted a compromise strategy, constructing moderately scaled predicted spectral libraries for selected proteins or PPIs, allowing the implementation of 4D DIA XL-MS analysis under current computational constraints.

The feasibility was first evaluated by using a predicted spectral library to identify intra-protein cross-links. For this purpose, 100 proteins identified from the HeLa cross-linking proteome with available 3D structures were selected to generate predicted spectral libraries for all possible intra-protein cross-linked peptides (Supplementary Data 3). The DIA analysis was then conducted on 20 cross-linked sample fractions. To manage the spectral library size, we randomly divided the 100 proteins into five equal groups and constructed separate predicted spectral libraries for each group. The sizes of these libraries ranged from 1.06 to 1.94 million, with an average of 1.37 million. As a result, 90 intra-protein cross-linked peptides not identified by the DDA approach were identified using this workflow (Supplementary Data 3).

We then calculated the theoretical spatial distances of the cross-linked peptides within their 3D protein structures to assess the reliability of these identifications. Distance information was available for 88 peptides, of which 81 (92%) had residue distances within the constraints of the DSS cross-linker (Fig. 6a), comparable to the 85% observed for all identified cross-linked peptides with DDA analysis (Fig. 6b, Supplementary Data 3). The structures of two example proteins, GMPS and RUVBL1 and their cross-linking sites illustrate the advantage of the high coverage of the predictive library-based DIA approach in protein structure analysis (Fig. 6c). In addition, we

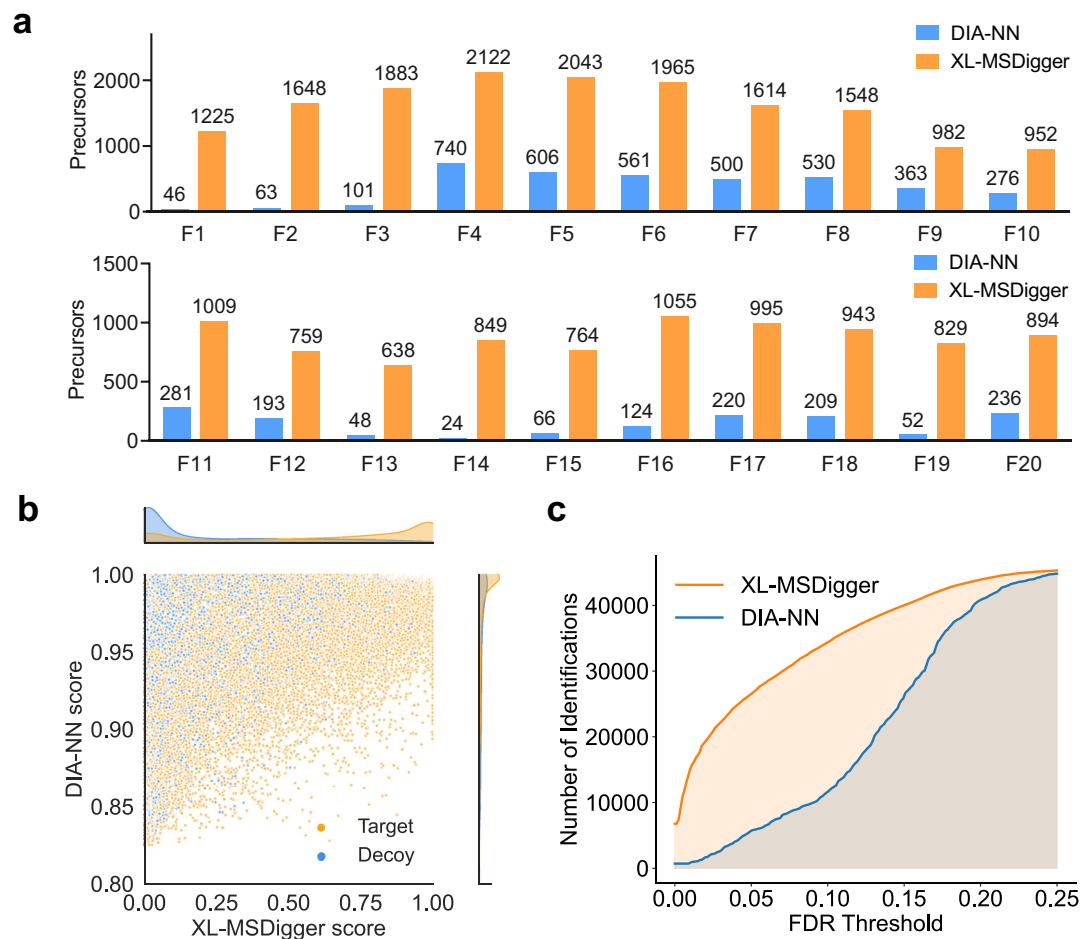


Fig. 5 | Effectiveness of the rescoring method in XL-MSDigger for DIA-based XL-MS analysis. **a** Comparison of identification counts between DIA-NN and XL-MSDigger for 20 HeLa cross-linked proteome fraction DIAsamples. **b** Scatter plot comparing target and decoy PSM scores between DIA-NN and XL-MSDigger, with

density plots displayed along the top (DIA-NN) and right (XL-MSDigger) axes. **c** FDR and identification number relationship curves for DIA-NN and XL-MSDigger. Source data are provided as a Source Data file.

calculated the distances between all possible intra-protein cross-linking sites for these 100 proteins and found that only 26.7% of the cross-linking sites met the distance constraint (Supplementary Fig. 9), which is much lower than our identified results. Furthermore, when constructing the predicted spectral library, we included an additional set of 100 randomly selected *E. coli* proteins as an entrapment control to further assess the quality of the identifications. The results showed that, at an FDR threshold of 5%, among 88 identified intra-protein cross-linked precursors, only two originated from *E. coli*. (Supplementary Data 3). These findings indicate that, with FDR assessment, DIA XL-MS using predicted spectral libraries achieves reliability and protein conformation analysis capabilities comparable to DDA.

To evaluate the performance of predicted spectral libraries in uncovering low-abundance cross-linked peptides, we selected 100 proteins with only one cross-linked peptide identified in DDA analysis and generated predicted spectral libraries for them (Supplementary Data 3). These libraries were then used for in-depth DIA analysis of 20 HeLa cross-linked protein samples. At a 5% FDR threshold, we identified 36 additional intra-protein cross-linked peptides that were not detected in the DDA results (Supplementary Data 3). Of these, 19 mapped to known 3D protein structures, and 15 met the theoretical distance constraints of the DSS cross-linker. These results demonstrate that the 4D DIA XL-MS with predicted spectral libraries enhances the detection of low-abundance intra-protein cross-linked peptides from DIA data.

We further evaluated the effectiveness of predicted spectral libraries in identifying inter-protein cross-links, focusing on HSP90AA1, a key molecular chaperone protein. Using the STRING database, we selected 17 proteins with the highest interaction confidence with HSP90AA1 and used these 17 PPIs to construct a predicted spectral library for DIA analysis. In the previous DDA results from the HeLa cross-linked proteome, only 3 of these 17 PPIs were identified. However, our analysis of DIA data from 20 HeLa samples using the predicted library identified 10 of these PPIs (Fig. 6d, Supplementary Data 4). To further assess the reliability of the identifications, we introduced an additional 17 *E. coli* proteins to form entrapment PPIs with HSP90AA1 during library construction. The entrapment experiment was randomly repeated ten times, and at an FDR threshold of 5%, only 5 out of a total of 109 identified cross-linked precursors (4.6%) corresponded to entrapment PPIs (Supplementary Data 4). These results further confirm the robustness and reliability of our DIA-based XL-MS analysis using predicted spectral libraries. Similarly, we studied HSPD1 and HSPA1B, selecting their interaction partners from STRING for spectral library construction and DIA analysis. While the DDA results show only 0 and 1 PPIs for HSPD1 and HSPA1B, respectively, the DIA analysis detected 4 and 7 PPIs (Supplementary Fig. 10, Supplementary Data 4). These results demonstrate the higher coverage of the DIA-based XL-MS analysis with predicted multidimensional spectral libraries compared to DDA, providing much more information on inter-protein cross-links and PPIs.

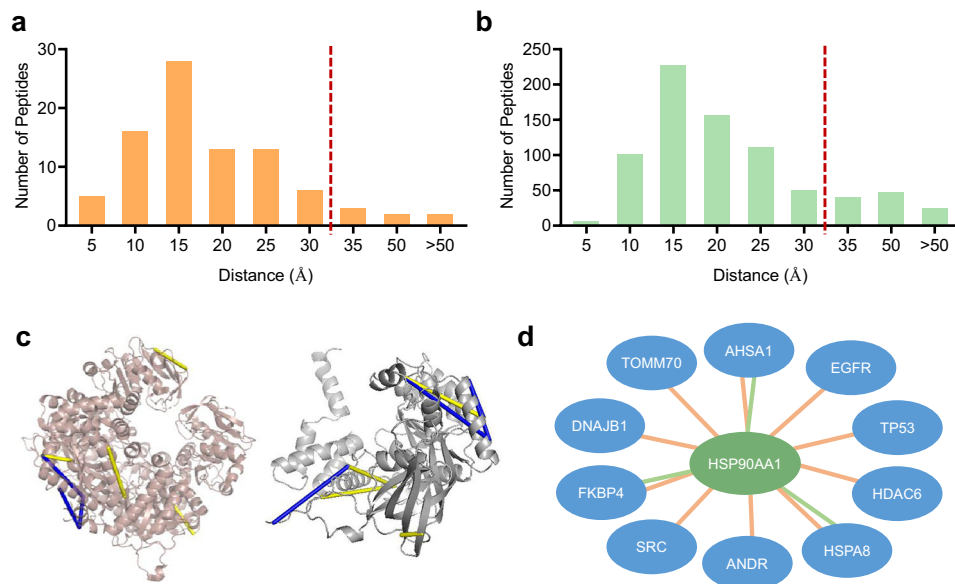


Fig. 6 | DIA-based XL-MS analysis based on predicted spectral libraries.

a Distribution of C α -C α distances between cross-linking sites of intra-protein cross-linked peptides uniquely identified in DIA analysis among 100 proteins with known structures, with the red line representing the theoretical distance constraint of the DSS cross-linker. **b** Distribution of C α -C α distances between cross-linking sites of intra-protein cross-linked peptides identified in DDA analysis among 100 proteins

with known structures. **c** Mapping schematic of identified cross-linked peptides in DDA results (yellow) and those unique in DIA results (blue) on two proteins (GMPS and RUVBL1) crystal structures. **d** Interactions between the HSP90AA1 protein and other 10 proteins identified in both DIA and DDA analyses, with lines indicating identified cross-linked peptides between two proteins, brown for DIA results, and green for DDA results. Source data are provided as a Source Data file.

Discussion

In this study, we developed Deep4D-XL, a deep learning tool for predicting multidimensional information of cross-linked peptides, including RT, CCS and MS/MS spectra. Based on Deep4D-XL, we further developed XL-MSDigger, a deep learning-driven pipeline for comprehensive analysis of both DDA- and DIA-based XL-MS data. By integrating Deep4D-XL with a deep neural network rescoring algorithm, we improved the identification depth of inter-protein cross-linked peptides in DDA analysis for both cleavable and uncleavable crosslinkers. We then constructed a spectral library of decoy cross-linked peptides using Deep4D-XL, which enables systematic FDR evaluation in DIA-based XL-MS analysis. Additionally, the application of a deep neural network-based rescoring algorithm to DIA-based XL-MS reduced the high FDR rates of existing DIA-based XL-MS identifications. We also demonstrated that the 4D DIA XL-MS with predicted spectral libraries achieves higher coverage of cross-linked peptide identifications compared to conventional DDA methods.

Although the training dataset for our model represents a large-scale collection of cross-linked peptide 4D information, it remains much smaller than the millions or even tens of millions of datasets available in traditional proteomics, leaving room for further improvements in the predictive accuracy of Deep4D-XL. As more XL-MS data becomes available, we plan to update and refine the model to improve performance.

In the field of XL-MS data analysis, the vast search space remains a major bottleneck. For DDA-based XL-MS, software such as pLink2 can already perform XL-MS data analysis at the proteome level within a few hours. However, for peptide-centric DIA XL-MS, although our approach using predicted spectral libraries for selected proteins improves identification depth, full proteome-scale DIA-based XL-MS analysis remains computationally infeasible. The n -square growth of the spectral library size poses a fundamental challenge. Accordingly, the current DIA implementation of XL-MSDigger is best suited for targeted or moderate-scale interaction analyses, where predicted spectral libraries can be constructed within practical computational

limits. A spectrum-centric⁴⁴ approach may offer a solution by converting DIA data into pseudo-MS/MS spectra and then using DDA-based algorithms for cross-linked peptide identification.

Methods

Sample preparation of cross-linked peptides

HeLa cells and yeast were each analyzed as a single biological sample. Briefly, the cell pellet was resuspended in cross-linking buffer (20 mM HEPES, 150 mM NaCl, 2.5 mM MgCl₂, pH 7.5) containing 1 \times EDTA-free protease inhibitor cock-tail and lysed by sonication on ice. Yeast was first ground in cross-linking buffer using a grinder before sonication. Cell debris was removed by centrifugation, protein concentration of the supernatant was determined by Bradford assay and adjusted to 1 mg/mL. Cross-linking reaction was performed using 2 mM DSS at 37 °C for 1 h and quenched with 50 mM Tris-HCl (pH 8.0). Cross-linked proteins were denatured with 8 M urea followed by reduction and alkylation. Samples were diluted by four times using 50 mM NH₄HCO₃ and digested with trypsin at an enzyme-to-protein ratio of 1:50 (w/w) at 37 °C overnight. The resulting peptides were desalted using C18 cartridges (Waters, USA), dried and stored at -80 °C for further use.

Strong cation exchange (SCX) fractionation of cross-linked peptide was performed on a PolySULFOETHYLA SCX column (3 μ m, 300 Å, 2.1 mm i.d. \times 150 mm, PolyLC, USA). SCX solvent A consisted of 0.1% TFA in 30% ACN, SCX solvent B consisted of 0.1% TFA, 0.5 M NH₄Cl in 30% ACN. The flow rate was set as 200 μ L/min. The digested products of 2 mg protein were dissolved with SCX solvent A and then loaded to the column. The SCX gradient was set as follows: 0–0.01 min (0–2% B); 0.01–8.01 min (2–3% B); 8.01–14.01 min (3–8% B); 14.01–28 min (8–20% B); 28–38 min (20–40% B); 38–48 min (40–90% B); 48–54 min (90% B); and 54–60 min (0% B). A total of 40 fractions with high proportions of cross-linked peptide were collected, with the eluent of 1 min as an individual fraction from 18 to 38 min. The fractions were named as F1, F2, F3 until to the last fraction of F40 in order of their elution time. All fractions were dried, de-salted using home-made C18 tips, dried and stored at -80 °C until LC-MS/MS analysis.

LC-MS/MS analysis

All the LC-MS/MS analysis of peptides were performed on a timsTOF Pro mass spectrometer (Bruker Daltonics, Germany) in PASEF or diaPASEF mode separately for DDA or DIA analysis via a CaptiveSpray nano-electrospray ion source (Bruker Daltonics, Germany) coupled with an EASY-nLC 1200 system (Thermo Fisher Scientific, USA). Samples were re-dissolved in mobile phase A and loaded onto a 20 cm in-house packed capillary column (75 μm inner diameter packed with 1.9 μm Venusil XBP C18 silica beads, Agela Technologies, China). The mobile phase for gradient elution consisted of mobile phase A (ddH₂O mixed with 0.1% formic acid) and mobile phase B [ACN-H₂O (8:2, v/v) mixed with 0.1% formic acid]. The chromatographic gradient program for both PASEF and diaPASEF analysis was set as follows: 2–28% B at 0–90 min, 28–46% B at 91–100 min, 46–95% B at 101–110 min, 95% B at 111–120 min.

For MS analysis operated in ddaPASEF mode, 1 survey TIMS-MS and 10 PASEF MS/MS scans were acquired during cycle time of 100 ms. Precursor ions for MS/MS analysis were isolated with a 2 Da window for $m/z < 700$ and 3 Da for $m/z > 800$ in a total m/z range of 100–1700. The collision energy was lowered linearly as a function of increasing ion mobility starting from 75 eV at $1/K_0 = 1.6 \text{ Vs}\cdot\text{cm}^{-2}$ to 25 eV at $1/K_0 = 0.6 \text{ Vs}\cdot\text{cm}^{-2}$. For analysis of cross-linked peptides, precursors with charge states of $\geq 3+$ and $\leq 6+$ were selected for fragmentation, while precursors with charge states $\geq 1+$ and $\leq 6+$ were selected for fragmentation of regular peptides.

For MS analysis operated in diaPASEF mode, a 2D polygonal filtering window was used for precursor selection, with m/z ranging from 400–1600, $1/K_0$ ranging from 0.74–0.84 at m/z of 400 and 1.22–1.58 at m/z of 1600. The m/z windows were divided to 80 sub-windows with m/z width set as 15. The collision energy was lowered linearly as a function of increasing mobility starting from 75 eV at $1/K_0 = 1.6 \text{ Vs}\cdot\text{cm}^{-2}$ to 25 eV at $1/K_0 = 0.6 \text{ Vs}\cdot\text{cm}^{-2}$.

DDA and DIA software parameters

pLink2 and pLink3 were used with the following parameters: precursor tolerance, 20 ppm; fragment tolerance, 20 ppm; enzyme, trypsin; number of missed cleavages, 3; peptide mass, [600, 6000]; peptide length, [6, 60]; fixed modification, carbamidomethylation (+57.021 Da) for cysteines; no variable modifications; filter tolerance, ± 10 ppm; FDR, 1% at PSM level.

Scout (1.6.3) were used with the following parameters: fragment m/z , [200, 1800]; crosslinker, DSSO; reaction residuals, k; enzyme, trypsin; enzyme specificity, FullySpecific; Isotopic possibilities precursor, 1; peptide length, [6, 60]; max variable modification per peptide, 2; peptide mass, [500, 6000]; max miscleavages, 3; ppm error on MS1 level, 10; ppm error on MS2 level, 20; static modification, Carbamidomethyl; variable modification, Oxidation on Methionine; FDR 100% on all level.

DIA-NN (1.8.1) was run with the following parameters: Mass accuracy, 10 ppm; MS2 accuracy, 15 ppm; scan window, 6; use isotopologues; no shared spectra; neural network class, double-pass model; quantification strategy, robust LC; precursor FDR, 1%.

Peptide encoding

We encoded each peptide of a cross-linked peptide separately. Each peptide was converted into a letter sequence where each amino acid was represented by its corresponding one letter code, such as A for alanine and K for lysine. Each amino acid was then encoded as a 23-dimensional one-hot vector, with the first twenty dimensions representing the 20 standard amino acids, the 21st and 22nd dimensions reserved for potential PTMs which were not used in this paper, and the 23rd dimension indicating crosslinking sites. The maximum length of a peptide was set to 50. For a peptide containing M amino acids, the $M \times 23$ one-hot tensors were zero-padded to 50×23 .

Deep4D-XL model

Deep4D-XL builds upon the Deep4D model with enhancements tailored to cross-linked peptides, incorporating siamese network structures and cross-attention mechanisms. The framework comprises encoding and decoding modules. Specifically, shared parameter encoding modules characterize each peptide in a cross-linked pair. A cross-attention mechanism then integrates the outputs of the encoding modules of both peptides. Finally, the decoding module processes the output from the cross-attention module to generate predictions. Detailed specifications of the encoding and decoding modules are provided below.

Encoding module. The encoded representation tensor of the peptide is first embedded through a 500-dimensional fully connected layer. Then the embedding tensor is put into several serial sub-encoders, two sub-encoders for CCS prediction and three sub-encoders for retention time and fragment ion intensity prediction. Each sub-encoder consists of 5 layers, and each layer is composed of multi-head attention and a fully connected feed-forward network, with residual connection and layer normalization. The number of heads and the dimension of the feed-forward network are set to 5 and 1200.

Decoding module. The decoder consists of a convolutional neural network and a fully connected feed-forward network. The output tensor of the Encoder was put into a convolutional neural network. For CCS and intensity ion prediction, the convolutional neural network contains 2 batch normalization layers, 2 convolution layers and 2 max pooling layers. For retention time prediction, the convolutional neural network contains 3 batch normalization layers, 3 convolution layers and 3 max pooling layers. Then, the 3D output tensor of the convolutional neural network was converted to 1D and put into a fully connected feedforward network to obtain the final predicted CCS, RT or fragment ion intensity.

We trained or fine-tuned our models on a machine with one NVIDIA Tesla V100 GPU. The Adam optimizer was used with $\text{beta1} = 0.9$, $\text{beta2} = 0.999$ and $\text{eta}_{\text{min}} = 10^{-9}$. Learning rate was set to $1e-4$ for pre-training and $1e-5$ for fine-tuning, with adjustment managed by the Cosine Annealing Warm Restarts method. The CCS and RT models were trained using the MSE loss function, while the MS2 model utilized a normalized spectral contrast loss function. Training was conducted for 70 epochs with early stopping based on the validation set.

DDA rescoring

The workflow includes 4D prediction, feature extraction, rescoring and FDR evaluation. 4D prediction of cross-linked peptides using Deep4D-XL: The Deep4D-XL model is first fine-tuned using 4D information of all intra-protein cross-linked peptides. Subsequently, PSMs of all cross-linked peptides, including TD and DD decoys, were extracted from the pLink2 result file. CCS, RT, and MS/MS spectra of these PSMs are predicted using a fine-tuned Deep4D-XL model.

Feature extractions. All 4D information predicted values and experimental values of cross-linked peptide PSM are used for feature extraction. There are a total of 16 features, including Charge, Original_score, Length 1, Length 2, RT_AE, CCS_RE, Match_num, Match_num1, Match_num2, Both_num, Both_num1, Both_num2, Cosine, SA, Pearson, and Spearman (Supplementary Table 1).

Rescoring using deep neural networks. The deep neural network consists of four layers of fully connected networks, with corresponding neuron numbers of 100, 1000, 100, and 1. First, all target cross-linked peptides that initially meet the FDR criteria of the pLink2 software were labeled as 1, and all decoy cross-linked peptides as 0. A subset of target cross-linked peptides that did not meet the FDR criteria remained temporarily unlabeled. Subsequently, all target and decoy cross-linked peptides labeled as 1 or 0 are evenly divided into n parts for n-fold cross-validation. Using this approach, n models are

trained, with each model being trained on n-1 parts and tested on the remaining part. During training, part of the data is selected from the training set as the validation set, and the evaluation metric of the validation set is the median absolute error. Based on the evaluation metric, the best model parameters are selected for the rescoring of the test set and unlabeled cross-linked peptides.

DIA FDR evaluation

The FDR evaluation was based on the identification results of diaPASEF cross-linked mass spectrometry data in DIA-NN software. A sample-specific 4D experimental spectral library was constructed based on the DDA identification results of cross-linked peptides. Here, we defined the cross-linked peptides in the experimental spectral library as TT, meaning that both peptides in the cross-linked pair were target peptides. For each target cross-linked peptide, its corresponding decoy cross-linked peptides are generated by reversing one or two independent linear peptide sequences it contains (excluding the terminal amino acid). The decoy cross-linked peptides include two types: TD, which contains one target and one decoy peptide, and DD, which contains two decoy peptides. Next, use the Deep4D-XL model to construct a decoy spectral library, and merge the experimental spectral library and the decoy spectral library into a comprehensive spectral library. Finally, use this spectral library with DIA-NN software to perform DIA analysis on the diaPASEF cross-linked mass spectrometry data. Calculate the FDR based on the numbers of target and decoy peptides at a given score threshold, using the formula:

$$\text{FDR} = \frac{(\#TD - \#DD)}{\#TT} \quad (1)$$

#TD and #DD respectively represent the number of two types of decoy peptides passing the score threshold, while #TT represents the number of target cross-linked peptides identified above the same threshold.

DIA rescoring

The workflow includes the construction of a spectral library and DIA analysis, feature extraction of cross-linked peptides, and rescore based on deep neural networks and FDR assessment.

Construction of spectral library. The spectral library consists of three parts: the sample-specific experimental spectral library, the decoy spectral library, and the theoretically predicted spectral library. Among these, the sample-specific experimental spectral library and decoy spectral library are mandatory options, while the theoretical predicted spectral library is optional and only needed for DIA analysis based on theoretical predictions.

Feature extraction. Based on the experimental information of the cross-linked peptide in the identification results and the prediction information of Deep4D-XL, the following features are extracted: Evidence, CScore, Q.Value, PEP, Average_corr, Pep1_num, Pep2_num, Pep1_num_matched, Pep2_num_matched, Spec_frag_num, Spec_frag_num1, Spec_frag_num2, Pep_cosine, Pep_entropy_smi, Pep1_cosine, Pep2_cosine, Pep1_entropy_smi, Pep2_entropy_smi, RT_AE, and CCS_RE (Supplementary Table 2).

Rescoring based on deep neural networks. Similar to the rescoring method in the previous DDA, a deep neural network is also used here as the rescoring model. The model consists of four fully connected layers with neuron counts of 100, 2000, 100, and 1, respectively. Initially, all cross-linked peptides identified by DIA-NN software are labeled, marking all decoy cross-linked peptides as 0. Then, an equal number of target cross-linked peptides from the experimental spectral library are randomly selected and marked as 1. These two datasets are combined for training the rescoring model. Remaining cross-linked peptides from the experimental spectral library are not labeled. If a theoretical prediction spectral library is added to the combined spectral library, then the cross-

linked peptides from the theoretical prediction spectral library are marked as 'test'. The model training and rescoring are performed using cross-validation. In each training session, 10% of the training set is selected as a validation set, with the evaluation metric being the median absolute error. The best model parameters, selected based on the evaluation metric of the validation set, are used after training to rescore the cross-linked peptides in the test set. The cross-linked peptides labeled as 1 or 0 are rescored using the model output when they were part of the test set in cross-validation. For all unlabeled or 'test' cross-linked peptides, the rescoring result is based on the average output of the model during cross-validation.

Validation and analysis of the cross-links. The high-resolution structures of the protein were obtained from RCSB PDB⁴⁵. The latest known PPI database STRING⁴⁰ was used for PPI validation. 3D protein structure visualizations have been generated with PyMOL Open-Source 2.5.0.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All the LC-MS/MS raw data, FASTA files, spectral library, training and test datasets have been deposited to the ProteomeXchange Consortium with the dataset identifier [PXD063217](https://doi.org/10.6017/PXD063217). The Orbitrap-based XL-MS datasets used in this study and generated using DSS, PhoX, and DSBU cross-linkers are available in the ProteomeXchange database under accession codes [PXD017620](https://doi.org/10.6017/PXD017620), [PXD029034](https://doi.org/10.6017/PXD029034), and [PXD012546](https://doi.org/10.6017/PXD012546), respectively. Source data are provided with this paper.

Code availability

The source code of XL-MSDigger is stored on GitHub (<https://github.com/Chen-micrab/XL-MSDigger>).

References

1. Rolland, T. et al. A proteome-scale map of the human interactome network. *Cell* **159**, 1212–1226 (2014).
2. Hegyi, H. & Gerstein, M. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288**, 147–164 (1999).
3. Bartolec, T. K. et al. Cross-linking mass spectrometry discovers, evaluates, and corroborates structures and protein–protein interactions in the human cell. *Proc. Natl. Acad. Sci. USA* **120**, e2219418120 (2023).
4. Piersimoni, L., Kastiris, P. L., Arlt, C. & Sinz, A. Cross-linking mass spectrometry for investigating protein conformations and protein–protein interactions—a method for all seasons. *Chem. Rev.* **122**, 7500–7531 (2022).
5. Iacobucci, C., Götz, M. & Sinz, A. Cross-linking/mass spectrometry to get a closer view on protein interaction networks. *Curr. Opin. Biotech.* **63**, 48–53 (2020).
6. Hauri, S. et al. Rapid determination of quaternary protein structures in complex biological samples. *Nat. Commun.* **10**, 192 (2019).
7. Ferber, M. et al. Automated structure modeling of large protein assemblies using cross-links as distance restraints. *Nat. Methods* **13**, 515–520 (2016).
8. Götz, M. et al. StavroX—a software for analyzing crosslinked products in protein interaction studies. *J. Am. Soc. Mass Spectrom.* **23**, 76–87 (2012).
9. Hoopmann, M. R. et al. Kojak: efficient analysis of chemically cross-linked protein complexes. *J. Proteome Res.* **14**, 2190–2198 (2015).
10. Yılmaz, Ş et al. Xilmass: a new approach toward the identification of cross-linked peptides. *Anal. Chem.* **88**, 9949–9957 (2016).

11. Götze, M. et al. Automated assignment of MS/MS cleavable cross-links in protein 3D-structure analysis. *J. Am. Soc. Mass Spectrom.* **26**, 83–97 (2015).
12. Dai, J., Jiang, W., Yu, F. & Yu, W. Xolik: finding cross-linked peptides with maximum paired scores in linear time. *Bioinformatics* **35**, 251–257 (2019).
13. Chen, Z.-L. et al. A high-speed search engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides. *Nat. Commun.* **10**, 3404 (2019).
14. Clasen, M. A. et al. Proteome-scale recombinant standards and a robust high-speed search engine to advance cross-linking MS-based interactomics. *Nat. Methods* **21**, 2327–2335 (2024).
15. Yang, K. L. et al. MSBooster: improving peptide identification rates using deep learning-based features. *Nat. Commun.* **14**, 4539 (2023).
16. Li, K., Jain, A., Malovannaya, A., Wen, B. & Zhang, B. DeepRescore: leveraging deep learning to improve peptide identification in immunopeptidomics. *Proteomics* **20**, 1900334 (2020).
17. Zhou, X.-X. et al. pDeep: predicting MS/MS spectra of peptides with deep learning. *Anal. Chem.* **89**, 12690–12697 (2017).
18. Gessulat, S. et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* **16**, 509–518 (2019).
19. Tiwary, S. et al. High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nat. Methods* **16**, 519–525 (2019).
20. Yang, Y. et al. In silico spectral libraries by deep learning facilitate data-independent acquisition in proteomics. *Nat. Commun.* **11**, 146 (2020).
21. Chen, M. et al. High-coverage four-dimensional data-independent acquisition proteomics and phosphoproteomics enabled by deep learning-driven multidimensional predictions. *Anal. Chem.* **95**, 7495–7502 (2023).
22. Ma, C. et al. Improved peptide retention time prediction in liquid chromatography through deep learning. *Anal. Chem.* **90**, 10881–10888 (2018).
23. Bouwmeester, R., Gabriels, R., Hulstaert, N., Martens, L. & Degroev, S. DeepLC can predict retention times for peptides that carry as-yet unseen modifications. *Nat. Methods* **18**, 1363–1369 (2021).
24. Wen, B., Li, K., Zhang, Y. & Zhang, B. Cancer neoantigen prioritization through sensitive and reliable proteogenomics analysis. *Nat. Commun.* **11**, 1759 (2020).
25. Meier, F. et al. Deep learning the collisional cross sections of the peptide universe from a million experimental values. *Nat. Commun.* **12**, 1185 (2021).
26. Giese, S. H., Sinn, L. R., Wegner, F. & Rappsilber, J. Retention time prediction using neural networks increases identifications in crosslinking mass spectrometry. *Nat. Commun.* **12**, 3237 (2021).
27. Chen, Z.-L., Mao, P.-Z., Zeng, W.-F., Chi, H. & He, S.-M. pDeepXL: MS/MS spectrum prediction for cross-linked peptide pairs by deep learning. *J. Proteome Res.* **20**, 2570–2582 (2021).
28. Ludwig, C. et al. Data-independent acquisition-based SWATH - MS for quantitative proteomics: a tutorial. *Mol. Syst. Biol.* **14**, e8126 (2018).
29. Hao, Y. et al. 4D-diaXLMS: proteome-wide four-dimensional data-independent acquisition workflow for cross-linking mass spectrometry. *Anal. Chem.* **95**, 14077–14085 (2023).
30. Rojas Echeverri, J. C. et al. A workflow for improved analysis of cross-linking mass spectrometry data integrating parallel accumulation-serial fragmentation with merx and skyline. *Anal. Chem.* **96**, 7373–7379 (2024).
31. Li, M. D. et al. Siamese neural networks for continuous disease severity evaluation and change detection in medical imaging. *npj Digit. Med.* **3**, 48 (2020).
32. Vaswani, A. et al. Attention is All You Need. *Advances in neural information processing systems* **30**, (2017).
33. Meier, F. et al. Online Parallel Accumulation–Serial Fragmentation (PASEF) with a novel trapped ion mobility mass spectrometer. *Mol. Cell. Proteom.* **17**, 2534–2545 (2018).
34. Linden, A. et al. A cross-linking mass spectrometry approach defines protein interactions in yeast mitochondria. *Mol. Cell. Proteomics* **19**, 1161–1178 (2020).
35. Jiang, P. et al. A membrane-permeable and immobilised metal affinity chromatography (IMAC) enrichable cross-linking reagent to advance in vivo cross-linking mass spectrometry. *Angew. Chem. Int. Ed.* **61**, e202113937 (2022).
36. Götze, M., Iacobucci, C., Ihling, C. H. & Sinz, A. A simple cross-linking/mass spectrometry workflow for studying system-wide protein interactions. *Anal. Chem.* **91**, 10236–10244 (2019).
37. 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* **17**, 41–44 (2020).
38. Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J. & Scholkopf, B. Support vector machines. *IEEE Intell. Syst. Their Appl.* **13**, 18–28 (1998).
39. Sengupta, S. et al. A review of deep learning with special emphasis on architectures, applications and recent trends. *Knowl-Based Syst.* **194**, 105596 (2020).
40. Szklarczyk, D. et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638–D646 (2023).
41. Chen, Z. A. & Rappsilber, J. Protein dynamics in solution by quantitative crosslinking/mass spectrometry. *Trends Biochem. Sci.* **43**, 908–920 (2018).
42. Müller, F., Kolbowski, L., Bernhardt, O. M., Reiter, L. & Rappsilber, J. Data-independent acquisition improves quantitative cross-linking mass spectrometry. *Mol. Cell. Proteom.* **18**, 786–795 (2019).
43. Meier, F. et al. diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition. *Nat. Methods* **17**, 1229–1236 (2020).
44. Tsou, C.-C. et al. DIA-Umpire: a comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods* **12**, 258–264 (2015).
45. Berman, H. M. The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

Acknowledgments

This work was financially supported by the Noncommunicable Chronic Diseases-National Science and Technology Major Project [2024ZD0532100 (S.M.C.)], National Natural Science Foundation of China [22474098 (S.M.C.), 22074111 (S.M.C.)], and National Key Research Development Program of China [2021YFC2700700 (S.M.C.)].

Author contributions

M.R.C., Y.H.H., and S.M.C. conceived and designed the project. M.R.C. performed computational work and wrote the manuscript. Y.H.H. conducted the cross-linking experiments and mass spectrometric experiments. X.H. and P.F.W. contributed to sample preparation. J.S. assisted in data analysis. B.Z. and S.M.C. revised and edited the manuscript. All authors participated in result discussions and provided critical feedback on the manuscript. S.M.C. supervised the overall research.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-026-69489-8>.

Correspondence and requests for materials should be addressed to Suming Chen.

Peer review information *Nature Communications* thanks Swantje Lenz and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026