Retention Time Prediction Using Neural Networks Increases Identifications in Crosslinking Mass Spectrometry

Sven H. Giese[*1,3,4], Ludwig R. Sinn[*1], Fritz Wegner[1], and Juri Rappsilber[†1,2]

[1] Bioanalytics, Institute of Biotechnology, Technische Universität Berlin, 13355 Berlin, Germany

[2] Wellcome Centre for Cell Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3BF, United Kingdom

[3] Data Analytics and Computational Statistics, Hasso Plattner Institute for Digital Engineering

[4] Digital Engineering Faculty, University of Potsdam

Abstract

Crosslinking mass spectrometry (Crosslinking MS) has developed into a robust technique that is increasingly used to investigate the interactomes of organelles and cells. However, the incomplete and noisy information in the spectra limits the numbers of protein-protein interactions (PPIs) that can be confidently identified. Here, we successfully leveraged chromatographic retention time (RT) information to aid the identification of crosslinked peptides from spectra. Our Siamese machine learning model xiRT achieved highly accurate RT predictions of crosslinked peptides in a multi-dimensional separation of crosslinked *E. coli* lysate. We combined strong cation exchange (SCX), hydrophilic strong anion exchange (hSAX) and reversed-phase (RP) chromatography and reached R^2 0.94 in RP and a margin of error of 1 fraction for hSAX in 94%, and SCX in 85% of the predictions. Importantly, supplementing the search engine score with retention time features led to a 1.4-fold increase in PPIs at a 1% false discovery rate. We also demonstrate the value of this approach for the more routine analysis of a crosslinked multiprotein complexes. An increase of 1.7-fold in heteromeric crosslinked residue-pairs was achieved at 1% residue-pair FDR for Fanconi anaemia monoubiquitin ligase complex, solely using reversed-phase RT. Retention times are a powerful complement to mass spectrometric information to increase the sensitivity of Crosslinking MS analyses.

[*] authors contributed equally

[†] corresponding author: juri.rappsilber@tu-berlin.de

Introduction

Crosslinking mass spectrometry (Crosslinking MS) reveals the topology of proteins, protein complexes, and protein-protein interactions.[1] Fueled by experimental and computational improvements, the field is moving towards the analyses of interactomes of organelles and cells.[1–3] The identification of crosslinked peptides poses three major challenges. First, the low abundance of crosslinked peptides compared to linear peptides decreases their chance for mass spectrometric observation. Second, the unequal fragmentation of the two peptides leads to a biased total crosslinked peptide spectrum match (CSM) score[4,5]. Third, the combinatorial complexity from searching all the possible peptide pairs in a sample increases the chance for random matches. These challenges increase from the analysis of individual proteins to organelles and cells.

To address the challenge of low abundance, Crosslinking MS studies routinely rely on chromatographic methods to enrich and fractionate crosslinked peptides[1,2,6]. Essentially all analyses contain at least one chromatographic step, by directly coupling reversed-phase (RP) chromatography separation to the mass spectrometer (LC-MS). Additional separation is frequently employed when more complex systems are being analysed. Strong cation exchange chromatography (SCX)[7,8] was used for the analysis of HeLa cell lysate[9] or murine mitochrondria[10]. Size-exclusion chromatography (SEC)[11] was used to fractionate crosslinked HeLa cell lysate[12] and *Drosophila melanogaster* embryos extracts[13]. Multi-dimensional peptide pre-fractionation was used for the analysis of crosslinked human mitochondria (SCX-SEC)[14] and *M. pneumoniae* (SCX-hSAX)[15]. Such multi-dimensional chromatography workflows can yield in the order of 10,000 CSM at 1-5% false discovery rate (FDR).[14–17]

The identification of crosslinked peptides from spectra is however still challenged by the uneven fragmentation of the two peptides and the large search space that increase the odds of random matches. This is especially the case for heteromeric crosslinks as the size of their search space exceeds that of self-links, i.e. links falling within a protein or homomer[16]. Typically, database search tools use the precursor mass and fragmentation spectrum for the identification of peptides to compute a single

final score for each CSM. For linear peptides, post-search methods such as Percolator[18] have been developed that train a machine learning predictor to discriminate correct from incorrect peptide identification. Percolator uses additional spectral information (features) such as charge, length, and other enzymatic descriptors of the peptide[19] to compute a final support vector machine (SVM) score. Similarly, the crosslink search engine Kojak[20] supports the use of PeptideProphet[21,22] and XlinkX[23] supports Percolator[18], while pLink2[24] and ProteinProspector[4] have a built-in SVM classifier to re-rank CSMs. Although RT data is readily available, none of these tools use the, often multi-dimensional, RT information for improved identification in crosslinking studies. A prerequisite for this would be that retention times could be predicted reliably.

For linear peptides, RT prediction has been implemented under various chromatographic conditions.[25–31] In contrast, RTs of crosslinked peptides have not been predicted yet. A suitable machine learning approach for this could be deep learning[32]. Deep neural networks have been successfully applied in proteomics, for example for de novo sequencing[33] or for the prediction of retention times[29,34] and fragment ion intensities[35]. Deep learning allows encoding peptide sequences very elegantly through, for example, recurrent neural network (RNN) layers. These layers are especially suited for sequential data and are common in natural language processing[32]. RNNs use the order of amino acids in a peptide to generate predictions without additional feature engineering. However, it is unclear how to encode the two peptides of a crosslink.

Moreover, it is also unclear whether the knowledge of RTs could improve the identification of crosslinked peptides. A common scenario for an identified crosslink is that one of its peptides was matched with high sequence coverage, while the other was matched with poorer sequence coverage.[4] Such CSMs unfortunately resemble matches where one peptide is correct and the other is false (i.e. a target-decoy match or a true target and false target match). Another consequence of coverage gaps is the misidentification of noncovalently associated peptides as crosslinks.[36] The severity of this coverage issue depends on the applied acquisition strategy[37], crosslinker chemistry[38], and the details of the implemented scoring in the search engine. Nevertheless, assuming RT predominantly depends on both

peptides of a crosslink, it could complement mass spectrometric information and thus improve existing scoring routines and lead to more crosslinks at the same confidence (i.e. constant FDR).

In this study, we prove that analytical separation behavior carries valuable information about both crosslinked peptides and can improve the identification of crosslinks. For this we built a multi-dimensional RT predictor for crosslinked peptides based on a proteome-wide crosslinking experiment comprising 144 acquisitions on an Orbitrap mass spectrometer from extensively fractionated peptides of the soluble high-molecular weight proteome of *E. coli*. We then investigated the benefits of incorporating the derived RT predictions into the identification process. In addition, we demonstrate the value of RT prediction for a purified multiprotein complex using the reversed-phase chromatography dimension only.

Material and Methods

Sample Preparation

Crosslink samples were processed exactly as described in Lenz *et al.*[16] with the exception that the crosslinker DSS was used. Briefly, cells were lysed by sonication, cleared from debris and the high-molecular weight proteome enriched by ultrafiltration. This sample was then fractionated by size-exclusion chromatography to give 44 fractions. The proteins of each fraction were crosslinked at 0.75 mM DSS. The crosslinked samples were pooled and precipitated using acetone. Upon resuspending, the samples were derivatized by incubating 30 minutes at room temperature with 10 mM dithiothreitol followed by 20 mM iodoacetamide and proteolyzed using LysC and Trypsin. The digests were fractionated, first, by strong cation exchange chromatography (9 fractions) and the obtained fractions separated by hydrophilic strong anion exchange chromatography as the second separation dimension (10 pools). Samples were cleaned up in between and at the end of the procedures following the StageTip protocol[39].

Spectra & Peptide Spectrum Match Processing

All raw spectra were converted to Mascot generic format (MGF) using msConvert[40]. The database search with Comet[41] (v. 2019010) was done with the following settings: peptide mass tolerance 3 ppm; isotope_error 3; fragment bin 0.02; fragment offset 0.0; decoy_search 1; fixed modification on C (carbamidomethylation, +57.021 Da); variable modifications on M (oxidation, +15.99 Da). False discovery rate (FDR) estimation was performed for each acquisition. First, the highest scoring PSM for a modified peptide sequence was selected, then the FDR was computed based on Comet's e-value. Spectra were searched using xiSEARCH (v. 1.6.753)[12], after recalibration of precursor and fragment m/z values, with the following settings: precursor tolerance, 3 ppm; fragment tolerance, 5 ppm; missed cleavages, 2; missed monoisotopic peaks[42], 2; minimum peptide length, 7; variable modifications: oxidation on M, mono-links for linear peptides on K,S,T,Y, fixed modifications: carbamidomethylated C. The specificity of the crosslinker DSS was configured to link K, S, T, Y, and the protein N-terminus with a mass of 138.06807 Da. The searches were run with the workflow system snakemake[43]. The FDR on CSM-level was defined as FDR = TD - DD / TT[44], where TD indicates the number of target-decoy matches, DD the number of decoy-decoy matches and TT the number of target-target matches. Crosslinked peptide spectrum matches (CSMs) with non-consecutive peptide sequences were kept for processing[45]. PPI level FDR computation was done using xiFDR[44] (v. 2.1.3 and 2.1.5 for writing mzIdentML) to an estimated PPI-FDR of 1%, disabling the boosting and filtering options. CSM, peptide and residue-level FDR were fixed at 5%, protein group FDR was set to 100%. FDR estimations for self and heteromeric links were done separately. In xiFDR a unique CSM is defined as a combination of the two peptide sequences including modifications, link sites and precursor charge state. For the assessment of identified CSMs an entrapment database (described in the next section) as well as decoy identifications were used on both, CSM and PPI levels. PPI results were also compared against the APID[46] and STRING[47] databases (v11, minimal combined confidence of 0.15).

Database Creation

The database of potentially true crosslinks was defined as *Escherichia coli* proteome (reviewed entries from Uniprot release 2019-08). This database was filtered further to proteins identified with at least a single linear peptide at a q-value[48] threshold of 0.01, $q(t) = \min_{s \leq t} FDR\ (s)$, with the threshold t and score s. This resulted in 2850 proteins. In addition to the FDR estimation through a decoy database, we used an entrapment database. The proteins from the entrapment database represent the search space of false positive CSMs independent of *E. coli* decoys and were sampled from human proteins (UP000005640, retrieved 2019-05). *E. coli* decoys might fail in this task after machine learning if overfitting should have taken place. So, entrapment targets allow control for overfitting. For this, human target peptides were treated as targets and human decoy peptides as decoys. To avoid complications through false spectrum matches due to homology, we used blastp[49] (BLAST 2.9.0+, blastp-short mode, word size 2, e-value cutoff 100) and aligned all *E. coli* tryptic peptides (1 missed cleavage, maximum length 100) to the human reference. All proteins that showed peptide alignments with a sequence identity of 100% were removed from the human database. Only the remaining 9990 sequences were used as candidates in the entrapment database. For each of the 2850 *E. coli* proteins a human protein was added to the database. To reduce search space biases from protein length and thus different number of peptides for the two organisms, we followed a special sampling strategy. The human proteins were selected by a greedy nearest neighbor approach based on the K/R counts and the sequence length. The final number of proteins in the combined database (*E. coli* & human) was 5700 (2850*2).

Fanconi anaemia monoubiquitin ligase complex data processing

The publicly available raw files from an analysis of the BS3-crosslinked Fanconi anemia monoubiquitin ligase complex[50] (FA-Complex) were downloaded from PRIDE together with the original FASTA file (PXD014282). The raw files were processed as described for the *E. coli* data (m/z recalibration and searched with xiSEARCH), followed by an initial 80% CSM-FDR filter for further processing. Due to the much smaller FASTA database (8 proteins), the entrapment database was constructed more conservative than for the proteome-wide *E. coli* experiment, i.e. for each of the target proteins, the amino acid composition was used to retrieve the nearest neighbor in an *E. coli* database. The FDR settings to evaluate the rescoring were set to 5% CSM- and peptide-pair level FDR, 1% residue-pair- and 100% PPI-FDR using xiFDR without boosting or additional filters. The resulting links were visualized (circular view) and mapped to an available 3D structure (final refinement model 'sm.pdb')[51,52] using xiVIEW[53]. To ease the comparison of identified and random distances, a random Euclidean distance distribution was derived in three steps: first, all possible crosslinkable residue-pair distances in the 3D structure were computed. Second, 300 random 'bootstrap' samples with n distances were drawn (n= the number of identified residue-pairs at a given FDR) and third, the mean per distance bin was computed across all 300 samples.

xiRT - 3D Retention Time Prediction

The machine learning workflow was implemented in python and is freely available from https://github.com/Rappsilber-Laboratory/xiRT. xiRT is the successor of DePART[29], which was developed for the retention time (RT) prediction of hSAX fractionated peptides based on pre-computed features. xiRT makes use of modern neural network architectures and does not require feature engineering. We used the popular python packages sklearn[54] and TensorFlow[55] for processing (section S1 for more details). xiRT consists of five components (Fig. 1d, Fig. S1, Section S1): (1) The input for xiRT are amino acid sequences with arbitrary modifications in text format (e.g. Mox for oxidized Methionine). xiRT uses a similar architecture for linear and crosslinked peptide RT prediction. Before the sequences can be used as input for the network, the sequences are label

encoded by replacing every amino acid by an integer and further 0-padded to guarantee that all input sequences have the same length. Modified amino acids as well as crosslinked residues are encoded differently than their unmodified counterparts. (2) The padded sequences were then forwarded into an embedding layer that was trained to find a continuous vector representation for the input. (3) To account for the sequential structure of the input sequences, a recurrent layer was used (either GRU or LSTM). Optionally, the GRU/LSTM layers were followed by batch normalization layers. For crosslinked peptide input, the respective outputs from the recurrent layers were then combined through an additive layer (default setting). (4) Task-wise subnetworks were added for hSAX, SCX, and RP retention time prediction. All three subnetworks had the same architecture: three fully connected layers, with dropout and batch normalization layers between them. The shape of the subnetworks is pyramid-like, i.e. the size of the layers decreased with network depth. (5) Each subnetwork had its own activation function. For the RP prediction, a linear activation function was used and mean squared error (*MSE*) as loss function. For the prediction of SCX and hSAX fractions we followed a different approach. The fraction variables were encoded for ordinal regression in neural networks[56]. For example, in a three-fraction setup, the fractions (f) were encoded as $f_1 = [0,0,0], f_2 = [1,0,0]\ and\ f_3 = [1,1,0]$. Subsequently, we chose sigmoid activation functions for the prediction layers and defined binary cross entropy (*BC*) as loss function. To convert predictions from the neural network back to fractions, the index of the first entry with a predicted probability of less than 0.5 was chosen as the predicted fraction. The overall loss was computed by a weighted sum of the MSE_{RP}, BC_{SCX} and BC_{hSAX}. The weight parameters are only necessary when xiRT is used to predict multiple RT dimensions at the same time (multi-task). To predict a single dimension (single-task, e.g. RP only), the weight can be set to 1. The number of neurons, dropout rate, intermediate activation functions, the weights for the combined loss, number of epochs and other parameters in xiRT were optimized on linear peptide identification data. Reasonable default values are provided within the xiRT package. For optimal performance, further optimization might be necessary for a given task.

Cross-Validation and Prediction Strategy

Cross-validation (CV) is a technique to estimate the generalization ability of a machine learning predictor[57] and is often used for hyper-parameter optimization. We performed a 3-fold CV for the hyper-parameter optimization on the linear peptide identification data from xiSEARCH, excluding all identifications to the entrapment database (section S2 and Fig. S2 for details). We defined a coarse grid of parameters (Tab. S1) and chose the best performing parameters based on the average total (unweighted) loss, R^2_{RP} and accuracy across the CV folds. Further, we define the relaxed accuracy (racc) to measure how many predictions show a lower prediction error than $|1|$. We then repeated the process with an adapted set of parameters (Tab. S2). In addition to the standard CV strategy, we used a small adjustment: per default, in k-fold cross-validation, the training split consists of k-1 parts of the data (folds) and a single testing fold. However, we additionally used a fraction (10%) from the training folds as extra validation set during training. The validation set was used to select the best performing classifier over all epochs. The model assessment was strictly limited to the testing folds. This separation into training, validation and testing was also used for the semi-supervised learning and prediction of RTs, i.e. when xiRT was used to generate features to rescore CSMs previously identified from mass spectrometric information. In this scenario, the CV strategy was employed to avoid the training and prediction on the same set of CSMs. In xiRT, a unique CSM is defined as combination of the two peptide sequences, ignoring link sites and precursor charge.

Supervised Peptide Spectrum Match Rescoring

To assess the benefits of RT predictions, we used a semi-supervised support vector (SVM) machine model. The implementation is based on the python package scikit-learn[58] in which optimal parameters are determined via cross-validation. The input features were based on the initial search score (for FA-complex only) and differences between predicted and observed RTs. For each crosslinked peptide, three predictions were made per chromatographic dimension: for the crosslinked peptide, for the alpha peptide and the beta peptide. Additional features were engineered depending on the number of chromatographic dimensions and included the summed, absolute or

squared values of the initial features (Tab. S3 for all features). For example, for three RT dimensions, the total number of features was 43. The data for the training included all CSMs that passed the 1% CSM-FDR cutoff (self / heteromeric, TT, TD, DDs) and TD/DD identifications that did not pass this cutoff. TTs were labeled as positive training examples, TD and DDs (DXs) were labeled as negative training examples.

To stratify the *k*-folds during CV, the CSMs were binned into *k* xiSCORE percentiles. Afterwards, they were sampled such that each score range was equally represented across all CV folds. When the positive class was limited to the TT identifications at 1% CSM-FDR, the number of negative classes was usually larger than the number of positive classes. To circumvent this, for each CV split, a synthetic minority over-sampling technique (SMOTE)[59] was used to generate a balanced number of positive and negative training samples (here only used for the FA-complex data). SMOTE was applied within each CV fold to avoid information leakage. A 3-fold CV was performed for the rescoring. In each iteration during the CV, two folds were used for the training of the classifier and the third fold was used to compute an SVM score. During this CV step, a total of three classifiers were trained. The scores for all TT-CSMs that did not pass the initial FDR cutoff were computed by averaging the score predictions from the three predictors. For all CSMs passing the initial FDR cutoff, rescoring was performed when the CSM occurred in the test set during the CV. The final score was defined as: $xi_{rescored} = xi_{SCORE} + xi_{SCORE} \times SVM_{score}$, where SVM_{score} was the output from the SVM classifier and xi_{SCORE} the initial search engine score.

Feature Analysis

The KernelExplainer from SHAP[60] (Shapley Additive exPlanations) was used to analyze the importance of features derived from the SVM classifier. SHAP estimates the importance of a feature by setting its value to "missing" for an observation in the testing set while monitoring the prediction outcome. We used a background distribution of 200 samples (100 TT, 100 TD) from the training data to simulate the "missing" status for a feature. SHAP values were then computed for 200 randomly selected TT (predicted to be TT) that were not used during the SVM training. SHAP values allow to directly

estimate the contributions of individual features towards a prediction, i.e. the expected value plus the SHAP values for a single CSM sums to the predicted outcome. For a selected CSM, a positive SHAP value contributes towards a true match prediction. For the interpretability analysis (SHAP) of the learned features in xiRT, the DeepExplainer was used (section S3).

In addition, we performed dimensionality reduction using UMAP[61] on the RT feature space for visualization purposes (excluding the search engine score). UMAP was run with default parameters (n_neighbors=15, min_dist=0.1) on the standardized feature values. The list of used features for the multi-task learning setup is available in Tab. S3.

Statistical Analysis

Significance tests were computed using a two-sided independent t-test with Bonferroni correction. The significance level α was set to 5%.

Data Availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the jPOST partner repository[62‡] with the dataset identifier PXD020407 and DOI 10.6019/PXD020407. Raw data of the FA-Complex is available via the previously published PRIDE identifier (PXD014282). Additional files and intermediate results are available via Zenodo (10.5281/zenodo.4270324). Source data are provided with this manuscript.

Code Availability

The developed python package is available on the python package index and on GitHub (https://github.com/Rappsilber-Laboratory/xiRT).

‡ Access: https://repository.jpostdb.org/preview/1564483676042143f9d3ae, key: 3194

Results and Discussion

This section covers 1) a description of the experimental workflow and the motivation, 2) the evaluation of the developed retention time predictor, 3) an interpretability analysis of the deep neural network, 4) an analysis of the RT features and their importance for rescoring, 5) the evaluation of the rescoring results from an *E. coli* lysate, and 6) the evaluation of the rescoring results from a routine crosslinking MS experiment, i.e. the analysis of a multiprotein complex (FA-complex).

A Substantial Fraction of Crosslinks below the Confidence Threshold are Correct

Crosslinked peptides belonging to the high-molecular weight *E. coli* proteome were deep-fractionated along three chromatographic dimensions (hSAX, SCX and RP). This 3D fractionation approach led to 144 LC-MS runs as some of the 90 fractions contained enough material for repeated analysis. The resulting data were searched with an entrapment database approach (Fig. 1a) leading to 11196 CSMs (11072 TT, 87 TD, 37 DD, Fig. S3) at 1% CSM-FDR, separating self and heteromeric CSMs[16,44,63]. The human entrapment database allows to assess error, independently of the target-decoy approach. This will play a critical role here as *E. coli* decoys will be used for the machine learning-based rescoring (but not for the RT prediction). Judged by a set of peptide characteristic metrics (e.g. peptide length, pI, GRAVY) the human entrapment database resembles the properties of the *E. coli* target database (Fig. S4).

Before attempting RT prediction and subsequent complementation of search scores, we investigated the extent of false negatives, approximated here by PPIs present in STRING[47] or APID[46] database. At 1% CSM-FDR, 110 such 'validated' (val) protein-protein interactions were identified. 10%, 30% and 50% CSM-FDR returned 226, 278 and 418 validated PPIs, respectively (Fig. 1b). When raising the CSM-FDR from 1% to 50% we thus saw a nearly 4-fold increase in the detectable number of validated PPIs. In contrast, using a pessimistic approach of semi-randomly drawing pairs of *E. coli* proteins from the STRING/APID (first protein) and the search database (second protein) yielded purely by chance 10, 22, 44, and 91 overlapping PPIs with STRING or APID for 1%, 10%, 30% and 50% CSM-FDR cutoffs,

293 respectively. While this shows that loosening the FDR threshold increases validated PPIs also by
294 chance, the actually observed number is much higher (418 versus 91 at 50% CSM-FDR). This means
295 that there is a substantial number of valid PPIs with insufficient match confidence.

296 The underlying scoring challenge is essential to the identification of peptides in general. The plethora
297 of search engines for linear[64] and crosslinked peptides[65] use spectral characteristics differently for
298 their scoring. In xiSEARCH, the final score is a composite that incorporates spectral metrics such as
299 explained intensity and matched number of fragments. Empirically, we observe a fast decrease in the
300 search engine score (Fig. 1c) with increasing FDR. This indicates that at higher FDRs spectral matching
301 metrics might be suboptimal. Poor spectral quality, inefficient peptide fragmentation or random
302 fragment matching all influence the search engine score negatively. RT information could
303 complement MS information but this would require accurate RT prediction of crosslinked peptides.

304 Accurate Multi-dimensional Retention Time Prediction for Crosslinked Peptides

305 RT prediction for crosslinked peptides has not yet been achieved. One reason for this is the challenge
306 of encoding a crosslinked pair of peptides for machine learning. We overcame this here using a
307 Siamese neural network as part of a new machine learning application, xiRT (Fig. 1d), which allowed
308 the incorporation of RTs into a rescoring workflow (Fig. 1e). The Siamese part of the network
309 (embedding layer and recurrent layer) shares the same weights for both peptides. Practically, the
310 sharing of weights leads to consistent predictions, independent of the peptide order. After the
311 recurrent layer, the two outputs were combined and passed to three subnetworks consisting of
312 dense layers with individual prediction layers (details on the architecture are available in Fig. S1). In
313 this multi-task learning setup, the network simultaneously learned to predict the hSAX, SCX and RP
314 RT through a single training step. Multi-task learning can improve the overall performance of
315 predictors by forcing the network to learn a robust representation of the input data.[66]

316 The training and evaluation of xiRT followed a CV strategy that avoided the simultaneous learning
317 and prediction on overlapping parts of the data (Methods, Fig. 2a). We used a 3-fold CV strategy

where two folds were used for training (excluding 10% for the validation throughout the training epochs) and one fold for testing/prediction. All CSMs with an FDR < 1% were used during CV. For the remaining CSMs, the best predictor (with the lowest total loss) was used to predict the RTs.

To achieve the best possible prediction performance, hyper-parameters of the network were optimized. Since extensive hyper-parameter optimization on a small data set can lead to overfitting, we initially optimized a large part of hyper-parameters using 20,802 unique linear peptide identifications at 1% FDR. The final parameters for the Siamese network architecture for crosslinks were obtained by a small grid-search (6,453 unique peptide-pairs at 1% CSM-FDR; Fig. S5).

Using these parameters, we evaluated the learning behavior during the training time (epochs) across the CV folds. The training behavior on the three CV folds was similar and reached a stable trajectory after approximately 15 epochs (Fig. 2b). Based on very similar error trends on validation and training sets, we concluded to have reached a state where neither overfitting nor underfitting occurred. The overall performance across the prediction folds was comparable in terms of accuracy (hSAX: 61% $\pm$ 1.1, SCX: 47% $\pm$ 1.7) and MSE (11.58 $\pm$ 2.0)(Fig. 2c). Comparing single-task and multi-task configurations of xiRT revealed no significant differences in the prediction accuracy but greatly reduced run times (Fig. S6-7). Note that we estimated the theoretical boundaries given the ambiguous elution behavior (i.e. peptide elution across multiple chromatographic fractions) for SCX at 65% accuracy and for hSAX at 73% accuracy (Tab. S4, Fig. S8). Most of the predictions showed only a small error, and thus a high relaxed accuracy: for hSAX 94% $\pm$ 0.0 and for SCX 87% $\pm$ 1.15 of the predictions were within a range of $\pm$ 1 fraction (Fig. 2d-e). The R^2_{RP} of 0.94 $\pm$ 0.01 also showed a predictable relationship for the RP dimension (Fig. 2f). The consistent accuracy and R^2 results across CV folds demonstrates reproducible training and prediction behavior which reduces unwanted biases from the different CV folds. In conclusion, RTs of crosslinked peptides can robustly be learned within a data set, making them available as features in a CSM rescoring framework.

It was difficult to compare our RT predictions to other studies which used SCX[67] or hSAX[29] for multiple reasons: 1) there is currently no other model that predicts the RT of crosslinked peptides, 2) the recent SSRCalc[67] study (SCX) for linear peptides used a much larger data set of 34,454 unique peptides and the fractionation was much more fine-grained (30 - 50 fractions). Similarly, the hSAX[29] study on linear peptides used a much finer fractionation (30 fractions) and a different methodology to encode the loss function during the machine learning. 3) Applied gradients and liquid chromatography conditions can change the elution behavior quite drastically. In our study, the number of observations was neither for hSAX nor for SCX equally distributed but varied between ~200 and ~2000 CSMs per fraction (Fig. S3). Since we employed a partially exponential gradient during the chromatographic fractionation, the degree of peptide separation varied for earlier and later fractions.

Given that we had less data to train on than recent RT predictions of linear peptides, we evaluated how the numbers of observations influenced the prediction accuracy ($R^2_{RP} + \mathrm{Acc}_{\mathrm{hsax}} + Acc_{scx}$, Fig. 2g). The learning curve showed two important characteristics: first, the prediction performance over CV folds was very reproducible. This means that predictions were robust even with very moderate data quantity. Second, the maximal performance was achieved with approximately 70%-100% of the data points (100% corresponding to 6453 total CSMs, 3871 for training, 431 for validation, 2151 for prediction). Given that a first plateau was reached with 30% of the data, it is unclear if the final prediction accuracy constitutes another local optimum or the limit of the prediction accuracy. The individual task metrics showed that the RP behavior seemed to be easier for the model to learn than the ordinal regression tasks (SCX, hSAX, Fig. S9). The RP behavior could be accurately predicted from approximately 60% of the data points, while the maximum accuracy for hSAX and SCX dimensions was only achieved by using 80% - 100% of the data. In other words, while using even fewer CSMs might be possible when predicting RP RTs, one would expect a reduced accuracy in the hSAX/SCX dimensions.

An approach to reduce the number of required CSMs would be to leverage the abundantly available data on linear peptides for transfer learning. Indeed, a recent study showed that transfer learning across different peptide identification results works well for linear peptides[34]. However, in our hands, pretraining a network from linear peptides and applying the same weights to the Siamese part of the network neither improved the performance nor reduced the training time for crosslink RT predictions (data not shown). In contrast, a robust and accurate RT prediction could be achieved on a multiprotein complex crosslinking study (FA-complex, see below) when first training on the *E. coli* CSMs (Fig. S10). Another possibility to increase the training data size and robustness during CV is to increase the number of folds, e.g. 5- or 10-fold, at the cost of runtime. Increasing the expedience of xiRT, we also implemented transfer learning for cases when the number of fractions differs between the initial model and the new prediction task.

Explainable Deep Learning Reveals Amino Acid Contributions

Using the SHAP package, we set out to explain predictions made by xiRT. For instance, when a specific crosslinked peptide was analyzed, residue-specific contributions towards the predicted RT could be computed (Fig. S11). The residues D, E, Y and F displayed high SHAP values indicating a stronger retention during hSAX separation in a randomly chosen peptide. Looking at a specific crosslinked peptide in SCX (Fig. S12), the SHAP values highlighted that K and R were the most important residues contributing towards later peptide elution. As one might expect, crosslinked K residues contributed much less towards later elution times than the stronger charged, unmodified K residues. Investigating the SHAP values for a collection of CSMs revealed additional contributions from W for hSAX and H for SCX while returning hydrophobic residues Y, F, W, I, L, V and M for RP (Fig. S13), revealing residue contributions in crosslinked peptides as seen in the respective analyses of linear peptides[29,67,68]. In summary, the SHAP values were good estimates for the individual RT contributions of the amino acid residues.

Next, we investigated the network architecture and the learned feature representations more closely (Section S4). As first analysis, the dimensionality reduced embedding space across the network was

393 analyzed (Fig. S14). This revealed that the shared sequence-specific layer already captured the RP
394 properties quite well, while the hSAX and SCX properties were not as clearly captured. As expected,
395 the separation of CSMs according to RT increased the further the features propagated through the
396 network. In the last layer, the RP and hSAX sub-networks reached a very good separation, while in
397 the SCX subtask CSMs remained moderately separated in two dimensions.

398 RT Characteristics for Unsupervised Separation of True and False CSMs

399 Now that we established the RT prediction of crosslinked peptides, we computed a set of
400 chromatographic features to explore their ability to separate true from false CSMs (Tab. S3).
401 Dimensionality reduction was computed for RP only (13 chromatographic features) and for SCX-
402 hSAX-RP (43 chromatographic features) predictions (Fig. 3a-b). Both chromatographic feature sets
403 revealed good separation possibilities for confident TT (99% true, given 1% CSM-FDR) and TD (100%
404 false) identifications in two-dimensional space. For the RP analysis, the TD *E. coli* CSMs and TT Mix /
405 TD Mix CSMs were enriched in one area of the plot (the lower right part, Fig. 3a). In contrast, the
406 subset of confident TT *E. coli* CSMs were distributed outside this area. As one would expect for two
407 sets of random matches, the CSMs from the entrapment database (TT Mix, TD Mix) closely followed
408 the distribution of TD *E. coli* CSMs. The areas populated by the known false matches were also
409 populated by an equal number of presumably false TT matches. When the features of all three RT
410 dimensions were considered, the separation of true and false CSMs further improved (Fig. 3b). Again,
411 the distributions of TD *E. coli* CSMs and entrapment CSMs behaved similarly. Interestingly, few CSMs
412 that passed the 1% FDR threshold were located in regions dominated by false identifications. This
413 might identify them as part of the expectable fraction of 1% false positive identifications.
414 Importantly, the described separation was achieved unsupervised on RT features alone, i.e. without a
415 search engine score or target-decoy labels.

416 To test the transferability of our findings, we also ran xiRT with unfiltered pLink2 results (section S4
417 and Fig. S15). The prediction performance from Q-value filtered CSMs was similar to the results with
418 xiSEARCH (Fig. S15a-c). A two-sided t-test between hSAX, SCX and RP errors for TT and TDs revealed

significant differences in the respective error distributions of the pLink2 predictions (Fig. S15d). Importantly, the separation of true and false matches in two-dimensional space was also possible with pLink2 identifications (Fig. S15e). In summary, xiRT can learn retention times irrespective of the used search engine and the learned chromatographic features alone carry substantial information to separate true from false matches.

To investigate the relevance of multi-dimensional RT predictions for the identification of crosslinked peptides, we first supplemented each CSM with RT features. Then, we performed a semi-supervised rescoring and evaluated the trained SVM model using the SHAP framework. We chose to analyze SHAP values for the 15 most important features for TT observations (FDR > 1%) that were predicted to be a correct TT identification (Fig. 3c). This analysis revealed a similar magnitude for all 15 SHAP values implying that a single feature alone is insufficient to recognize false matches. Notably, the top 5 features contained features from RP, hSAX and SCX predictions which indicates that each chromatographic dimension carried relevant information for the rescoring. Because 11 of the 15 features were predictions considering only one of the two peptides and not directly derived from peptide-pairs, the predicted RTs displayed a larger error. This analysis suggests that an RT prediction model for linear peptides can add valuable information for crosslink analysis. In general, the model learned mostly that low errors in the RT dimensions indicate true positive identifications. Thus, the model implicitly learned that the RT of a crosslinked peptide should differ from the RT of the individual peptides. This might become useful especially for distinguishing consecutive[45] from crosslinked peptides or when dealing with gas-phase associated peptides[36].

Rescoring Crosslinked Peptides Enhances their Identification

Before computing a combined score, we compared the CSM scores based on mass spectrometric information (xiSCORE) and RT features (SVM score, Fig. 4a). Both scores largely agreed. Heteromeric CSMs passing 1% CSM-FDR yielded high SVM scores. Also, most target-decoy CSMs achieved a low SVM score (Fig. 4a, right) and a low xiSCORE (Fig. 4a, top). The SVM score distribution of the TDs matched closely the distribution of TTs in the low scoring area, which indicated that they still

modeled random TT matches and that overfitting was avoided. Interestingly, the TTs were overrepresented in the low scoring area for the xiSCORE but not for the SVM score, suggesting that true TTs remained hidden among the random matches when using xiSCORE alone. The broad SVM score distribution of TTs indicated that the rescoring process could be optimized. In conclusion, neither of the mass spectrometric information (xiSCORE) nor the RT information (SVM score) seem to reveal all true CSMs.

As a combination of both approaches should yield better results than either alone, we combined the SVM score with the xiSCORE. We evaluated the impact of rescoring CSMs on the number and quality of identified PPIs, as PPIs are typically the objective of large-scale crosslinking MS experiments. Heteromeric CSMs increased by 1.7-fold and heteromeric PPIs increased by 1.4-fold (Fig. 4b). Self-links increased only marginally in agreement with their smaller search space and accordingly lower random match frequency. Essentially, nearly all self-links were identified exhaustively based on mass spectrometric data alone. In contrast, RT information substantially improved the identification of heteromeric CSMs. Further gains might be possible by directly combining RT features with mass spectrometric features (and possibly also other) for supervised scoring.

Likely, the benefits of RT predictions for the rescoring depend on the data set and applied chromatographic separations. On the *E. coli* data, we therefore performed additional analyses where we limited the rescoring to only use a subset of the chromatographic dimensions (Tab. S5). The number of identified CSMs for heteromeric links increased from 724 in the reference to 902 (RP only), 977 (SCX-RP), 1092 (hSAX-RP) and 1199 (SCX-hSAX-RP). Likewise, PPIs increased from 109 to 135, 131, 157, 152, respectively (Tab. S5). As observed above, gains can be expected from each chromatographic dimension. When having to choose one ion chromatography, the hSAX dimension seemed more useful than the SCX dimension which could arise from the better prediction performance or more complex separation mechanisms. Importantly, even using RP RT alone already led to a marked gain in heteromeric PPIs (see also next section).

To systematically evaluate the additionally identified PPIs from all three RT dimensions, we compared them to the originally identified PPIs based exclusively on xiSCORE. In addition, the STRING/APID databases and a larger set of PPIs from a larger study[16] served as extra references for validation. Almost all PPIs found in the original dataset by xiSCORE were also contained in the rescored data set (91%). 85% of the newly identified PPIs were either found in the data set from Lenz *et al.,* in STRING/APID or both. Among the eight PPIs unique to the rescored data set, only one involved a human protein from the entrapment database (Fig. 4c). The remaining seven PPIs might constitute genuine PPIs. Note that the overall percentage of PPIs involving human proteins was reduced by rescoring. Since all human target proteins were included in the positive training data, this is an important indicator of a well-behaved model. Deepening trust further, almost all novel PPIs were identified with multiple CSMs (Fig. 4d). Finally, we selected the subnetwork of the RNA polymerase to investigate the additionally identified PPIs in a well-characterized interaction landscape (Fig. 4e). Indeed, all interactions added by RT-based rescoring were already reported in APID. In summary, all our evidence points at the successful complementation of MS information by RT, at least for a proteome-wide crosslinking analysis. It remained to be seen, however, if this could also be leveraged in more routine multiprotein complex analyses.

Multiprotein Complex Studies Also Benefit from the RT Prediction

Many crosslinking MS studies investigate multiprotein complexes and rely on only few chromatographic dimensions. We therefore evaluated the benefit of predicted RTs for the analysis of the FA-complex, an eight-membered multiprotein complex that was crosslinked using BS3. Here, the search engine score was supplemented exclusively with RP RT predictions during the rescoring. By using transfer learning, the small number of CSMs (692 unique CSMs, without considering charge states) found in this multiprotein complex analysis were sufficient to achieve accurate RP predictions (Fig. S10). The resulting crosslinks at 1% residue-pair FDR (lower levels set to 5%) showed an increase of 36 (+10%) self- and 53 (+70%) heteromeric residue-pairs. Importantly, the rescored links showed no indication of increased hits to the entrapment database (Fig. 5a) indicating that no overfitting

496 occurred during the rescoring. At the same time, heteromeric PPIs already identified before rescoring
497 received additional support. For example, the number and sequence coverage of links increased
498 between FAAP100 ("100") and FANCB ("B"), FANCA ("A") and FANCB, and FANCA and FANCG ("G").
499 Overall, the heteromeric links increased 1.7-fold with an even higher proportional increase in
500 'verified' links, i.e. fitting the available structure, by 1.9-fold (Fig. 5b). The derived distance
501 distribution of newly identified links is dissimilar from a random distribution and shows no
502 indications of reduced quality (Fig. 5c). Applying this 'structural validation' on its own might be
503 optimistic[69], however, in summary our rigorous quality control ensures trustworthy results. It is
504 currently unclear in how far even smaller datasets could benefit from xiRT. Generally, to improve
505 prediction performance, pre-training on larger data sets will lead to better generalization abilities of
506 the predictor. Subsequently, also smaller data sets can be used for accurate RT prediction. To
507 additionally benefit from sample specific information, increasing the cross-validation splits will utilize
508 larger parts of the data during training. In any case, our successful analysis of a multiprotein complex
509 supplemented with only RP features highlights the broad applicability of xiRT.

510 Conclusion

511 Using a Siamese network architecture, we succeeded in bringing RT prediction into the Crosslinking
512 MS field, independent of separation setup and search software. Our open source application xiRT
513 introduces the concept of multi-task learning to achieve multi-dimensional chromatographic
514 retention time prediction, and may use any peptide sequence-dependent measure including for
515 example collision cross section or isoelectric point. The black-box character of the neural network
516 was reduced by means of interpretable machine learning that revealed individual amino acid
517 contributions towards the separation behavior. The RT predictions – even when using only the RP
518 dimension – complement mass spectrometric information to enhance the identification of
519 heteromeric crosslinks in multiprotein complex and proteome-wide studies. Overfitting does not
520 account for this gain as known false target matches from an entrapment database did not increase.

Leveraging additional information sources may help to address the mass-spectrometric identification challenge of heteromeric crosslinks.

Acknowledgements

We thank Edward Rullmann, Andrea Graziadei and Francis O'Reilly for critical reading of the manuscript, and Jakub Bartoszewicz (RKI / HPI) for fruitful discussions. We are grateful to Tabea Schütze for help with fermenting *E. coli*.

Funding Sources

This work was supported by NVIDIA with hardware from the grant "Artificial Intelligence for Deep Structural Proteomics", by the Wellcome Trust through a Senior Research Fellowship to JR (103139) and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2008 - 390540038 – UniSysCat, and by grant no. 392923329/GRK2473. The Wellcome Centre for Cell Biology is supported by core funding from the Wellcome Trust (203149).

Additional Information

Abbreviations

FDR, false discovery rate; RT, Retention time; hSAX, hydrophilic strong anion exchange chromatography; SCX, strong cation exchange chromatography; RP, Reversed-phase chromatography; CV, cross-validation; acc, accuracy; racc, relaxed accuracy; TT, target-target; TD, target-decoy; DD, decoy-decoy; PPI, protein-protein interaction.

Supplementary Information

Identifications over the different gradients, details on the network architecture/the hyper-parameter optimization, machine learning results for linear peptide identifications, learning curves, RT error

542 characteristics, xiRT explainability analysis, results from combining pLink2 and xiRT are available in

543 the supplementary material.

544 ## Corresponding Author

545 * Juri Rappsilber: juri.rappsilber@tu-berlin.de

546 # References

547 1. O'Reilly, F. J. & Rappsilber, J. Cross-linking mass spectrometry: methods and applications in
548 structural, molecular and systems biology. *Nat. Struct. Mol. Biol.* **25**, 1 (2018).

549 2. Yu, C. & Huang, L. Cross-Linking Mass Spectrometry: An Emerging Technology for
550 Interactomics and Structural Biology. *Anal. Chem.* **90**, 144–165 (2018).

551 3. Leitner, A., Faini, M., Stengel, F. & Aebersold, R. Crosslinking and Mass Spectrometry: An
552 Integrated Technology to Understand the Structure and Function of Molecular Machines.
553 *Trends in Biochemical Sciences* (2016) doi:10.1016/j.tibs.2015.10.008.

554 4. Trnka, M. J., Baker, P. R., Robinson, P. J. J., Burlingame, a L. & Chalkley, R. J. Matching Cross-
555 linked Peptide Spectra: Only as Good as the Worse Identification. *Mol. Cell. Proteomics* **13**,
556 420–434 (2014).

557 5. Giese, S. H., Fischer, L. & Rappsilber, J. A Study into the Collision-induced Dissociation (CID)
558 Behavior of Cross-Linked Peptides. *Mol. Cell. Proteomics* **15**, 1094–1104 (2016).

559 6. Barysz, H. M. & Malmström, J. Development of Large-scale Cross-linking Mass Spectrometry.
560 *Molecular and Cellular Proteomics* (2018) doi:10.1074/mcp.R116.061663.

561 7. Rinner, O. *et al.* Identification of cross-linked peptides from large sequence databases. *Nat.*
562 *Methods* **5**, 315–8 (2008).

563 8. Chen, Z. A. *et al.* Architecture of the RNA polymerase II-TFIIF complex revealed by cross-
564 linking and mass spectrometry. *EMBO J.* **29**, 717–26 (2010).

565 9. Liu, F., Rijkers, D. T. S., Post, H. & Heck, A. J. R. Proteome-wide profiling of protein assemblies
566 by cross-linking mass spectrometry. *Nat. Methods* **12**, 1179–1184 (2015).

567 10. Schweppe, D. K. *et al.* Mitochondrial protein interactome elucidated by chemical cross-linking
568 mass spectrometry. *Proc. Natl. Acad. Sci.* **114**, 1732–1737 (2017).

569 11. Leitner, A., Walzthoeni, T. & Aebersold, R. Lysine-specific chemical cross-linking of protein
570 complexes and identification of cross-linking sites using LC-MS/MS and the xQuest/xProphet
571 software pipeline. *Nat. Protoc.* **9**, 120–137 (2014).

572 12. Mendes, M. L. *et al.* An integrated workflow for crosslinking mass spectrometry. *Mol. Syst.*
573 *Biol.* **15**, e8994 (2019).

574 13. Götze, M., Iacobucci, C., Ihling, C. H. & Sinz, A. A Simple Cross-Linking/Mass Spectrometry
575 Workflow for Studying System-wide Protein Interactions. *Anal. Chem.* **91**, 10236–10244
576 (2019).

577 14. Ryl, P. S. J. *et al.* In Situ Structural Restraints from Cross-Linking Mass Spectrometry in Human
578 Mitochondria. *J. Proteome Res.* **19**, 327–336 (2020).

15. O'Reilly, F. J. *et al.* In-cell architecture of an actively transcribing-translating expressome. *Science (80-.).* **369**, 554–557 (2020).

16. Lenz, S. *et al.* Reliable identification of protein-protein interactions by crosslinking mass spectrometry. *BioRxiv* 1–10 (2020) doi:10.1101/2020.05.25.114256.

17. Gonzalez-Lozano, M. A. *et al.* Stitching the synapse: Cross-linking mass spectrometry into resolving synaptic protein interactions. *Sci. Adv.* **6**, eaax5783 (2020).

18. The, M., MacCoss, M. J., Noble, W. S. & Käll, L. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. *J. Am. Soc. Mass Spectrom.* **27**, 1719–1727 (2016).

19. Granholm, V., Noble, W. S. & Käll, L. A cross-validation scheme for machine learning algorithms in shotgun proteomics. *BMC Bioinformatics* **13 Suppl 1**, S3 (2012).

20. Hoopmann, M. R. *et al.* Kojak: efficient analysis of chemically cross-linked protein complexes. *J. Proteome Res.* **14**, 2190–8 (2015).

21. Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. *Anal. Chem.* **74**, 5383–5392 (2002).

22. Ma, K., Vitek, O. & Nesvizhskii, A. I. A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet. *BMC Bioinformatics* **13**, S1 (2012).

23. Liu, F., Lössl, P., Scheltema, R., Viner, R. & Heck, A. J. R. Optimized fragmentation schemes and data analysis strategies for proteome-wide cross-link identification. *Nat. Commun.* (2017) doi:10.1038/ncomms15473.

24. 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).

25. Klammer, A. A., Yi, X., MacCoss, M. J. & Noble, W. S. Improving tandem mass spectrum identification using peptide retention time prediction across diverse chromatography conditions. *Anal. Chem.* **79**, 6111–8 (2007).

26. Dwivedi, R. C. *et al.* Practical implementation of 2D HPLC scheme with accurate peptide retention prediction in both dimensions for high-throughput bottom-up proteomics. *Anal. Chem.* **80**, 7036–42 (2008).

27. Krokhin, O. V. Sequence-Specific Retention Calculator. Algorithm for Peptide Retention Prediction in Ion-Pair RP-HPLC: Application to 300- and 100-Å Pore Size C18 Sorbents. *Anal. Chem.* **78**, 7785–7795 (2006).

28. Pfeifer, N., Leinenbach, A., Huber, C. G. & Kohlbacher, O. Improving peptide identification in proteome analysis by a two-dimensional retention time filtering approach. *J. Proteome Res.* **8**, 4109–15 (2009).

29. Giese, S. H., Ishihama, Y. & Rappsilber, J. Peptide Retention in Hydrophilic Strong Anion Exchange Chromatography Is Driven by Charged and Aromatic Residues. *Anal. Chem.* acs.analchem.7b05157 (2018) doi:10.1021/acs.analchem.7b05157.

30. Alpert, A. J. *et al.* Peptide orientation affects selectivity in ion-exchange chromatography. *Anal. Chem.* **82**, 5253–9 (2010).

31. Yeung, D., Klaassen, N., Mizero, B., Spicer, V. & Krokhin, O. V. Peptide retention time prediction in hydrophilic interaction liquid chromatography: Zwitter-ionic sulfoalkylbetaine and phosphorylcholine stationary phases. *J. Chromatogr. A* (2020)

622 doi:10.1016/j.chroma.2020.460909.

623 32. Ba, L. J. & Caruana, R. Do Deep Nets Really Need to be Deep? *Nature* **521**, 436–444 (2013).

624 33. Tran, N. H., Zhang, X., Xin, L., Shan, B. & Li, M. De novo peptide sequencing by deep learning.
625 *Proc. Natl. Acad. Sci. U. S. A.* (2017) doi:10.1073/pnas.1705691114.

626 34. Ma, C. *et al.* Improved Peptide Retention Time Prediction in Liquid Chromatography through
627 Deep Learning. *Anal. Chem.* **90**, 10881–10888 (2018).

628 35. Gessulat, S. *et al.* Prosit: proteome-wide prediction of peptide tandem mass spectra by deep
629 learning. *Nat. Methods* **16**, 509–518 (2019).

630 36. Giese, S. H., Belsom, A., Sinn, L., Fischer, L. & Rappsilber, J. Noncovalently Associated Peptides
631 Observed during Liquid Chromatography-Mass Spectrometry and Their Affect on Cross-Link
632 Analyses. *Anal. Chem.* **91**, 2678–2685 (2019).

633 37. Giese, S. H., Belsom, A. & Rappsilber, J. Optimized fragmentation regime for diazirine photo-
634 cross-linked peptides. *Anal. Chem.* **88**, 8239–8247 (2016).

635 38. Liu, F., Lössl, P., Scheltema, R., Viner, R. & Heck, A. J. R. Optimized fragmentation schemes and
636 data analysis strategies for proteome-wide cross-link identification. *Nat. Commun.* **8**, 15473
637 (2017).

638 39. Rappsilber, J., Ishihama, Y. & Mann, M. Stop and go extraction tips for matrix-assisted laser
639 desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics.
640 *Anal. Chem.* **75**, 663–70 (2003).

641 40. Kessner, D., Chambers, M., Burke, R., Agus, D. & Mallick, P. ProteoWizard: open source
642 software for rapid proteomics tools development. *Bioinformatics* **24**, 2534–6 (2008).

643 41. Eng, J. K. *et al.* A Deeper Look into Comet - Implementation and Features. *J. Am. Soc. Mass*
644 *Spectrom.* (2015) doi:10.1007/s13361-015-1179-x.

645 42. Lenz, S., Giese, S. H., Fischer, L. & Rappsilber, J. In-Search Assignment of Monoisotopic Peaks
646 Improves the Identification of Cross-Linked Peptides. *J. Proteome Res.* **17**, 3923–3931 (2018).

647 43. Koster, J. & Rahmann, S. Snakemake--a scalable bioinformatics workflow engine.
648 *Bioinformatics* **28**, 2520–2522 (2012).

649 44. Fischer, L. & Rappsilber, J. Quirks of Error Estimation in Cross-Linking/Mass Spectrometry.
650 *Anal. Chem.* **89**, 3829–3833 (2017).

651 45. Iacobucci, C. & Sinz, A. To Be or Not to Be? Five Guidelines to Avoid Misassignments in Cross-
652 Linking/Mass Spectrometry. *Anal. Chem.* **89**, 7832–7835 (2017).

653 46. Alonso-López, Di. *et al.* APID database: Redefining protein-protein interaction experimental
654 evidences and binary interactomes. *Database* **2019**, 1–8 (2019).

655 47. Szklarczyk, D. *et al.* STRING v11: Protein-protein association networks with increased
656 coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic*
657 *Acids Res.* (2019) doi:10.1093/nar/gky1131.

658 48. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad.*
659 *Sci. U. S. A.* (2003) doi:10.1073/pnas.1530509100.

660 49. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search
661 tool. *J. Mol. Biol.* (1990) doi:10.1016/S0022-2836(05)80360-2.

662 50. Shakeel, S. *et al.* Structure of the Fanconi anaemia monoubiquitin ligase complex. *Nature* **575**,

234–237 (2019).

51. Farrell, D. P. *et al.* Deep learning enables the atomic structure determination of the Fanconi Anemia core complex from cryoEM. *IUCrJ* **7**, 881–892 (2020).

52. farrell, daniel. Deep learning enables the atomic structure determination of the Fanconi Anemia core complex from cryoEM. (2020) doi:10.5281/ZENODO.3998806.

53. Graham, M. J., Combe, C., Kolbowski, L. & Rappsilber, J. xiView: A common platform for the downstream analysis of Crosslinking Mass Spectrometry data. *bioRxiv* (2019) doi:10.1101/561829.

54. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

55. Abadi, M. *et al.* TensorFlow: A system for large-scale machine learning. in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016* (2016).

56. Cheng, J., Wang, Z. & Pollastri, G. A neural network approach to ordinal regression. in *Proceedings of the International Joint Conference on Neural Networks* (2008). doi:10.1109/IJCNN.2008.4633963.

57. Berrar, D. Cross-validation. in *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics* (2018). doi:10.1016/B978-0-12-809633-8.20349-X.

58. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

59. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* (2002) doi:10.1613/jair.953.

60. Lundberg, S. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Nips* **16**, 426–430 (2017).

61. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3**, 861 (2018).

62. Okuda, S. *et al.* JPOSTrepo: An international standard data repository for proteomes. *Nucleic Acids Res.* (2017) doi:10.1093/nar/gkw1080.

63. Walzthoeni, T. *et al.* False discovery rate estimation for cross-linked peptides identified by mass spectrometry. *Nat. Methods* **9**, 901–903 (2012).

64. Xu, C. & Ma, B. Software for computational peptide identification from MS-MS data. *Drug Discovery Today* (2006) doi:10.1016/j.drudis.2006.05.011.

65. Yılmaz, Ş. *et al.* Cross-linked peptide identification: A computational forest of algorithms. *Mass Spectrom. Rev.* **37**, 738–749 (2018).

66. Ruder, S. An Overview of Multi-Task Learning in Deep Neural Networks. *arXiv* 1706.05098 (2017).

67. Gussakovsky, D., Neustaeter, H., Spicer, V. & Krokhin, O. V. Sequence-Specific Model for Peptide Retention Time Prediction in Strong Cation Exchange Chromatography. *Anal. Chem.* **89**, 11795–11802 (2017).

68. Guo, D., Mant, C. T., Taneja, A. K., Parker, J. M. R. & Rodges, R. S. Prediction of peptide retention times in reversed-phase high-performance liquid chromatography I. Determination of retention coefficients of amino acid residues of model synthetic peptides. *J. Chromatogr. A* (1986) doi:10.1016/0021-9673(86)80102-9.

705 69. Yugandhar, K., Wang, T. Y., Wierbowski, S. D., Shayhidin, E. E. & Yu, H. Structure-based
706 validation can drastically underestimate error rate in proteome-wide cross-linking mass
707 spectrometry studies. *Nat. Methods* (2020) doi:10.1038/s41592-020-0959-9.

708

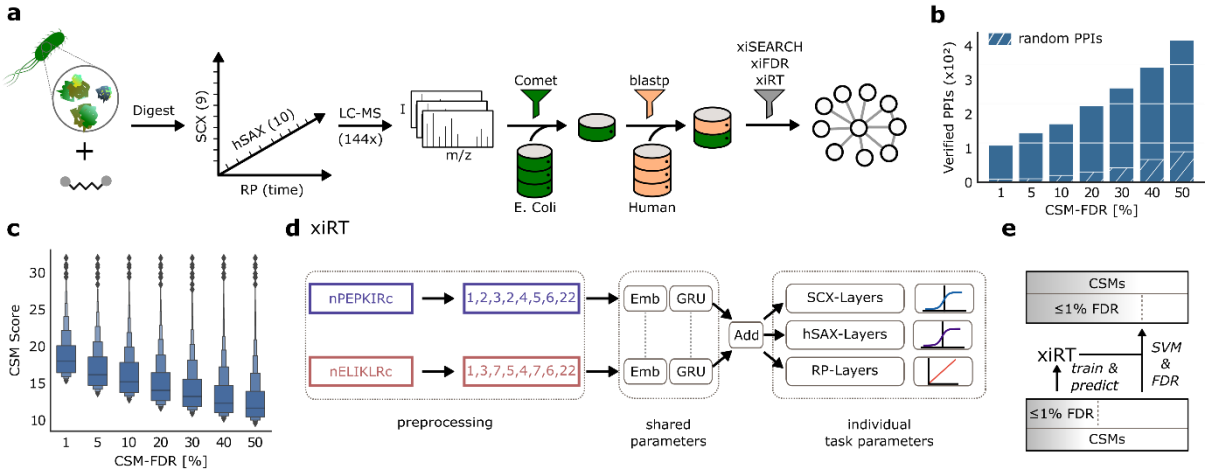


Figure 1: Workflow overview. a) Experimental and data analysis workflow. The soluble high-molecular weight proteome of *E. coli* lysate was crosslinked with disuccinimidyl suberate (DSS) and the digest sequentially fractionated by SCX (9 fractions collected), hSAX (10 pools collected) and finally by RP coupled to the MS. The protein database for the crosslink search was created by a linear peptide search with Comet and a sequence-based filter using BLAST. For each *E. coli* protein in the final database a human protein was added as a control. b) Potential for false negative PPI identifications. Verified PPIs are estimated from matches to the STRING/APID databases. PPIs are computed based on CSM-level FDR. Estimated random hits correspond to the average number of semi-randomly drawn pairs (first protein was randomly selected from the STRING/APID DB and second protein was drawn from the FASTA file). Gained PPIs accentuate the additional information that is available in the data at higher FDR. c) Decrease of CSM scores based on spectral evidence with increased FDRs. Boxenplot shows the median and 50% of the data in the central box. d) xiRT network architecture to predict multi-dimensional retention times. A crosslinked peptide is represented as two individual inputs to xiRT. xiRT uses a Siamese network architecture that shares the weights of the embedding and recurrent layers. Individual layers for the prediction tasks are added with custom activation functions (sigmoid / linear functions for fractionation / regression tasks, respectively). e) Rescoring workflow. The predictions from xiRT are combined with xiSCORE's output to rescore CSMs using a linear SVM, consequently leading to more matches at constant confidence.

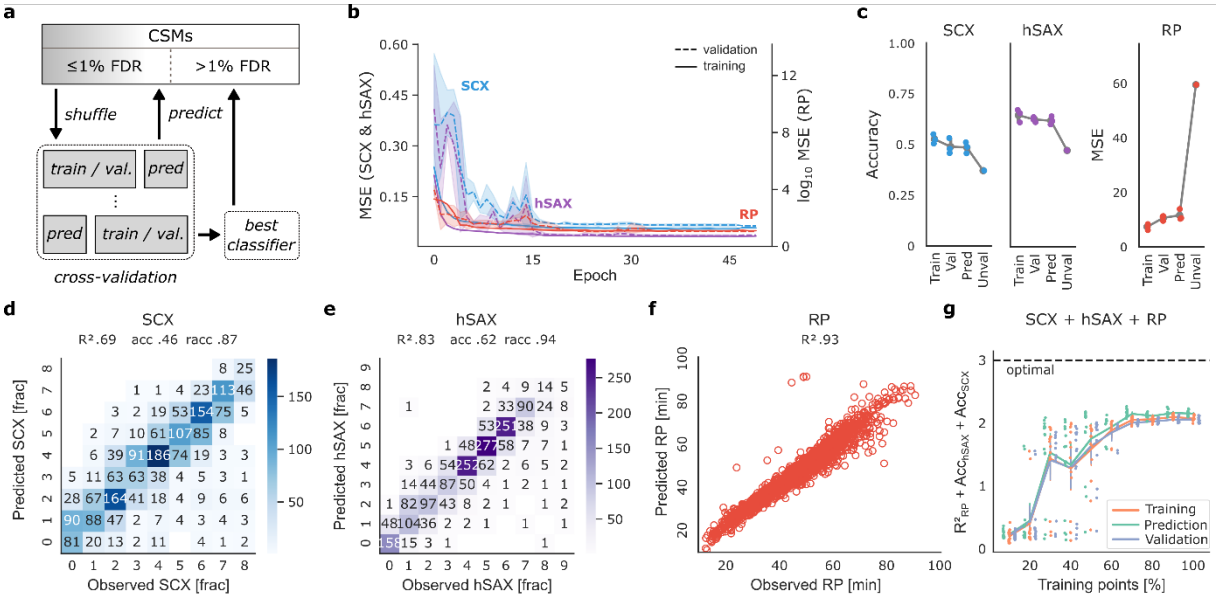
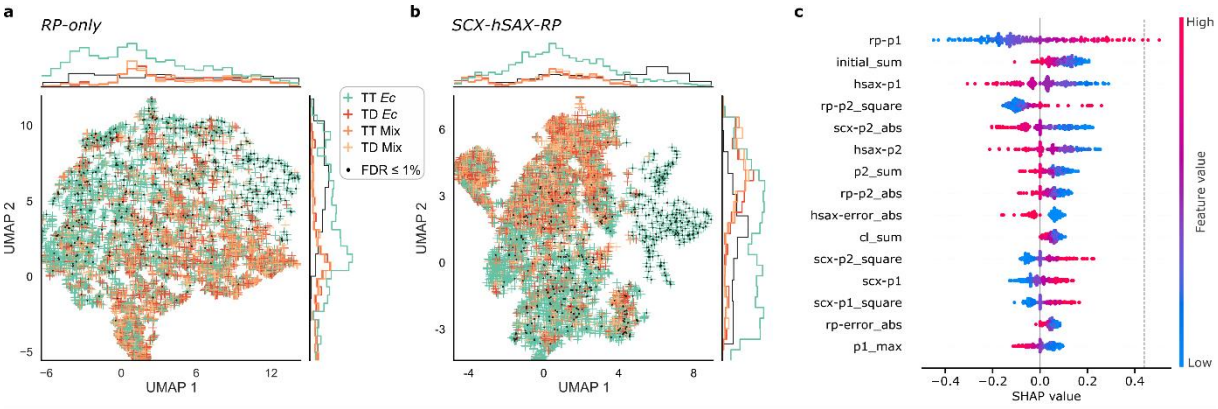


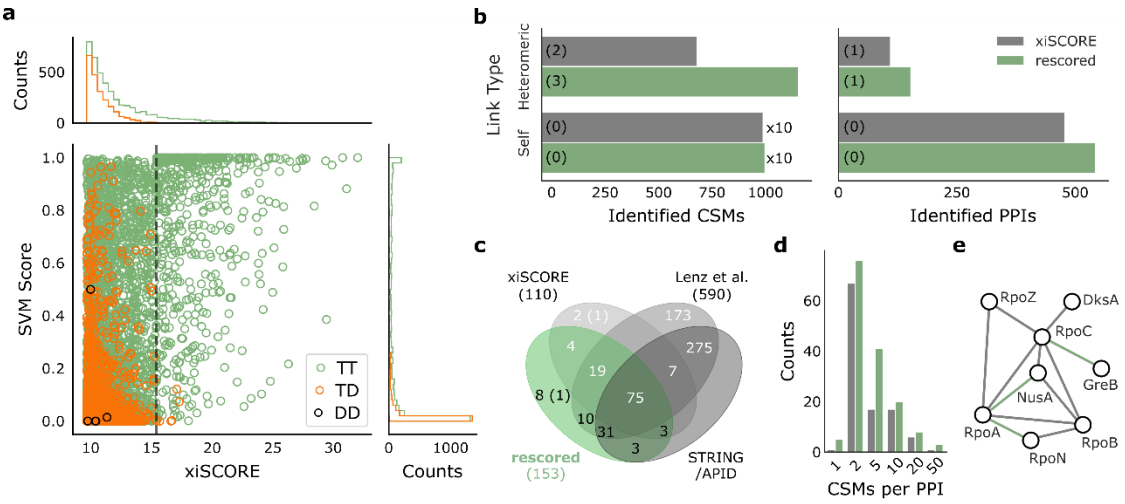
Figure 2: Cross-validation of retention time prediction. a) Applied cross-validation strategy in xiRT. To predict the RTs of CSMs excluded from training, the best CV classifier is used. b) xiRT performance over training epochs. Shaded areas show the estimated 95% confidence interval. c) xiRT performance across different metrics (error bars show standard deviation). Prediction for the 'unvalidated' data was only performed once. d-f) Prediction results from a representative CV iteration for SCX, hSAX and RP at 1% CSM-FDR. g) Learning curve with increasing number of CSMs, e.g. 10% (645 total CSMs, 387 for training, 43 for validation, 215 for prediction), 50% (3226, 1935, 216, 1075), 100% (6453, 3871, 431, 2151); bars indicate standard deviation.



737

738 **Figure 3: Visualization of RT features.** a) xiRT-based features from RP dimension only (13 features)
739 after dimensionality reduction with UMAP. b) xiRT-based feature from SCX-hSAX-RP dimensions (43
740 features) after dimensionality reduction with UMAP. Input data for a) and b) were CSMs of
741 heteromeric links in the proteome-wide crosslinking dataset (Ec = E. coli, Mix = match between E.
742 Coli and human peptides), filtered to 50% CSM-FDR. Identifications passing 1% CSM-FDR are
743 highlighted. DD identifications are not shown. c) SHAP analysis of RT feature importance for CSM-
744 rescoring (using a linear SVM) including SCX, hSAX and RP features (Tab. S4). Each dot represents a
745 previously identified CSM from 200 randomly chosen TTs that were excluded from training (i.e. CSM-
746 FDR > 1%). The background data set consists of 100 TT and TD CSMs each. Dashed line indicates the
747 base value for a prediction based on the background data alone (0.44).

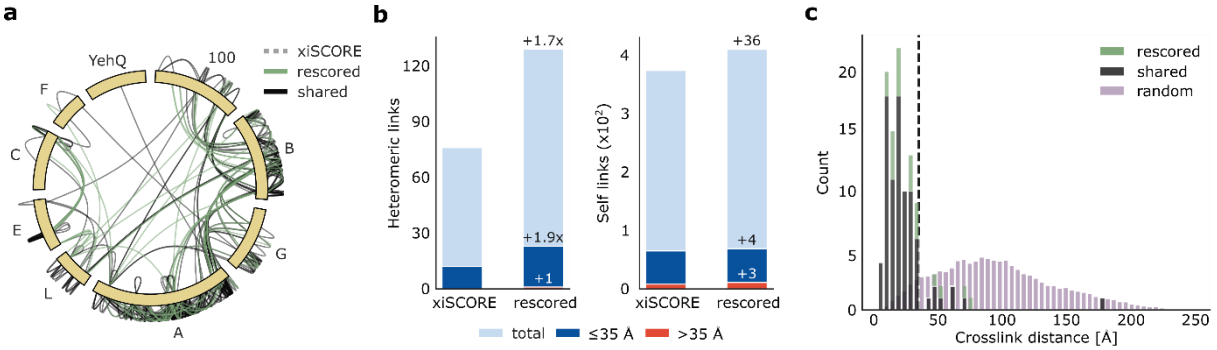
748



749

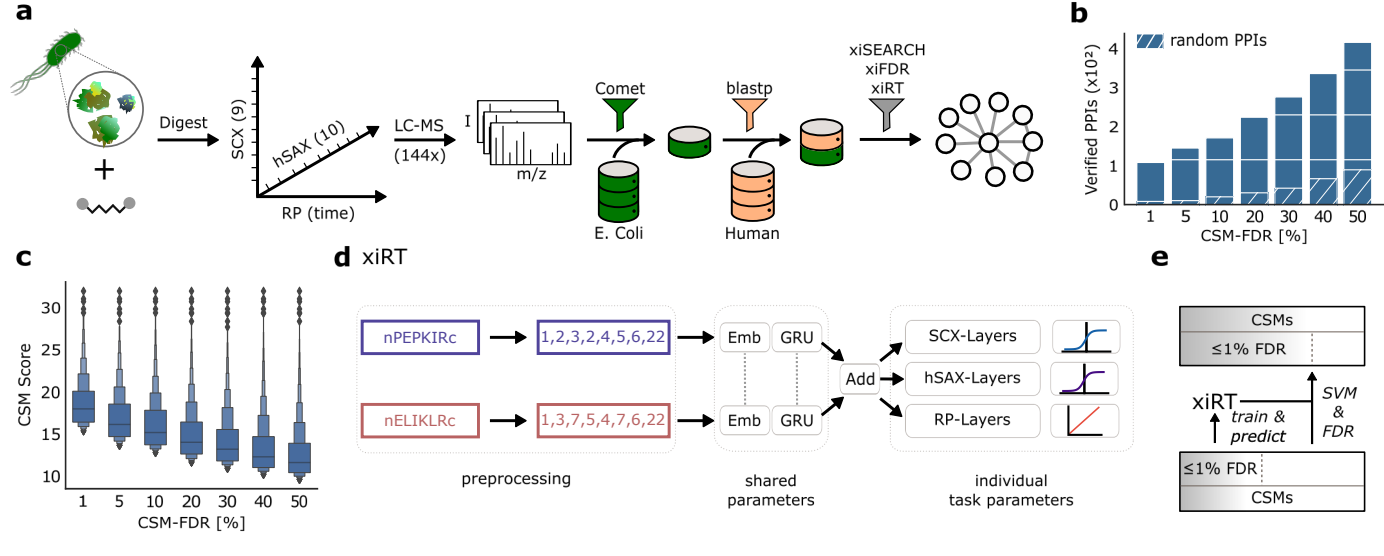
750 **Figure 4: Incorporation of RT prediction to CSM-scoring increases crosslink identification**. a) Score
751 distributions of heteromeric CSMs based on mass spectrometric information (xiSCORE) and RT
752 features (SVM score). The dashed line indicates the xiSCORE-based CSM-FDR threshold of 1%. b)
753 Increase in identification of TT-CSMs and PPIs at constant FDR. Numbers in brackets indicate
754 identifications involving a human protein. c) Overlap of observed PPIs (at 1% heteromeric PPI-FDR) to
755 external references. Numbers in the Venn diagram represent the identified PPIs among *E. coli*
756 proteins or PPIs involving human proteins (in brackets). Black numbers highlight the added benefit
757 from combining xiSCORE with xiRT's SVM score for PPI identification. d) Distribution of CSMs per PPI
758 before (grey) and after CSM-rescoring (green). e) Selected subnetwork of the RNA polymerase with
759 PPIs only identified after the rescoring connected in green. Data in b-e corresponds to a 1% PPI-FDR
760 (prefiltered at a 5% CSM-FDR).

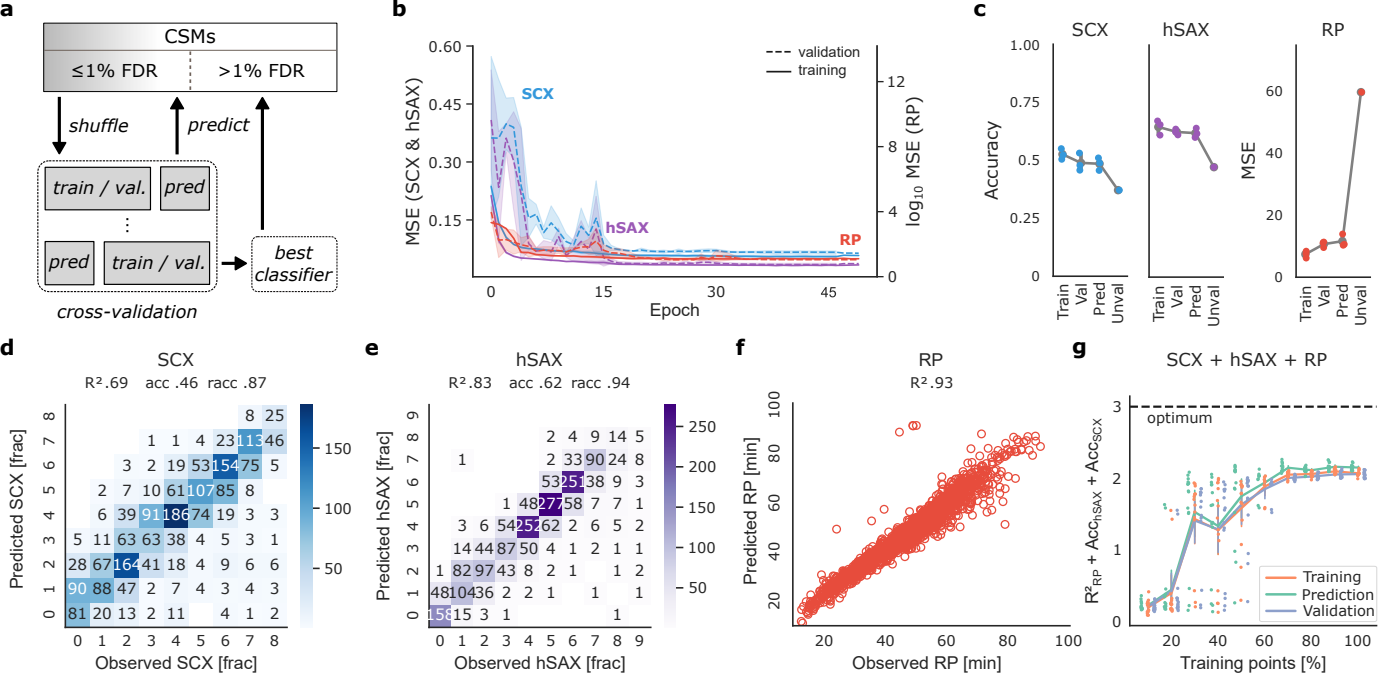
761

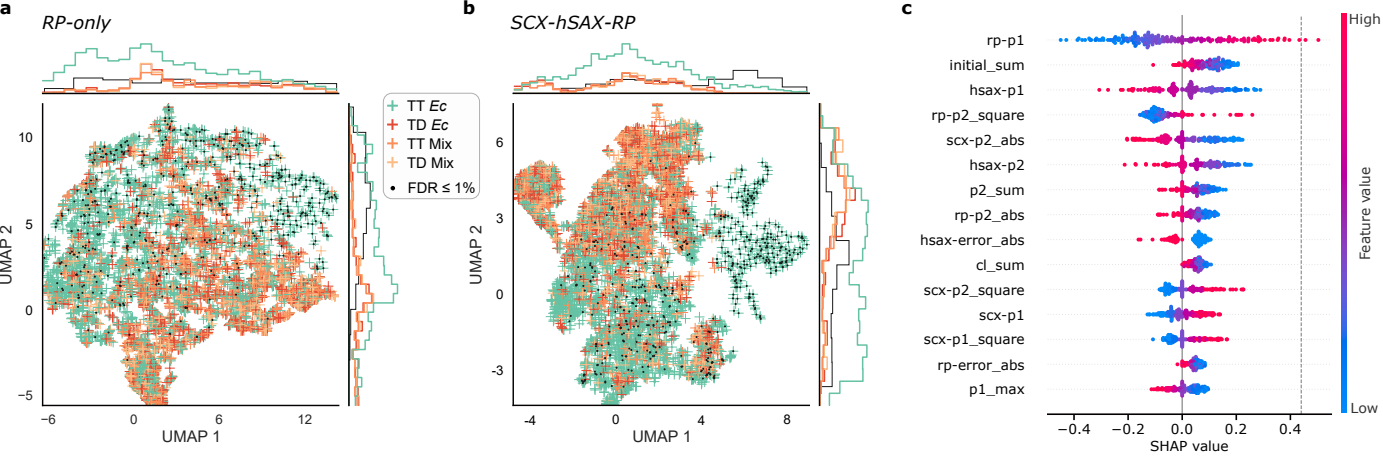


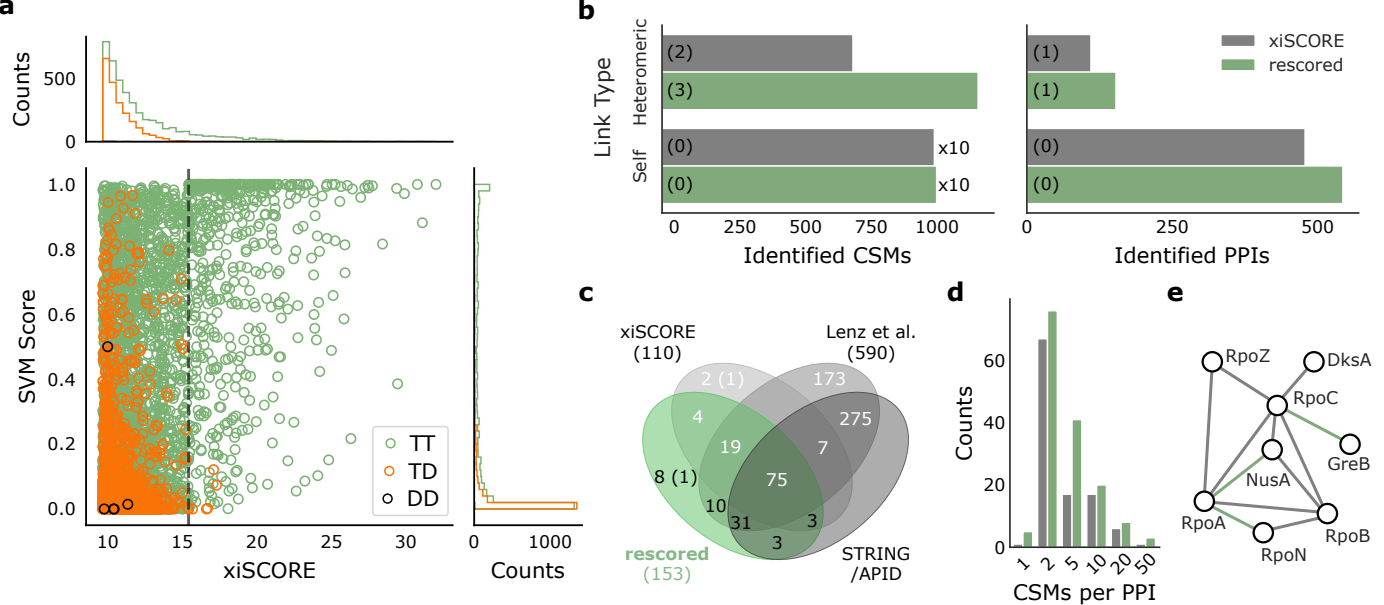
762

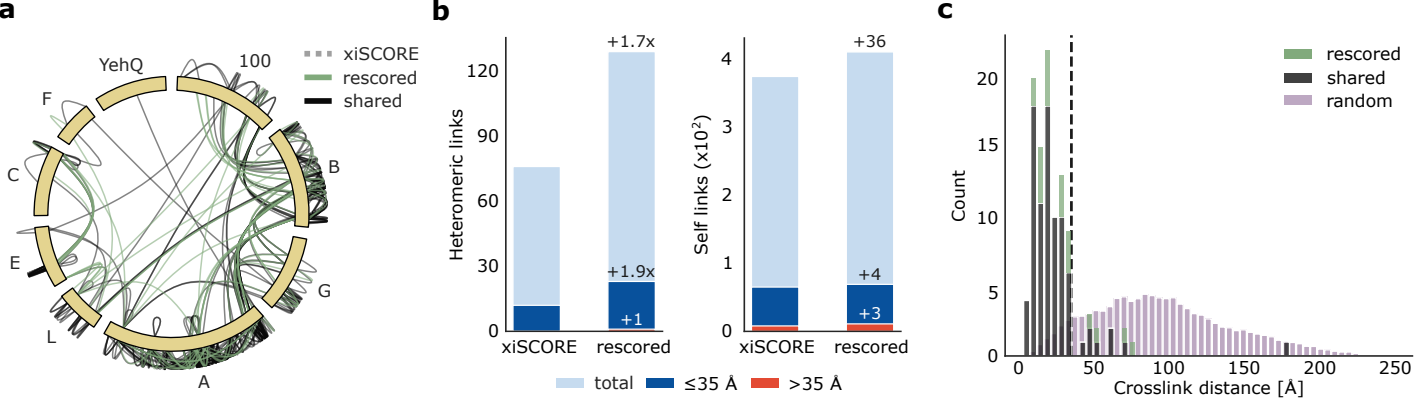
763 **Figure 5: Benefit of RT prediction for multiprotein complex crosslink analysis.** a) Crosslink network
764 from the Fanconi anemia complex analysis, shown in circular view. Unique residue pairs from
765 xiSCORE, after rescoring and shared between these analyses are depicted (1% residue-pair FDR).
766 Proteins associated to the Fanconi anemia core complex are indicated with their gene name suffix.
767 The *E. coli* protein YehQ represents a match from the entrapment database. b) Quantitative
768 assessment of residue-pairs with and without rescoring, and including calculated distances in the
769 model. c) Distribution of crosslink distances from identified residue-pairs (n=105) following rescoring,
770 shared between rescoring and xiSCORE (since no crosslinks unique to xiSCORE), and theoretically
771 possible residue-pairs that could be mapped to the model.


a
Digest
SCX (9)
hSAX (10)
RP (time)
LC-MS
(144x)
I
m/z
Comet
E. Coli
blastp
Human
xiSEARCH
xiFDR
xiRT
b
random PPIs
Verified PPIs (x10²)
CSM-FDR [%]
c
CSM Score
CSM-FDR [%]
d xiRT
nPEPKIRc
1,2,3,2,4,5,6,22
nELIKLRc
1,3,7,5,4,7,6,22
Emb
GRU
Add
SCX-Layers
hSAX-Layers
RP-Layers
preprocessing
shared parameters
individual task parameters
e
CSMs
≤1% FDR
xiRT
train & predict
SVM & FDR


a
CSMs
≤1% FDR
>1% FDR
shuffle
predict
train / val.
pred
pred
train / val.
best classifier
cross-validation
b
validation
training
SCX
hSAX
RP
MSE (SCX & hSAX)
log10 MSE (RP)
Epoch
c
SCX
hSAX
RP
Accuracy
MSE
Train
Val
Pred
Unval
d
SCX
R².69 acc .46 racc .87
Predicted SCX [frac]
Observed SCX [frac]
e
hSAX
R².83 acc .62 racc .94
Predicted hSAX [frac]
Observed hSAX [frac]
f
RP
R².93
Predicted RP [min]
Observed RP [min]
g
SCX + hSAX + RP
optimum
Training
Prediction
Validation
Training points [%]


a
RP-only
b
SCX-hSAX-RP
TT Ec
TD Ec
TT Mix
TD Mix
FDR ≤ 1%
UMAP 1
UMAP 2
c
rp-p1
initial_sum
hsax-p1
rp-p2_square
scx-p2_abs
hsax-p2
p2_sum
rp-p2_abs
hsax-error_abs
cl_sum
scx-p2_square
scx-p1
scx-p1_square
rp-error_abs
p1_max
SHAP value
Feature value
High
Low


a
Counts
500
0
SVM Score
1.0
0.8
0.6
0.4
0.2
0.0
10
15
20
25
30
xiSCORE
TT
TD
DD
0
1000
Counts
b
Link Type
Heteromeric
Self
(2)
(3)
(0)
(0)
x10
x10
0
250
500
750
1000
Identified CSMs
xiSCORE
rescored
(1)
(1)
(0)
(0)
0
250
500
Identified PPIs
c
xiSCORE
(110)
Lenz et al.
(590)
2 (1)
173
4
275
19
7
8 (1)
75
10
31
3
3
rescored
(153)
STRING
/APID
d
Counts
60
40
20
0
1
2
5
10
20
50
CSMs per PPI
e
RpoZ
DksA
RpoC
GreB
NusA
RpoA
RpoB
RpoN


a
YehQ
100
xiSCORE
rescored
shared
F
B
C
E
G
L
A
b
Heteromeric links
+1.7x
+1.9x
+1
xiSCORE
rescored
Self links (x10²)
+36
+4
+3
total
≤35 Å
>35 Å
c
rescored
shared
random
Count
Crosslink distance [Å]