Data analysis strategies for proteome-wide crosslinking mass spectrometry

vorgelegt von

M. Sc.

Swantje Lenz

ORCID: 0000-0002-8839-5371

an der Fakultät III - Prozesswissenschaften

der Technischen Universität Berlin

zur Erlangung des akademischen Grades

Doktorin der Naturwissenschaften

- Dr. rer. nat. -

genehmigte Dissertation

Promotionsausschuss:
Vorsitzende: Sina Bartfeld
Gutachter: Bernhard Renard
Gutachter: Juri Rappsilber

Tag der wissenschaftlichen Aussprache: 24. Mai 2022

Berlin 2023

Declaration of Authorship

I, Swantje Lenz, declare that this thesis titled, “Data analysis foundation of proteome-wide crosslinking mass spectrometry” and the work presented in it are my own.

I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. Apart from such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signature

Date

Table of contents

Abstract

As enzymes, building blocks or messengers, proteins are essential molecules in life. They often function together with other proteins in complexes or in transient interactions. A crucial part of understanding their function is knowing their structure and their interaction partners.

Crosslinking mass spectrometry (crosslinking MS) has by now been established as a method to study protein interactions and structures by delivering medium-resolution inter-residue distances. Initially, crosslinking MS studies were limited to single proteins or complexes, but the technology has the potential to be used in more complex samples, with the goal to detect protein-protein interactions at a proteome-wide scale. To realise this, the technology requires development and optimization of multiple steps of the workflow.

In this thesis, I focus on the data analysis of crosslinking MS data at different points of the workflow. The work of this thesis increased the number of identifications during database search and provides the groundwork for further optimisation of the crosslinking MS workflow. It demonstrates data-driven evaluation of experimental tests and provides a reliable procedure for error estimation.

First, I show that for crosslinked peptides, due to their low abundance and large size, the monoisotopic precursor mass is often misassigned by the mass spectrometer software. We implemented a solution into our database search, where multiple masses are searched. This increased the number of crosslinked identifications significantly.

Another important factor in MS acquisition is the fragmentation of crosslinked peptides. I therefore analysed the fragmentation behaviour of the MS-cleavable crosslinker DSSO, which is commonly used for large-scale crosslinking MS studies. We analyse commonly used workflows regarding the peptide fragmentation and utilisation of the characteristic peaks during database search. This showed that the advantage of MS-cleavable crosslinkers lies in the improved fragmentation and showed that some workflows are suboptimal in their speed.

Finally, we use a controlled sample of *E. coli* lysate to demonstrate a reliable procedure to estimate the error of crosslinked PPIs. The study was set up to allow for an experimental control of the error. With this and three other controls we show that for a reliable error estimation in crosslinked PPIs, the FDR needs to be calculated separately for self and heteromeric matches and on the PPI-level. This error estimation was applied to our *E. coli* lysate and provided a reliable network of protein-protein interaction. Here, we found an unknown binder to RNA polymerase which we map to its binding site with use of the structural information of the crosslinks.

Overall, the results of this work allowed us to use crosslinking MS on the scale of proteome-wide, in-cell studies. The next challenge will be increasing the depth to allow detection of low abundant proteins, which will require further optimisation of crosslinking MS.

Zusammenfassung

In Form von Enzymen, Bausteinen oder Botenstoffen agieren Proteine als wesentliche Moleküle des Lebens. Gemeinsam mit anderen Proteinen wirken sie oft als Protein-Komplexe oder in kurzlebigen Interaktionen. Dabei ist die Kenntnis ihrer Struktur und Interaktionspartner für das Verständnis ihrer Funktion entscheidend.

Als Methode zur Untersuchung von Proteininteraktionen und -strukturen hat sich die Crosslinking-Massenspektrometrie (Crosslinking MS) etabliert, da sie Distanzen zwischen Aminosäureresten in mittlerer Auflösung liefert. Ursprünglich waren Crosslinking MS Studien auf einzelne Proteine oder Komplexe beschränkt, aber die Technologie hat das Potenzial, in komplexeren Proben eingesetzt zu werden und Protein-Protein-Interaktionen auf einem proteomweiten Maßstab zu erkennen. Um dieses Potenzial auszuschöpfen, muss die Technologie in mehreren Schritten des Arbeitsablaufs weiterentwickelt und optimiert werden.

In dieser Arbeit konzentriere ich mich auf die Datenanalyse von Crosslinking-MS-Daten an verschiedenen Stellen des Arbeitsablaufs. Die Arbeit hat die Zahl der Identifizierungen der Datenbanksuche erhöht und liefert die Grundlage für die weitere Optimierung des Arbeitsablaufs bei der Crosslinking MS. Sie demonstriert die datengesteuerte Auswertung experimenteller Tests und liefert ein zuverlässiges Verfahren zur Fehlerabschätzung.

Zunächst zeige ich, dass bei vernetzten Peptiden aufgrund ihrer Größe und geringen Abundanz die monoisotopische Vorläufermasse von der Software des Massenspektrometers oft falsch zugeordnet wird. Wir haben eine Lösung in unsere Datenbanksuche implementiert, bei der nach mehreren Massen gesucht wird. Dadurch kann die Zahl der vernetzten Identifizierungen erheblich gesteigert werden.

Ein weiterer wichtiger Faktor bei der MS Akquisition ist die Fragmentierung von vernetzten Peptiden. Daher haben wir das Fragmentierungsverhalten des MS-spaltbaren Crosslinkers DSSO analysiert, der üblicherweise für groß angelegte Crosslinking MS Studien verwendet wird. Wir analysierten gängige Arbeitsabläufe hinsichtlich der Peptidfragmentierung und der Nutzung der charakteristischen Peaks bei der Datenbanksuche. Dabei zeigte sich, dass der Vorteil von MS-spaltbaren Crosslinkern in der verbesserten Fragmentierung liegt und dass einige Arbeitsabläufe in ihrer Geschwindigkeit suboptimal sind.

Schließlich demonstrieren wir anhand einer kontrollierten Probe von *E. coli* Lysat ein zuverlässiges Verfahren zur Abschätzung des Fehlers von vernetzten PPIs. Die Studie ermöglicht uns eine experimentelle Kontrolle des Fehlers. Mit dieser und drei weiteren Kontrollen zeigen wir, dass für eine zuverlässige Fehlerabschätzung für vernetzte PPIs die FDR separat für Selbst- und Heteromere PPIs und auf PPI-Ebene berechnet werden muss. Diese Fehlerabschätzung wird auf unser *E. coli* Lysat angewandt und liefert ein zuverlässiges PPI-Netzwerk. Wir haben einen unbekannten Binder für die RNA-Polymerase gefunden, den wir mit Hilfe der Strukturinformationen der Crosslinks auf seine Bindungsstelle abbilden.

Insgesamt ermöglichen uns die Ergebnisse dieser Arbeit, Crosslinking MS auf die Ebene von proteomweiten, zellinternen Studien zu bringen. Die nächste Herausforderung besteht darin, die Tiefe der Analyse zu erweitern, um auch Proteine mit geringer Abundanz nachzuweisen, wofür weitere Optimierungen der Crosslinking MS erforderlich sind.

Introduction

Crosslinking MS as a technology to investigate protein-protein interactions

Proteins are one of the most important molecules in biology and are therefore a focus of molecular studies. Among the wide range of functions they can perform, they can work as catalysts, information messengers and structural units. They often do not function alone, but as part of multi-protein complexes with specific structures and dynamics fine-tuned by evolution. Identifying the members of these complexes and their three-dimensional architecture is important to understand their function. Until recently, this had to be done by purification and reconstitution outside of the cellular context, but recent technological developments are now allowing for *in situ* studies (Böhning & Bharat, 2021).

Traditional methods to study these protein-protein interactions (PPIs) are for example two-hybrid techniques (Parrish et al., 2006) or, using proteomics, affinity-purification MS (Rigaut et al., 1999) or co-fractionation MS (Kristensen et al., 2012). These techniques require genetic tagging of the protein coding genes, lysis of the cells, or purification of the proteins and therefore take the proteins outside of their natural context prior to analysis. This can introduce experimental biases into the results, for example losing transient or weak interactions by the lysis or washing steps.

In the past years, crosslinking MS has emerged as an alternative tool to study PPIs (O'Reilly & Rappsilber, 2018). Crosslinking introduces covalent bonds between amino acid residues of the proteins. Followed up by mass spectrometric analysis, the linked peptides, and thereby the linkage positions, can be determined. Since the crosslinker reagent has a known length, this leads to a known distance restraint between the two residues. In contrast to the aforementioned methods, crosslinking MS can be performed *in situ* and therefore can generate information on the protein interactions inside intact cells, including transient or weak interactions that are often lost during cell disruption.

Crosslinks between two proteins inform that they were close in space during the crosslinking reaction and therefore are very likely to be directly interacting. Since crosslinking delivers information on residue-level, it also provides structural information about binding sites of these interactions. These can be used as distance restraints for further structural studies, for example in integrative modelling in combination with electron-microscopy densities (Robinson et al., 2015; von Appen et al., 2015). Therefore, crosslinking MS is a versatile tool that can be used in both focused structural biology studies and also systems-wide studies to map PPIs and their topologies.

Prior to the work of this thesis, crosslinking MS experiments tended to be performed on single proteins or complexes as proof-of-concept studies or providing extra information to structural studies (Z. A. Chen et al., 2010; Lasker et al., 2012). In recent years, crosslinking MS has developed into a technology that can handle the complexity of proteome-wide protein-protein interaction networks (Chavez et al., 2018; Liu et al., 2018). To analyse these very complex protein mixtures, multiple adjustments and optimization steps in the workflow were needed. The focus of my PhD was to develop data analysis approaches to allow crosslinking MS to handle the complexity of a whole crosslinked cell. The advances that I describe here

have underpinned several biological discoveries and will underpin many more in the future as this technology becomes mainstream.

The crosslinking MS workflow

During bottom-up proteomics, whole proteins are digested to peptides, which are subsequently identified by mass spectrometry. Crosslinked peptides have certain properties that make their identification by mass spectrometry more challenging than for linear peptides. For one, they are always much less abundant than the background of linear peptides in digestion mixtures, which hinders their detection in the MS. In addition to this experimental challenge, identifying the crosslinked peptide is more complex since two peptides have to be matched successfully from the same fragmentation spectrum. Both of these problems are intensified as the sample complexity is growing to larger scales. During the past years, all steps of the workflow were further optimised to meet those challenges. Briefly, the experimental workflow has five distinct steps that have each undergone significant optimisations by many labs over the years.

Crosslinking reaction. As a first step, the crosslinker is added to the sample, for example a purified protein complex or, for *in situ* experiments, to cells. The reagents consist of two reactive groups, which react with amino acids, separated by a spacer (Belsom & Rappsilber, 2021). Commonly NHS-ester based crosslinker reagents are used, which react primarily with lysines or n-termini, as well as serine, threonine or tyrosine in a side reaction. These include for example disuccinimidyl suberate (DSS) or bis(sulfosuccinimidyl)suberate (BS3), crosslinkers with two NHS-ester groups. Another popular crosslinker succinimidyl 4,4-azipentanoate (SDA) consists of one NHS-group and a diazirine group as the second reactive group. The diazirine is activated by UV-light, resulting in a carbene intermediate which can rapidly react with any amino acid. In addition, crosslinker spacers have been modified to be cleavable in the MS, for example disuccinimidyl sulfoxide (DSSO) (Kao et al., 2011) or DSBU (Müller et al., 2010). Other crosslinkers contain an enrichment group on the spacer (Steigenberger et al., 2019; Tang & Bruce, 2010) to thereby allow targeted enrichment of crosslinked peptides.

The choice of crosslinker depends on the purpose and scale of the experiment. For example, SDA as a crosslinker resulting in a high density of crosslinks was shown to be beneficial in structural studies (Lee et al., 2020), but results in a highly complex set of peptide pairs that makes identification in large-scale samples challenging. Its fast reaction time makes it especially suitable for cleanly capturing conformational changes (Belsom & Rappsilber, 2021). For large-scale studies, cleavable crosslinkers like DSSO have been widely used, as it has been thought that cleavage in the mass spectrometers aids identification by leaving characteristic features in the spectrum. There remained questions on the essentiality of cleavable crosslinkers and their implementation, which we addressed with the work in this thesis.

Crosslink enrichment. After digestion of the proteins to peptides, crosslinked peptides are very low abundant in comparison to linear peptides. Since crosslinked peptides are larger and higher charged than linear peptides, these properties can be used to separate them both in the mass spectrometer, and by chromatographic methods prior to MS acquisition. Chromatographic fractionation is used to enrich crosslinks by separating the sample based on

size and or charge. Methods previously used are for example size exclusion chromatography, strong cation exchange or hydrophilic strong anion exchange (Z. A. Chen et al., 2010; Fritzsche et al., 2012; Leitner et al., 2012; O'Reilly et al., 2020). Enrichment becomes crucial for large-scale samples such as proteome-wide studies, as their complexity is much higher than that of simpler complexes. In these cases, multiple fractionations performed subsequently can be performed to improve the results (Lenz et al., 2021; O'Reilly et al., 2020).

Data Acquisition and MS data. Fractionated peptides are further separated by liquid-chromatography, typically reversed-phase, and injected in the mass spectrometer via electrospray ionisation. This derives m/z peak positions and their intensities and the charge of the peptides coming off the column in an MS1 scan. In addition to this, metadata for each scan is collected, for example the retention time, measurement settings or fill times. During acquisition, MS1 scans are continuously recorded. They provide information about the unfragmented peptides, which, for example, can be used for their quantification (Cox & Mann, 2008). During the MS1 scan, specific precursors are selected for further fragmentation. A subsequent MS2 scan is the representation of the resulting fragments and is used for peptide identification. In the most common fragmentation methods in proteomics, collision-induced dissociation (CID) and higher energy collision dissociation (HCD), b- and y-ions are the most commonly seen fragments.

Typically in crosslinking MS workflows, the precursor selection is dependent on the charge state to reduce sampling of lower charged linear peptides. The fragmentation methods have been optimised for crosslinked peptides (Kolbowski et al., 2017; Liu et al., 2017), and most commonly use HCD or CID fragmentation. In some crosslinking workflows performed with cleavable crosslinkers, MS3 fragmentation of the single peptides is done in addition to the MS2 (Liu et al., 2017).

Because crosslinked peptides consist of two covalently bound peptides instead of a single continuous peptide, their MS2 spectra are slightly different. For one, they present more complex, chimeric spectra with fragment peaks coming from both peptides. In addition, crosslinked spectra contain not only linear fragments (continuous chain of amino acids from a single peptide), but also fragments that include the crosslinker and second peptide (**Figure 1**). These crosslinking site containing fragments are of special interest because they give extra confidence of the peptides being crosslinked and the position of the link site.

These steps are followed by the computational parts of the workflow, i.e. database search and error estimation. Besides the data analysis of experimental data, these two steps are the focus of this thesis, therefore I will describe them in detail below.

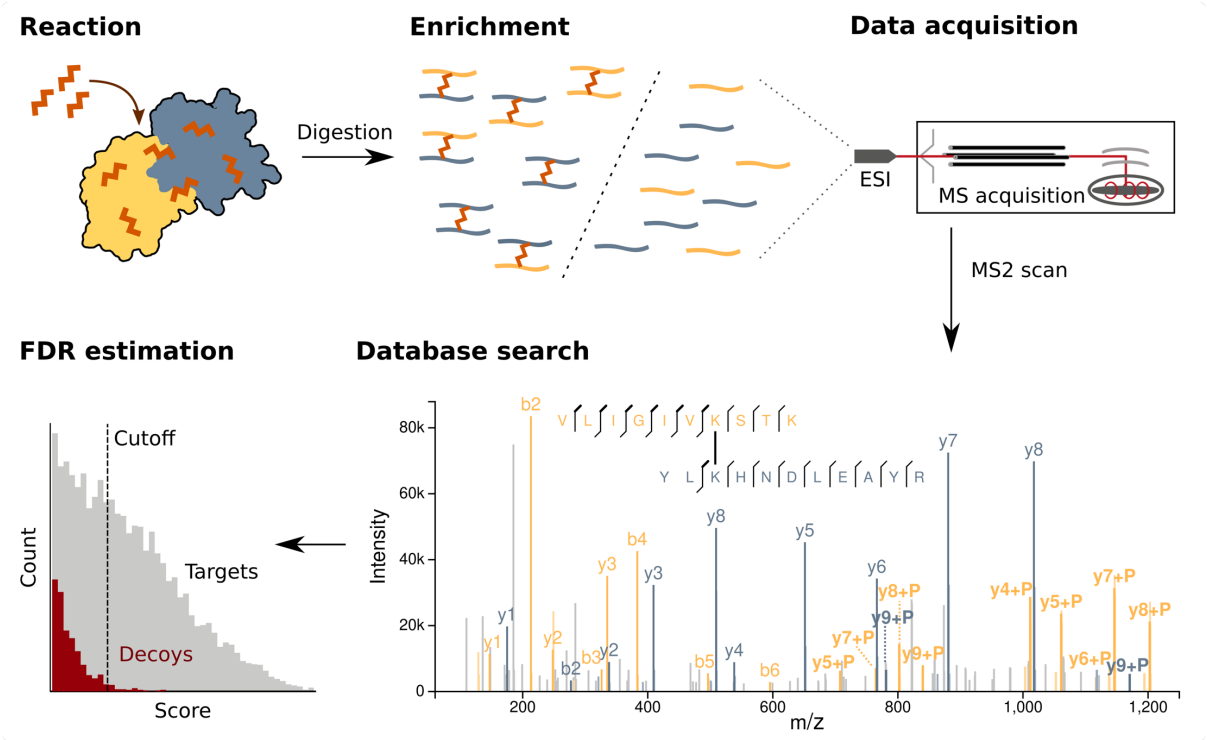


Figure 1. Crosslinking MS workflow. Crosslinker reagent is added to the sample (e.g. protein complexes or cells), proteins are digested and the sample is enriched for crosslinked peptides. After MS acquisition, MS2 scans are searched against a sequence database and the error is estimated and a cutoff applied. Adapted from (O'Reilly & Rappsilber, 2018). The spectrum is a representative spectrum of a BS3-crosslinked peptide with fragments containing the second peptide marked in bold.

Crosslinked peptide identification

MS acquisitions generate a large amount of raw data, from which the peptides need to be identified. In this section I describe how the spectra are searched during database search, with a focus on crosslinking MS.

Database search concepts. To identify the peptides from the MS2 spectra, database search algorithms can be employed. A user defined sequence database is *in silico* digested to create peptides which are likely present in the sample. During search, the algorithms are matching generated theoretical spectra of candidate peptides against the experimentally observed spectra. The candidates are then scored to represent the quality of the match. The scoring procedure and algorithm depend on the software used, but is typically a measure of the agreement between theoretical and experimental spectra in terms of matching of predicted and observed fragments and their intensities (Cox et al., 2011; Kong et al., 2017).

Fragment spectra of crosslinked peptides tend to be more complex than the spectra of linear peptides, as they are chimeric spectra of two peptides. This makes it more difficult to successfully identify the correct match. In addition, combinations of two peptides have to be considered during database search, which leads to a quadratic expansion of the search space, commonly referred to as the n-square problem. A larger search space tends to not only increase search time, but also adds noise, and therefore increases the probability for random matches during search. Several different search approaches exist specifically designed for crosslink identification (Z.-L. Chen et al., 2019; Götze et al., 2019; Hoopmann et al., 2015; Liu et al., 2017; Mendes et al., 2019). Typically, one of the peptides fragments less well and is harder to identify (Trnka et al., 2014). The better fragmenting peptide is utilised by some approaches, which simplify the search by first identifying one of the peptides in an open-modification-like search (Z.-L. Chen et al., 2019; Hoopmann et al., 2015).

One workflow commonly used in large-scale studies is based on MS-cleavable crosslinkers. Upon cleavage in the MS, they can produce signature peptide doublet peaks. These can be used to calculate the masses of the individual peptides, which in turn can be used to simplify the search (Götze et al., 2019). In addition, MS3 fragmentation can be triggered on the peptide doublets, which results in a linear spectrum of a crosslinked peptide. This simplifies the crosslinking search to that of two linear peptides (Liu et al., 2017). In my thesis work, I performed the first systematic evaluation of these workflows, providing conclusive evidence that the gains in identification do not stem from utilising the peptide masses, but instead from a better fragment coverage of the peptides.

xiSEARCH algorithm. Here, I will focus on describing xiSEARCH, the search software developed by the Rappsilber group that has been featured in a number of publications (Mendes et al., 2019; O'Reilly et al., 2020; Ryl et al., 2020) and used throughout this thesis. To circumvent the n-square search space problem, xiSEARCH first matches one of the crosslinked peptides as a linear peptide with an unknown modification. First, the MS2 spectrum is de-charged (all peaks are shifted to an m/z value as if they were singly charged) and linearized (crosslinked fragments get shifted to the m/z value of linear complement fragment) (Giese et al., 2016). Then, the most abundant peaks of the spectrum are used to identify candidate peptides (alpha candidates). The theoretical mass of the second peptide is calculated by subtracting the crosslinker and the alpha candidate mass from the experimental precursor

mass. This theoretical mass of the peptide has to match the experimental mass in a user defined mass tolerance, which is dependent on the resolution of the MS1 scan. Suitable beta candidates of a matching mass are derived for a defined number of the alpha candidates.

Both peptides are then scored together against the spectrum. xiSEARCH calculates multiple subscores which are combined to a final match score. These subscores represent different properties influencing the confidence of matches. For one, the number of fragments matched to one or both of the peptides is an important factor, as is also the coverage of the matched peptide sequence. In addition there are scores for how many and with which intensity the peaks in the spectrum are explained by the matched peptide. MS1 and MS2 errors of the match also influence the score. Other scores are based on the ranks during candidate selection.

Considerations for large-scale data. As the field is moving towards studies on proteome-wide scales, searches need to be adjusted to account for the increased complexity of the sample and the size of the database. In contrast to linear samples or even simple crosslinked samples, the chance of random matching is highly increased. For one, this is due to the quadratic increase of the search space itself, which makes it more likely to match another peptide by chance. In addition, the complexity of the sample decreases the spectral quality due to an increase of interfering species (Houel et al., 2010) and lower abundance of crosslinked peptides.

Search parameters have to be set carefully as the gain in true matches can be outweighed by their contribution to random matches. For example, the increase of the search space by an excess of modifications or proteins is usually tolerated in linear or simple crosslinked searches, but can reduce identifications in large-scale searches. An approach used in this thesis work is to only search the most abundant proteins (Lenz et al., 2021; Linden et al., 2020; Ryl et al., 2020) and to solely search common modifications. In addition, the tightening of error tolerances during search can be beneficial. Simple crosslinking data have been acquired with a high MS resolution already, and during the work of this thesis, we have increased the resolution further, which allowed to lower the error tolerance even more.

Recently developed computational approaches are being adapted to crosslinking search algorithms and await evaluation of their benefit in real world conditions. Retention time prediction has been previously established for linear proteomics (Klammer et al., 2007) and recently for crosslinking MS (Giese et al., 2021). In addition, rescoring the search results of linear searches via machine learning approaches has been established (The et al., 2016). For crosslinking, the search engine pLink2 has implemented a support vector machine to improve their results after search (Z.-L. Chen et al., 2019). However, machine learning approaches have to be performed with care to avoid overtraining on the decoy matches, which would lead to their removal, but not that of the false positive matches. This is especially the case in datasets with smaller numbers of identifications, which is the typical case in crosslinking MS.

Error estimation in crosslinking MS

Matches coming from the search engine will also contain random and therefore wrong matches, which need to be filtered out. Since scores can be dependent on the search algorithm, search database and search settings, a score cutoff generally can not be transferred between experiments. Instead, the error inside the data needs to be estimated reliably. This allows the user to apply a cutoff at the desired error, usually between 1 and 5%. The method for error estimation most common in proteomics is a false discovery rate (FDR) estimation using a target-decoy approach (Elias & Gygi, 2007). Here, random matches are modelled based on known wrong (decoy) sequences which are added to the database before search. This approach was first established in linear proteomics and has been transferred to crosslinking MS successfully (Fischer & Rappsilber, 2017; Walzthoeni et al., 2012).

Decoy generation and random space. The target-decoy approach for FDR calculation is based on the assumption that a random match will be matched to targets with the same likelihood as to decoys, i.e. that matches to the decoy database model the noise in matches to the target database. The proportion of decoys in the targets is then used to calculate the error rate depending on the score cutoff.

Decoy sequences are generated from the target database defined by the user. Multiple approaches have been proposed to create the decoys (Elias & Gygi, 2007; Wang et al., 2009), and typically the protein sequences are reversed or shuffled. The number of targets and decoys in the database should be equal and decoys should be similar to the targets in size and amino acid composition, however not contain the same peptides as in the targets. Target and decoy peptides need to be treated the same and be matched to spectra under the same scoring function, while competing for an identification.

FDR estimation in crosslinking MS is slightly different to linear approaches due to the combination of two peptides. Here, not only the simple case of correct or random needs to be considered, but also a match consisting of one correct and one random peptide. Therefore, not only target and decoy matches exist, but also their combinations, i.e. target-target, target-decoy (which is the same as decoy-target for a non-directional crosslinker) and decoy-decoy. For the typical case of a non-directional crosslinker, the random space for target-decoy matches is approximately double the size of target-target and decoy-decoy space (i.e. 1:2:1). A formula taking this difference in random spaces into account was previously derived (Fischer & Rappsilber, 2017). Apart from the FDR formula, two other considerations need to be taken into account for crosslinking MS, about which I will go into detail below.

Self and heteromeric matches. The two peptides of the crosslink can either stem from the same (self or intra crosslink) or from two different proteins (heteromeric or inter crosslink). While the first case also includes homodimers, the latter is the focus of most interaction studies. The chance of random matching is inherently different for both types of matches: Assuming protein A and protein B in the search, the crosslink A1-A2 is the same as A2-A1. In contrast, A1-B2 is different from B1-A2. This makes the number of possible random combinations different for self and heteromeric matches.

The estimated FDR is only valid on the total set of CSMs that was used for calculation. If the data are further subsetted after FDR estimation to heteromeric matches only (e.g. for reporting PPIs) this subset will be enriched in false positive identifications. In these

heteromeric matches, the error will then be higher than what was initially estimated. Therefore, FDR needs to be calculated separately for self and heteromeric matches. This splitting of the two data types has been discussed early on (Fischer & Rappsilber, 2017; Walzthoeni et al., 2012), but is not generally implemented in the field. In the work of this thesis, we demonstrate the relevance of the FDR separation for the derivation of crosslinked PPIs.

A common misconception about the splitting of the FDR estimation is that this is done to increase the number of identifications. While self matches indeed tend to increase in number as random heteromeric matches are removed, typically the number of heteromeric matches decreases as the error is estimated correctly. A separate FDR calculation is therefore not a way of increasing numbers of heteromeric matches, but ensures a reliable error estimation for them.

Result levels. Similar to linear proteomics, crosslinking data can be evaluated on different result levels (Fischer & Rappsilber, 2017) (**Figure 2A**). The initial match of peptide to spectrum is referred to as peptide spectrum match, or in the case of crosslinking MS, crosslinked spectrum match (CSM). These can exist in multiple copies if the same precursor was triggered for fragmentation multiple times, so often only the highest scoring unique CSM is considered, so as to not bias the FDR estimation. CSMs of different charge states are aggregated into peptide pairs. Multiple different peptide sequences can represent the same residues being crosslinked, for example due to missed cleavages or modifications. Therefore peptide pairs are further aggregated into residue pairs, only including the position of the crosslink on the protein(s). Finally, the residue pairs can merge further into PPIs, which are the main focus of proteome-wide interaction studies. Merging to higher levels is not commonly done, and so comparisons on the approaches of merging results to higher levels are rare and only existing for linear proteomics (Audain et al., 2017). xiFDR calculates the new score as $Score_{higher} = \sqrt{\sum Score_{lower}{}^{2}}$ and estimates the FDR based on that score. This has the implicit assumption of modelling error propagation from multiple matches as independent events.

For increasing result levels, true matches will corroborate and aggregate with each other. Random matches however will do so less often. Therefore, the error grows with increasing result levels, making it crucial that the FDR is calculated and the cutoff applied on the level of interest. The result level of choice depends on the purpose of the study. For analysis of search parameters or performance, it is usually sufficient and recommended to analyse (unique) CSMs, as these are the results directly stemming from the search. Structural studies of single proteins or protein complexes are usually interested in residue pairs as distance restraints, therefore the FDR should be estimated on residue pair level. For large-scale interaction studies the FDR needs to be estimated on PPI-level, as shown by work included in this thesis.

Pre-filtering of data. A common practice of increasing the confidence of results is the application of (pre-)filters to the data. With increasing confidence, the number of matches passing the FDR cutoff will also grow. Pre-filters are applied to target matches as well as decoy matches. A common filtering parameter is the delta score, as this describes the score difference to the next match and can therefore be related to the confidence of the match. To increase numbers of matches on higher result levels, it has been shown that a prefilter on the lower level results can be beneficial (Fischer & Rappsilber, 2017). Our in-house software xiFDR

implements a multistep grid search ("boosting") that is testing different cutoffs numerically to find the best pre-filter settings.

Of note, some studies employ a filter on the data after FDR estimation (Bartolec et al., 2020; Yugandhar, Wang, Leung, et al., 2020), since they notice an unexpected amount of noise in their results. A properly working FDR estimation should filter out noise in the data and result only in a reduction in number of matches. Instead of a sign for noisy data, the observations in these studies are more likely to stem from an unreliable FDR estimation. For a sensible statement about the error in the data, any pre-filters should be applied first with FDR estimated after that.

Validation. In the beginning, crosslinking results were commonly validated on structures. Based on the known distance the crosslinker can cover, a maximum plausible distance is calculated. The corresponding distances of the resulting residue pairs are calculated on solved structures and the percentage of violation compared to the maximum crosslinker distance is used as an estimate for false positive results. A downside of this approach is that flexibility of the proteins and conformational changes are ignored.

In addition, while this might have been a sensible approach for single proteins or complexes, on a larger scale this leads to an underestimation of the error (Yugandhar, Wang, Wierbowski, et al., 2020). The selection of certain complexes only allows the validation of a subset of the PPI network (**Figure 2B**) that in addition has already been validated by other methods (i.e. solving the structure experimentally). Therefore there is a bias in pre-selection for true matches. In addition, this can lead to a selection of complexes that are also highly abundant in the cell, leading to high abundant and therefore high quality links.

Another validation approach targeted at search softwares was proposed based on a synthetic library of crosslinked peptides (Beveridge et al., 2020). Based on different poolings of the synthetic peptides, certain peptide combinations were known to be wrong and could be used to validate the FDR. However, the sample size was relatively small to simulate proteome-wide studies and fails to simulate the complexity and noise of large-scale spectra. In the work of this thesis we have established a similar control based on experimentally impossible crosslinks, however on a proteome-wide scale.

A more suitable and universal control of search softwares and their FDR is the employment of an entrapment database, as demonstrated by work in this thesis. Here, in addition to the sequences expected in the sample, sequences known to be wrong are added to the database. It is important that the entrapment space is large enough to be matched at random sufficiently. The principle is highly similar to the target-decoy approach, however here the identification process is truly blind to the entrapment proteins. These matches can be used as a decoy independent control of the search results and FDR, and is particularly useful to check for wrong decoy generation or overtraining during machine-learning.

The work of this thesis includes improvements on the identification of crosslinks and demonstrates procedures to reliably estimate the error on large-scale data. Together, these advances have allowed for crosslinking MS to become a method suitable for proteome-wide analyses.

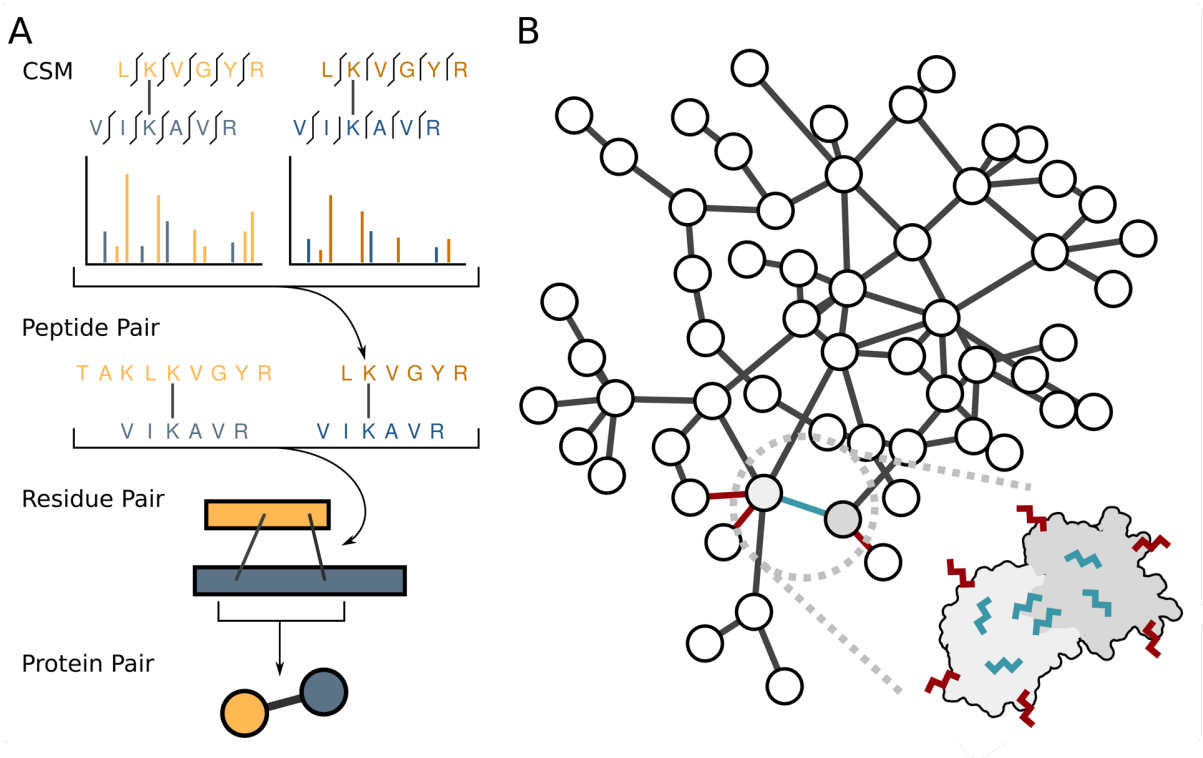


Figure 2. Error estimation and validation for crosslinked PPIs. A) Aggregation of matches for different result levels. B) Validation of PPI networks by mapping crosslinks to structures is unsuitable. It only considers a small subset of the network, while there is no information on the crosslinks not contained in the protein structure. Protein complex image adapted from (Yugandhar, Wang, Wierbowski, et al., 2020).

Contributions and Main Findings

The main focus of my work was data analysis of crosslink identifications at different points in the workflow. The overall goal was to increase the depth of identifications and their reliability to bring crosslinking MS from analysing complexes to proteome-wide analyses. In this cumulative thesis, I combine the manuscripts that are the result of this work. Manuscript 1 and 3 are accepted in a peer-reviewed journal. Manuscript 2 is included here as the preprint which was submitted to *Analytical Chemistry*, but is now accepted.

Manuscript 1 describes a problem arising specifically for crosslinked peptides during data acquisition, for which we implemented a solution into the database search (Lenz et al., 2018). The second manuscript is an evaluation of the cleavable crosslinker DSSO in its fragmentation behaviour as well as its effect on the database search. As cleavable crosslinkers are largely used and proposed to be beneficial for large-scale analyses, a better understanding is necessary to drive further optimisation. As more proteome-wide studies with a focus on PPIs are being performed, the error of this data needs to be evaluated. In manuscript 3, we therefore demonstrate a reliable procedure to estimate the error for crosslinked PPIs (Lenz et al., 2021).

The first manuscript **"In-Search Assignment of Monoisotopic Peaks Improves the Identification of Cross-Linked Peptides."** is a study into the monoisotopic peak detection and correction for crosslinked peptides. Due to the low abundance and large size of crosslinked peptides, the monoisotopic peak of the precursor isotope cluster is often small and therefore wrongly assigned by the vendor software. Other correction software, which detects precursor features on MS1 level, can correct the precursor mass only partially. We implement a solution into xiSEARCH by enabling it to search multiple monoisotopic masses. This led to an increase in the number of crosslinked identifications at constant FDR.

For this project, I performed the database searches, analysed the data and created the figures. Lutz Fischer implemented the algorithm into xiSEARCH. The manuscript was written by me and my co-authors.

The second manuscript **"Improved peptide backbone fragmentation is the primary advantage of MS-cleavable crosslinkers."** is a study into the fragmentation behaviour of the MS-cleavable crosslinker DSSO. Many of the proteome-wide crosslinking studies have used MS-cleavable crosslinkers and it was suggested that successful crosslinking of complex samples depends on them. We therefore used previously published DSSO datasets to derive statistical evidence of the suggested advantages. We found that almost all CSMs had the peptide doublet signature peaks, from which peptide masses could be calculated. However, knowing the peptide masses turned out not to be the main advantage, but instead an improved backbone fragmentation through cleavability of the crosslinker. In addition, we found that the commonly used MS3 approaches lack sensitivity and specificity and are therefore surpassed by steppedHCD methods. Our analysis of current methods demonstrates the importance of thorough data analysis in experimental optimisation and will allow for further improvements in crosslinking reagent development.

I performed the re-searches of the datasets and the analysis of their results, as well as the analysis of peak ranks, sequence coverages and the BS3-DSSO comparison. The

manuscript was written and the figures created in close collaboration with the second co-author Lars Kolbowski.

In the third manuscript **"Reliable identification of protein-protein interactions by crosslinking mass spectrometry."**, we use a controlled large-scale sample of *E. coli* lysate to demonstrate the error estimation for crosslinking MS derived PPIs. While FDR for crosslinking MS has been discussed, no consensus has emerged, especially for PPIs. Using proteome fractionation and crosslinking in separate fractions, we are able to define non-crosslinkable PPIs, and therefore pinpoint those that were falsely identified. We establish this as an experimental control to validate the PPI error in our data. With this and three additional computational controls, we demonstrate that other, previously used, FDR approaches drastically underestimate the error on PPIs. Only FDR estimation performed separately on self and heteromeric crosslinks, and estimated on the PPI level, can correctly estimate the error for PPIs. Using this heteromeric PPI-FDR, we derive an *E. coli* lysate interaction network of 590 PPIs at 1% FDR. In this network, we found a previously unknown interactor of RNA polymerase. With the structural information contained in the crosslinks, we map its binding site to the DNA exit channel of RNA polymerase.

I performed database searches and FDR estimations, established and implemented the controls of the target-decoy FDR and performed the binding site analysis of YacL. The manuscript was written with the two other co-authors, Francis O'Reilly and Ludwig Sinn. This work is also included in the thesis of Ludwig Sinn.

In addition I co-authored the following papers:

O'Reilly, F. J., Xue, L., Graziadei, A., Sinn, L., Lenz, S., Tegunov, D., Blötz, C., Singh, N., Hagen, W. J. H., Cramer, P., Stülke, J., Mahamid, J., & Rappsilber, J. (2020). In-cell architecture of an actively transcribing-translating expressome. *Science*, *369*(6503), 554–557. https://doi.org/10.1126/science.abb3758

Schäpe, P., Kwon, M. J., Baumann, B., Gutschmann, B., Jung, S., Lenz, S., Nitsche, B., Paege, N., Schütze, T., Cairns, T. C., & Meyer, V. (2019). Updating genome annotation for the microbial cell factory Aspergillus niger using gene co-expression networks. *Nucleic Acids Research*, *47*(2), 559–569. https://doi.org/10.1093/nar/gky1183

Ryl, P. S. J., Bohlke-Schneider, M., Lenz, S., Fischer, L., Budzinski, L., Stuiver, M., Mendes, M. M. L., Sinn, L., O'Reilly, F. J., & Rappsilber, J. (2020). In Situ Structural Restraints from Cross-Linking Mass Spectrometry in Human Mitochondria. *Journal of Proteome Research*, *19*(1), 327–336. https://doi.org/10.1021/acs.jproteome.9b00541

Manuscript 1:

In-Search Assignment of Monoisotopic Peaks Improves the Identification of Cross-Linked Peptides.

Manuscript available online
DOI: 10.1021/acs.jproteome.8b00600

Journal of proteome research

Article

Cite This: J. Proteome Res. 2018, 17, 3923−3931

pubs.acs.org/jpr

In-Search Assignment of Monoisotopic Peaks Improves the Identification of Cross-Linked Peptides

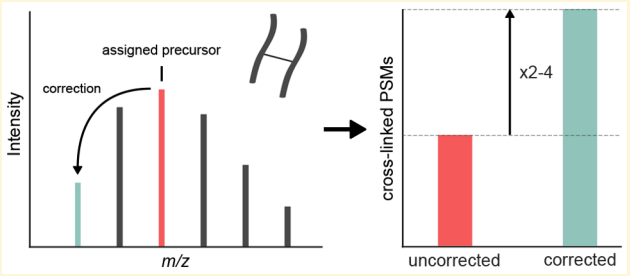
Swantje Lenz,[†] Sven H. Giese,[†] Lutz Fischer,[‡] and Juri Rappsilber*[,†,‡]

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

[‡]Wellcome Centre for Cell Biology, University of Edinburgh, Edinburgh EH9 3BF, United Kingdom

Supporting Information

ABSTRACT: Cross-linking/mass spectrometry has undergone a maturation process akin to standard proteomics by adapting key methods such as false discovery rate control and quantification. A poorly evaluated search setting in proteomics is the consideration of multiple (lighter) alternative values for the monoisotopic precursor mass to compensate for possible misassignments of the monoisotopic peak. Here, we show that monoisotopic peak assignment is a major weakness of current data handling approaches in cross-linking. Cross-linked peptides often have high precursor masses, which reduces the presence of the monoisotopic peak in the isotope envelope. Paired with generally low peak intensity, this generates a challenge that may not be completely solvable by precursor mass assignment routines. We therefore took an alternative route by '"in-search assignment of the monoisotopic peak" in the cross-link database search tool Xi (Xi-MPA), which considers multiple precursor masses during database search. We compare and evaluate the performance of established preprocessing workflows that partly correct the monoisotopic peak and Xi-MPA on three publicly available data sets. Xi-MPA always delivered the highest number of identifications with ∼2 to 4-fold increase of PSMs without compromising identification accuracy as determined by FDR estimation and comparison to crystallographic models.



KEYWORDS: *cross-linking, mass spectrometry, data processing, proteomics, software, structural proteomics, BS3, SDA, peptides, monoisotopic mass*

INTRODUCTION

Several approaches have been utilized to increase the numbers of identified cross-links, for example enriching for cross-linked peptides,[1−4] using different proteases[1,5,6] or optimizing fragmentation methods.[7,8] In parallel with experimental developments, data analysis has also progressed to extract even more cross-links to be used as distance restraints for modeling of proteins and their complexes.[9,10] Search software has been designed for the identification of cross-linked peptides, for example Kojak,[11] xQuest,[12] pLink,[13] XlinkX,[14] or Xi.[5] In addition, cross-linking workflows can make use of preprocessing methods to improve data quality and reduce file sizes,[15] as well as postprocessing methods to filter out false identifications[11,16] and custom-tailored false discovery rate (FDR) estimation.[17−19] Preprocessing can improve peptide identification by correcting the MS1 precursor ion m/z and simplifying MS2 fragment spectra. Established proteomics software perform such preprocessing, including MaxQuant[20,21] and OpenMS.[22,23] For example, MaxQuant performs a variety of preprocessing steps: it corrects the precursor m/z by an intensity-weighted average if a suitable peptide feature is found, reassigns the monoisotopic peak and contains options for intensity filtering of MS2 peaks. Despite such correction of the precursor mass, many linear search engines have integrated the possibility of considering multiple monoisotopic peaks during search.[24−26] However, the benefits of this search feature are currently unclear. It seems that the assignment of monoisotopic mass for tryptic peptides is already achieved adequately either during acquisition or as part of preprocessing.

Cross-linked peptides have characteristics that may render MS1 monoisotopic precursor mass assignment as used for linear peptides nonoptimal: high-charge states, large masses, and low abundances. Several cross-link search engines include MS1 correction in their pipeline: pLink corrects monoisotopic peaks based on previous work with linear peptides,[27] however does not include a parameter for searching multiple precursor masses. Kojak averages precursor ion signals of neighboring scans to create a composite spectrum and infer the true monoisotopic mass of the precursor. If this step fails, precursor masses up to −2 Da lighter are searched.[11] For previous searches in Xi, MaxQuant was used to perform preprocessing. Neither xQuest nor XLinkX describe precursor correction in their workflow documentation and there is no option for additionally searched masses available in the respective search

Received: August 4, 2018
Published: October 8, 2018

DOI: 10.1021/acs.jproteome.8b00600
J. Proteome Res. 2018, 17, 3923−3931

parameters. We are not aware of a detailed evaluation of the impact of different preprocessing techniques for cross-link identification, independent of the search software. Correcting the monoisotopic mass of precursors, although acknowledged as an issue,[11,28] awaits systematic evaluation.

In this study, we show that errors in assigning monoisotopic peaks during data acquisition are frequent for cross-linked peptides because of their size and generally low abundance. This adversely affects their identification. We show that standard software suites, MaxQuant and OpenMS correct monoisotopic precursor masses of cross-linked peptides with variable success. We then implement an option in Xi to consider multiple precursor masses during search, to minimize the impact of false monoisotopic precursor mass assignment on the identification of cross-links.

METHODS

Data Sets

In this study, we used three publicly available data sets (Table 1). The three data sets were chosen to reflect a range of applications of cross-linking mass spectrometry as well as a range of data complexity: the first data set is Human Serum Albumin (HSA) cross-linked with succinimidyl 4,4-azipentanoate (SDA) and fragmented using five different methods (PXD003737).[29] The second data set is a pooled pseudocomplex sample with seven separately cross-linked proteins with bis(sulfosuccinimidyl) suberate (BS3) (PXD006131).[7] This data set includes data from four different fragmentation methods. The third data set is the most complex sample, composed of 15 size exclusion chromatography fractions of *Chaetomium thermophilum* lysate cross-linked with BS3 and fragmented only with HCD (PXD006626).[30] The first and last size exclusion fractions were used to optimize the search parameters for this data set. All samples were analyzed on an Orbitrap Fusion Lumos Tribrid mass spectrometer (Thermo Fisher Scientific, San Jose, CA) using Xcalibur (version 2.0 and 2.1).

Table 1. Overview of Datasets Used

data set	sample	database size[a]	reference
1	HSA	1	Giese et al. 2016
2	pseudocomplex	7	Kolbowski et al. 2017
3	*C. thermophilum*	198−400[b]	Kastritis et al. 2017

[a]Database size refers to the number of proteins in the database. [b]Multiple size exclusion chromatography fractions (n = 15).

Preprocessing

Raw files were preprocessed independently using MaxQuant (1.5.5.30), OpenMS (2.0.1) and the ProteoWizard[31] tool msconvert (3.0.9576) for comparison. Scripts automating the preprocessing, search and evaluation were written in Python (2.7).

The essential steps during the preprocessing can be divided into two parts: (1) correction of the m/z or charge of the precursor peak for MS2 spectra and (2) denoising of MS2 spectra. MaxQuant and OpenMS both try to correct the precursor information via additional feature finding steps, i.e. identifying a peptide feature from the retention time, m/z and intensity domain of the LC-MS run. Additionally, denoising of the MS2 spectra is performed by simply filtering the most intense peaks in defined m/z windows. The preprocessing is by default enabled in MaxQuant and was run using the partial processing option (steps 1−5) with default settings except for inactivated "deisotoping" and "top peaks per 100 Da", which was set to 20. The OpenMS preprocessing workflow includes centroiding, feature finding,[32] precursor correction (mass and charge) using the identified features and MS2 denoising as described above (Supporting Information (SI) Figure S1). Msconvert was used to convert the raw files to mgf files without any correction. These peak files were denoted as "uncorrected" and used as our baseline to quantify improvements in the subsequent database search. For the "in-search assignment of the monoisotopic peak" in Xi (Xi-MPA), we used msconvert to convert raw files to mgf files and included a MS2 peak filter for the 20 most intense peaks in a 100 m/z window.

Data Analysis

Peak files were searched separately in Xi (1.6.731) with the following settings: MS accuracy 3 ppm, MS/MS accuracy 10 ppm, oxidation of methionine as variable modification, tryptic digestion, two missed cleavages. For samples cross-linked with SDA, linkage sites were allowed on lysine, serine, tyrosine, threonine, and protein n-terminus on one end and all amino acids on the other end of the cross-linker. Variable modifications were monolink SDA (110.048 Da), SDA loop-links (82.0419 Da), SDA hydrolyzed (100.0524 Da), SDA oxidized (98.0368 Da)[31] as well as carbamidomethylation on cysteine. For searches with BS3, linkage sites were lysine, serine, threonine, tyrosine, and the protein n-terminus. Carbamidomethylation on cysteine was set as fixed modification. Allowed variable modifications of the cross-linker were aminated BS3 (155.0946 Da), hydrolyzed BS3 (156.0786 Da) and loop-linked BS3 (138.0681 Da). For collision-induced dissociation (CID) and beam-type CID, also referred to as higher-energy C-trap dissociation (HCD), b- and y-ions were searched for, whereas for electron transfer dissociation (ETD) c- and z-ions were allowed. For ETciD and EThcD, b-, c-, z-, and y-ions were allowed. The HSA and pseudocomplex data sets were searched against the known proteins in the sample. For each protein fraction of the *C. thermophilum* data set, the databases of the original publication were used, where a database was created for each fraction by taking the most abundant proteins (iBAQ value above 10^6). For searches employing Xi-MPA, the parameter "missing_isotope_peaks" was set to the respective mass range searched. Data sets 1 and 2 were searched with a reversed decoy database, whereas data set 3 was searched with a shuffled decoy database due to palindromic sequences. For the reversed decoy database, lysines and arginines were swapped with the preceding amino acid before peptide generation.[17,20]

For cross-linking, there are different information levels: PSMs, peptide pairs, residue pairs (links) and protein pairs. The false discovery rate (FDR) can be calculated on each one of these levels and should be reported for the level at which the information is given.[12] The FDR was calculated as described in Fischer et al.[17] using xiFDR (1.0.14.34) according to the following equation: $FDR = \frac{TD - DD}{TT}$. A 5% PSM level cutoff was imposed. The setting "uniquePSMs" was enabled and the FDR was calculated separately on self-and between links. Minimal peptide length was set to 6. In data set 2, identified cross-linked residues were mapped to the crystal structure of the respective protein and the Euclidian distance between the alpha-carbons was calculated. Structures were downloaded

DOI: 10.1021/acs.jproteome.8b00600
J. Proteome Res. 2018, 17, 3923−3931

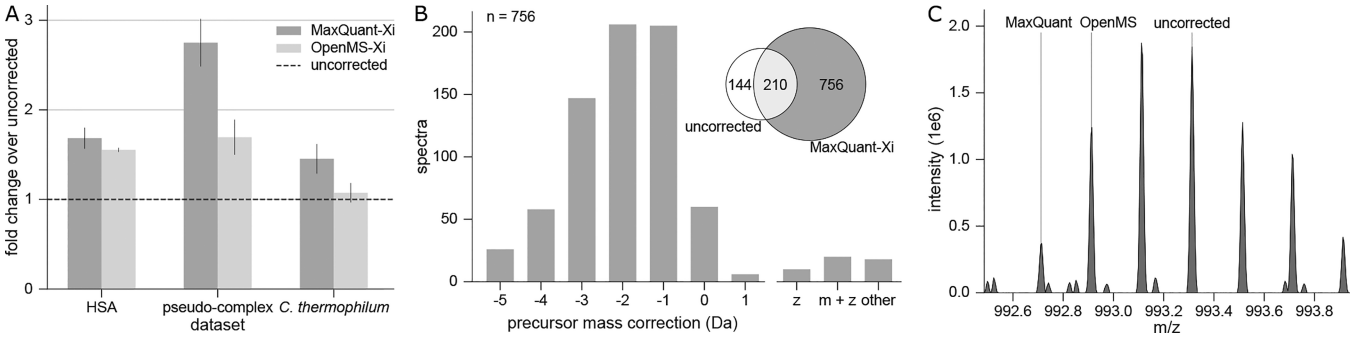


Figure 1. Correction of the monoisotopic peak is crucial in cross-link identification. (A) The data sets were preprocessed using MaxQuant and OpenMS, leading to more identified PSMs in all cases. Fold changes from uncorrected data (msconvert conversion of Xcalibur data) were calculated for each file separately and the mean plotted. Error bars represent the standard error of the mean between different acquisitions (HSA: n = 3, pseudocomplex: n = 3, *C. thermophilum*: n = 8). (B) The majority of additional identifications after preprocessing are due to correction of the precursor mass to lighter monoisotopic masses. Spectra that are unique to MaxQuant preprocessed searches of HCD acquisitions from data set 2 were evaluated in terms of precursor correction. The main proportion of the gain was corrected to lighter masses of up to −3 Da, while charge state correction or correction to heavier masses rarely occurred. (C) Isotope cluster of a corrected precursor of m/z 992.71 (z = 5, m = 4958.6 Da) was solely identified in MaxQuant preprocessed results. In OpenMS preprocessed and uncorrected data, the wrong monoisotopic mass was selected for unknown reasons.

from the PDB (IDs: 1AO6, 5GKN, 2CRK, 3NBS, 1OVT, 2FRJ). Kojak (1.5.5) was run via the Trans-Proteomic Pipeline (5.1.0)[33] using default settings except: MS1 resolution 120 000, BS3 allowed on lysine, serine, threonine, tyrosine, and the protein n-terminus, aminated BS3 (155.0946 Da) as variable modification of the cross-linker, 3 ppm mass tolerance on MS1 level. For the uncorrected search, the isotope error was set to 0 and precursor refinement was disabled. PSMs were validated using PeptideProphet[34] and FDR calculated as described above on the resulting probability.

The mass spectrometry data have been deposited to the ProteomeXchange Consortium via the PRIDE[35] partner repository with the data set identifier PXD011121. For transparency, python scripts are available on GitHub under: https://github.com/Rappsilber-Laboratory/Xi-MPA_scripts.

RESULTS AND DISCUSSION

We evaluated the impact on cross-link identification in Xi of changing the precursor monoisotopic mass that was initially assigned during data acquisition ("uncorrected"). In this analysis, MaxQuant and OpenMS were used as preprocessing tools. We used three different data sets that differ in complexity and fragmentation regimes. To measure the improvements from using the preprocessing tools, a simple conversion from raw files to mgf format was done with msconvert and used as a baseline. Note that in the spectrum header, there are two m/z values: the trigger mass of the MS2 and the assigned monoisotopic peak of the isotope cluster. Msconvert extracts the assigned monoisotopic mass. Processed data were searched separately in Xi and evaluated on PSM level or link (= unique residue pair) level, with a 5% FDR. Finally, the newly implemented in-search assignment of monoisotopic peaks in Xi was compared to the elaborate preprocessing pipelines in OpenMS and MaxQuant.

Preprocessing Increases the Number of Cross-Link PSMs by Finding the Correct Monoisotopic Peak

The data sets were preprocessed in MaxQuant and OpenMS and numbers of identified PSMs were compared to those obtained using uncorrected data. Data sets 1 (HSA) and 2 (pseudocomplex) were acquired with different acquisition methods. For comparability to data set 3 (complex mixture), we focused on the HCD acquired data. Cross-links between proteins were excluded, either because they were experimentally not possible (data set 2) or observed in too low numbers for reliable FDR calculation (data set 3).

For uncorrected data, 672, 354, and 2157 cross-link PSMs resulted for the HSA data set (data set 1), pseudocomplex (data set 2), and first and last fractions of *C. thermophilum* respectively (data set 3). Both preprocessing approaches improved numbers of identified PSMs for all data sets: Preprocessing in MaxQuant led to 1127 (68% increase), 966 (173% increase), and 2966 (38% increase), while for OpenMS, 1044 (55% increase), 598 (69% increase) and 2394 (11% increase) PSMs were identified (Figure 1A).

We assessed the gains in identified PSMs of preprocessed data compared to uncorrected data (focusing on data set 2) regarding three forms of precursor correction: (1) correction of the monoisotopic mass, (2) charge state correction, and (3) small corrections of the m/z value based on averaging the m/z values across the peptide feature (Figure 1B). Precursor mass and charge state of spectra identified solely in MaxQuant-Xi were compared to their counterparts when searching uncorrected data in Xi. Of the 756 newly identified spectra, 686 (91%) had a different monoisotopic precursor mass. Precursors were primarily corrected to lighter masses by MaxQuant, that is, the monoisotopic peak correction by −1 (208 spectra), −2 (215 spectra), −3 (149 spectra), and −4 Da (62 spectra). Greater shifts (−5 to −7 Da) only occurred 30 times, and corrections to heavier masses were observed 22 times. Only 30 spectra (4%) were corrected in their charge state. For the 60 spectra (8%) without correction in charge state or monoisotopic peak, we only identified nine spectra that had a higher error than 3 ppm before preprocessing, indicating a small correction of the initial precursor m/z (by averaging of peptide feature peaks). The main proportion of these identifications is likely a product of noise removal in MS2 spectra or small changes in the score distribution. Similarly, for OpenMS-Xi, the monoisotopic peak correction had the greatest impact: Of the 314 spectra that OpenMS added over uncorrected data, 139 were precursor corrected by −1 Da and 108 to −2 Da. In contrast to MaxQuant, corrections to −3 or lighter were not observed, which might explain the higher number of identifications obtained with MaxQuant-Xi.

DOI: 10.1021/acs.jproteome.8b00600
J. Proteome Res. 2018, 17, 3923−3931

For data set 1 and 3, the gains of preprocessing are smaller than for data set 2. The median peptide mass of data set 1 (3368 Da) is smaller than the median mass of data set 2 (3946 Da) and we later show that this is a major factor in precursor mass assignment. This reflects in the distribution of lighter masses assigned: 46%, 33%, and 21% were corrected by −1 Da, −2 Da, and to even lighter peaks, respectively. Data set 3 was acquired with a different version of Xcalibur, for which we saw better mass assignment than for earlier versions (data not shown). As an implication of this, already 67% of the lighter corrected masses are shifted by −1 Da, 21% by −2 Da, 12% to even lighter masses. However, we were not able to follow up on this to our content, since the source code of the vendor software is not available.

In summary, preprocessing, especially monoisotopic peak correction, leads to a notable increase in identifications. Using the 3-dimensional peptide feature is advantageous compared to on-the-fly detection of the monoisotopic peak. If the preceding MS1 spectrum was acquired during the beginning (or end) of the elution profile of a peptide, the intensity will be low. Thus, the monoisotopic peak might not even be detectable at the time of fragmentation. For large (cross-linked) peptides, this effect might be exacerbated by the monoisotopic peak usually being less intense than other isotope peaks. Therefore, using the additional information from the retention time domain will be beneficial. The same feature information can also be used to determine or validate the assigned charge state of the precursor. However, the instrument software almost always assigned the same charge state as MaxQuant or OpenMS. Thus, the main advantage for identifying cross-linked peptides arises from monoisotopic peak correction.

Interestingly, OpenMS and MaxQuant did not always agree on or find the same monoisotopic peak (Figure 1C). Of the total MaxQuant-corrected spectra with a different monoisotopic mass, 81% were not corrected and 6% corrected differently with OpenMS. Vice versa, 15% of the monoisotopic peaks corrected by OpenMS were not corrected by MaxQuant and 25% were corrected differently. Both MaxQuant and OpenMS have their own implementations for precursor correction - therefore, there might be instances where MaxQuant is able to find a corresponding peptide feature where OpenMS does not and vice versa. Although OpenMS did not lead to the same improvements in the number of identifications as MaxQuant, it did correct some precursors that the latter did not. We therefore suspect that there are also precursors with a falsely assigned monoisotopic peak that were corrected with neither algorithm. Furthermore, 3-dimensional detection of peptide features is challenging for low intensity peptides. In conclusion, there likely remain falsely assigned monoisotopic peaks in the data, ultimately leading to missed or false identifications.

In-Search Monoisotopic Peak Assignment Increases the Number of Identifications

We observed multiple cases where MaxQuant and OpenMS disagreed in their monoisotopic peak choice, indicating that the problem of monoisotopic peak assignment (MPA) cannot be solved easily at MS1 level. Indeed, we found instances where the monoisotopic peak is simply not distinguishable from noise, so a feature-based correction would not be feasible. Nevertheless, the associated MS2 spectra could be matched to a cross-linked peptide when considering multiple different monoisotopic masses during search. This shows that the extra information on obtaining a peptide-spectrum match is advantageous to MPA over considering MS1 information alone. Therefore, we implemented a monoisotopic peak assignment in Xi: for each MS2 spectrum, multiple precursor masses are considered during a single search and the highest scoring peptide-pair assigns the precursor mass. Note that this is different from simply searching with a wide mass error for MS1. The mass accuracy of MS1 is minimally compromised as multiple candidates for the monoisotopic mass are taken and considered with the original mass accuracy of the measurement.

To find a good trade-off between increased search space and sensitivity, we tested different mass range settings on the data sets. For data set 2 (HCD subset), the number of PSMs increased with ranges up to −5 Da on the considered monoisotopic masses (Figure 2A). However, the increase in identifications from −4 to −5 Da was only 3% and considering the increase in search time, we continued with a maximal correction to −4 Da as the optimal setting for this data set. Xi-MPA yielded 1508 PSMs, which is a 326% increase compared to searching uncorrected data and a 56% increase compared to MaxQuant-Xi. Similar improvements are observed for the other fragmentation methods in this data set (SI Figure S2). Additionally, we corrected up to −7 Da to test if a large increase in search space increases random spectra matches as measured by the target-decoy approach. The number of identifications at 5% FDR decreased only slightly compared to −5 Da (−1%), but still led to more identifications than up to −4 Da (3%). In the HSA data set, Xi-MPA with up to −4 Da increased the number of identified PSMs by 170% compared to uncorrected data (Figure S3).

As a final evaluation of in-search monoisotopic peak assignment, we searched the complete data set of *C. thermophilum*. We used 0 to −3 Da as the range of Xi-MPA, since an initial analysis of the first and last fraction of the *C. thermophilum* data set returned a similar number of identifications when running Xi-MPA up to −4 Da or −3 Da (Figure S4). As a comparison, we took the original peak files obtained from PRIDE. The FDR was calculated separately on self-and between links, enabled boosting (automatic prefiltering on PSM and peptide pair level[17]), with a minimum of three fragments per peptide and a minimal delta score of 0.5. For the original peak files, which were preprocessed in MaxQuant, we identified 3848 PSMs, 2594 peptide pairs and 1653 cross-links, with a 5% FDR on each respective level (Figure 2B). Xi-MPA resulted in 4952 PSMs (29% increase), 3566 peptide pairs (37% increase), and 2273 cross-links (38% increase).

Next, we looked into the complementarity of search results with the different approaches, using data set 2 at 5% link-FDR. Preprocessing via MaxQuant and OpenMS led to 172 and 158 links, respectively, while Xi-MPA resulted in 243 links. While the overlap between links of OpenMS-Xi and MaxQuant-Xi is only 50%, Xi-MPA identifications cover 76% of both searches (Figure 2C). Nineteen and 23 links are uniquely found in MaxQuant and OpenMS preprocessed data, respectively. However, there are five decoy links as well in each unique set (resulting in a link-FDR of 26% and 22%). For Xi-MPA, there are 75 unique target links with 12% link-FDR.

Identification-based monoisotopic peak assignment as employed by Xi-MPA results in more identifications than the feature-based assignment algorithms of OpenMS and MaxQuant. Neither OpenMS nor MaxQuant correct all precursor

DOI: 10.1021/acs.jproteome.8b00600
J. Proteome Res. 2018, 17, 3923−3931

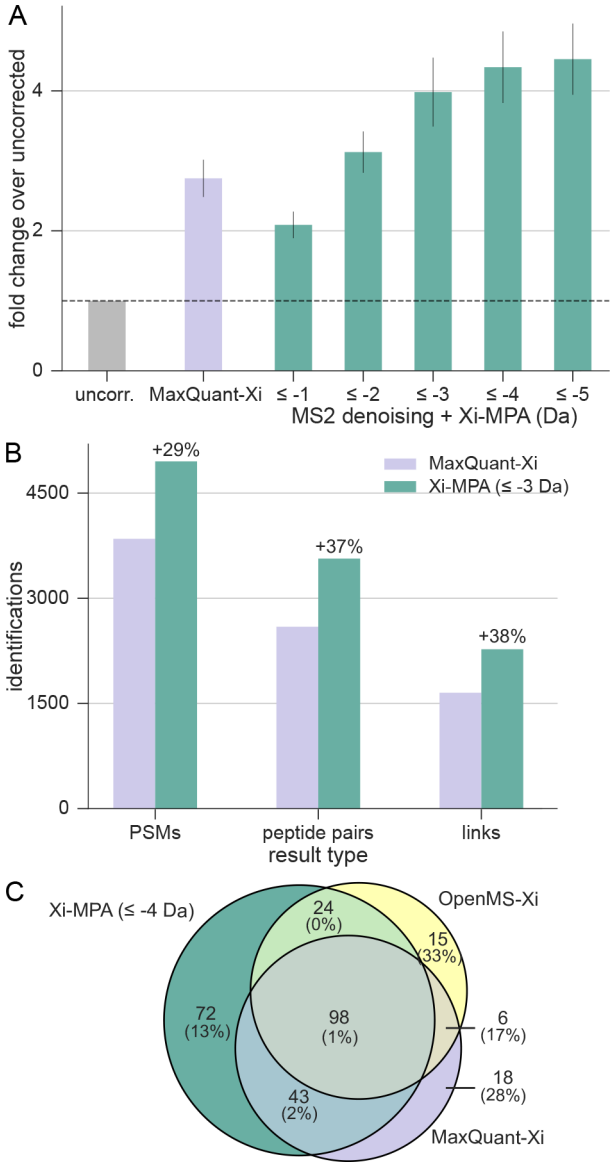


Figure 2. In-search monoisotopic peak assignment outperforms preprocessing. (A) Performance of Xi-MPA on data set 2. HCD data from the pseudocomplex data set were searched assuming different ranges of missing monoisotopic peaks. With increased ranges, the number of identified PSMs also increases. (B) Performance of Xi-MPA on the complete *C. thermophilum* data set. All 15 fractions were searched with the original preprocessed data as well as with Xi-MPA. (C) Overlap of identified residue pairs of MaxQuant-Xi and OpenMS-Xi to residue pairs gained from Xi-MPA (data set 2). Numbers in brackets are the proportion of decoys in the respective regions.

masses that are incorrectly assigned during data acquisition. In Xi-MPA, spectra are searched with multiple monoisotopic masses, thereby relying less on the MS1 information. The quality of the precursor isotope cluster does not contribute to the decision of monoisotopic mass and spectra for which correction failed will be identifiable. One could hypothesize that increasing the search space by considering multiple masses will lead to more false positives, thereby reducing the number of true identifications. This is not the case, as we match substantially more PSMs at constant FDR by considering alternative monoisotopic masses. As a second plausible caveat, this approach increases the search time. However, the use of relatively cheap computational time appears balanced by the notable increase in identified cross-links. The optimal range of additional monoisotopic peaks to search will however be dependent on complexity and quality of MS1 acquisition and the instrument software. To reduce the mass range considered in Xi-MPA, we developed a MS1 level-based approach. For each precursor, we search lighter isotope peaks in MS1 and use this to narrow the search space (explained in detail in the SI). This led to an average of 24% less values to be considered, while only reducing the number of identifications by 3%. We hope that our observation of the monoisotopic peak detection challenge in cross-linking together with our publicly available data sets will lead to further improvements in monoisotopic peak-assignment algorithms in the future, possibly tailored to cross-link data.

The cross-link search engine Kojak employs a precursor correction in its pipeline.[11] As we could not find a detailed evaluation of the impact of precursor correction in Kojak, we searched the HCD data of the pseudocomplex data set without correction as well as with their default correction settings. We focused on FDR 10% data as there were too few identifications for a reliable calculation of FDR 5%. Just 171 cross-linked PSMs passed for the unprocessed data, whereas for the default search, 1088 PSMs passed (536% increase). Of those, 862 (79%) were corrected in their monoisotopic precursor peak. These results support our observations with Xi.

In-Search Monoisotopic Peak Assignment Does Not Compromise Search Accuracy

Changing the search could lead to several problems. We already excluded that the increased search space leads to high-scoring decoy matches that in turn reduce the number of identifications at a given FDR cutoff. As an additional validation, we assessed our results against known PDB structures using the HCD data from the pseudocomplex data set (data set 2), at 5% link-FDR. Assuming a crystal structure is correct, a cross-link can be unexpectedly long either because the link is false or because of in-solution structural dynamics. If, however, the proportion of long-distance links in results of two approaches is identical, then at least the two results have equal quality.

We first tested the results of all three approaches against crystal structures. Residue pairs were mapped to PDB structures and the distance between the two alpha-carbons was calculated (see Methods). Thirty Å was set as the maximal distance for BS3, links with a greater distance were classified as long-distance. In this evaluation, we excluded the protein C3B because its flexible regions make it unsuitable for this analysis. For MaxQuant and OpenMS preprocessed results, 11.8% and 6.1% long-distance cross-links were identified, respectively. In Xi-MPA, 8.1% long distance cross-links were identified (Figure 3A). Of the links uniquely identified through Xi-MPA, only 5.3% were long distance links. Therefore, Xi-MPA as such does not lead to an enrichment in long-distance cross-links. However, it could be that mass-corrected precursors tend to have a higher proportion of long-distance links. We therefore split the Xi-MPA results into five groups corresponding to the monoisotopic mass change (0, −1, −2, −3, −4 Da) and looked at their match to crystal structures. If a link originated from PSMs with different mass corrections, all of those were considered. We conducted a "nonparametric ANOVA" (Kruskal−Wallis test) to detect any significant changes in the distance distributions of Xi-MPA identifications with different

DOI: 10.1021/acs.jproteome.8b00600
J. Proteome Res. 2018, 17, 3923−3931

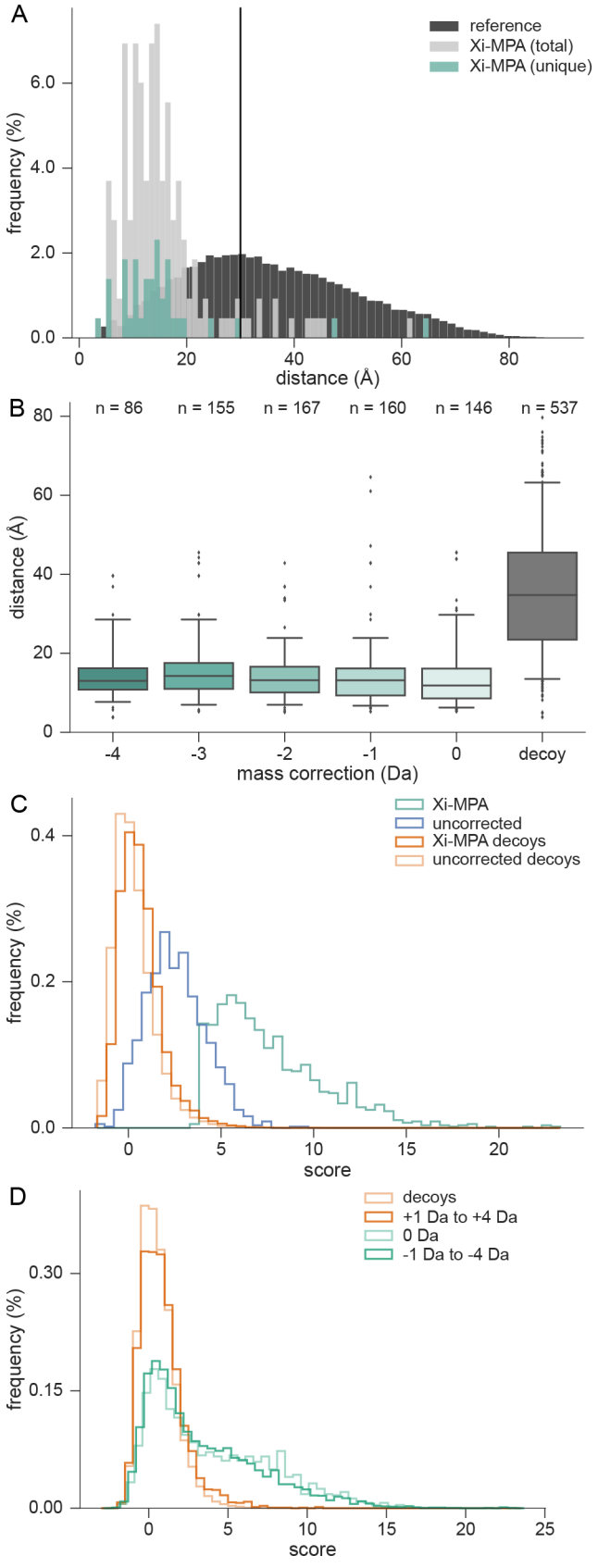


Figure 3. Matches with and without in-search mass shift show similar quality metrics. (A) Evaluation of Xi-MPA derived links on crystal structures (data set 2). Distances between α carbon atoms of identified cross-linked residues in the crystal structure of the proteins are shown in light gray while a reference distribution of all possible pairwise C-alpha distances of cross-linkable residues is shown in dark gray. Thirty Å is set as a limit, above which links are defined as long distance. (B) Distance distribution of identifications with different mass corrections. There was no significant difference between the different mass shifts, while all had a significant difference to the decoy distribution. (C) PSM scores of spectra identified with a mass shift are significantly higher than the corresponding score in uncorrected data. Shown are the score distributions of uncorrected and Xi-MPA results, as well as the corresponding decoy distribution. (D) Score distribution of PSM matches of the "decoy mass search". Identifications with a positive mass shift generally follow the decoy distribution (note that there are correct identifications with a positive mass shift, albeit few, see Figure 1B) while identifications with a negative shift resemble unshifted identifications. The scores of negative-shifted PSMs are significantly higher than those of positive-shifted PSMs (one-sided Wilcoxon test, p-value: $<2.2\times 10^{-16}$).

shifts and decoy distribution. However, we fail to reject the null hypothesis at the predetermined significance level of α = 0.05 (p-value: 0.13), indicating that the distance distributions for all subsets are similar. This matches the visual inspection of distance distributions (Figure 3B). Furthermore, all individual distance distributions were significantly smaller than the derived reference distribution (one-sided Wilcoxon test, see SI Table S5). In conclusion, we do not see any evidence of in-search monoisotopic mass assignment leading to increased conflicts with crystal structures. We then evaluated the effect of in-search monoisotopic mass assignment on PSM quality as assessed by the search score. First, we compared the scores of PSMs with a mass shift (Xi-MPA identifications) to the scores of the same spectrum without a mass shift (uncorrected data). While scores with shifted mass have a median of 6.7, the median score is 2.3 when using the uncorrected masses (Figure 3C). As one would expect from an increased search space, the scores of decoy hits also improve, albeit only marginally. We find that the score difference of target PSMs is significantly larger than of decoy PSMs (one-sided Wilcoxon test, p-value: $<2.2 \times 10^{-16}$). We then turned to a "decoy mass search" for which we not only searched the range from 0 Da to −4 Da, but also +1 Da to +4 Da. Assuming the monoisotopic peak in the uncorrected data is rarely lighter than the true monoisotopic peak, the new identifications should score like decoy identifications. Indeed, the resulting score distributions for targets with a positive mass shift follow the decoy distribution (Figure 3D). In contrast, identifications with a negative shift are distributed like the identifications without mass shift. In conclusion, in-search monoisotopic mass change leads to significantly improved scores with a distribution that resembles that of precursors that did not see a mass change (0 Da). Importantly, these improvements are not random since an equally large search space increase (+1 Da to +4 Da) results in a completely different score distribution that resembles the decoy distribution but not the distribution of identifications without a mass shift.

Heavy and Low Intensity Peptides Are Corrected More Frequently

One would especially expect to observe shifted mass assignment for peptides of high mass and low abundance. For large peptides (approximately >2000 Da), the monoisotopic peak will not be the most intense peak in the isotope cluster. If the peptide is of low abundance, the monoisotopic

DOI: 10.1021/acs.jproteome.8b00600
J. Proteome Res. 2018, 17, 3923−3931

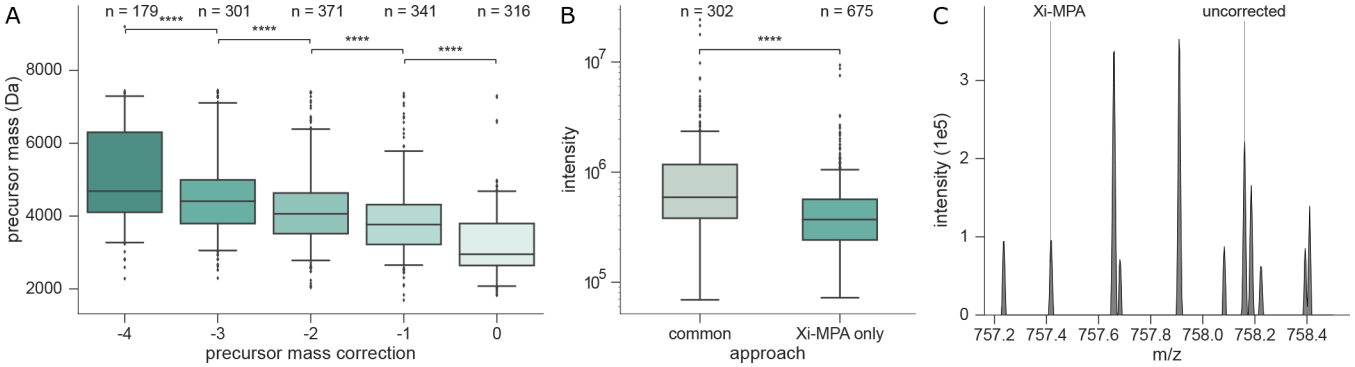


Figure 4. Correction is dependent on precursor mass and intensity. (A) Box plot of the precursor mass and monoisotopic mass correction of identified PSMs after Xi-MPA. PSMs with higher mass more often require monoisotopic mass correction to lighter masses. Whiskers show the 5 and 95% quantiles of the data. Asterisks denote the significance calculated by a one tailed *t* test (****: *p*-value: <0.0001). (B) Precursors of cross-linked PSMs identified in all three approaches, MaxQuant-Xi, OpenMS-Xi, and Xi-MPA ("common"), are more intense than precursors of PSMs that are only identified in Xi-MPA. In other words, successful correction happens more often for abundant precursors, whereas Xi-MPA identifies precursors of lower intensity. (C) MS1 isotope cluster of a cross-linked peptide. The monoisotopic peak of m/z 758.16 ($z = 4$, $m = 3028.6$ Da) was falsely assigned during acquisition and not corrected in any preprocessing approach. Xi-MPA identifies a PSM for a precursor with a mass that is 3 Da lighter.

peak may be of too low intensity to be detected. We therefore analyzed the monoisotopic peak assignment in Xi-MPA regarding the precursor mass and intensity. Indeed, precursors with higher masses are more often corrected to lighter monoisotopic peaks (Figure 4A). While the median precursor mass for uncorrected matches is 2952 Da, for matches corrected by −2 Da it is already 4062 Da and for −4 Da it is 4684 Da. Of the identifications with a mass above 3000 Da, 88% were identified with a lighter mass. For precursors lighter than 3000 Da, the proportion was 42%. Like for mass dependency, there is a trend toward larger correction ranges for lower intensity peptides (SI Figure S5). However, this is less strong than it is for precursor mass.

When evaluating the newly matched precursors of Xi-MPA, the advantage of not having to rely on MS1 identification is evident. Matches not made through any of the preprocessing methods are generally much less intense (Figure 4B) and larger (SI Figure S6) than matches that are common to all approaches. Manual analysis of isotope clusters of corrected precursors from data set 2 revealed many cases where the monoisotopic peak was present in the MS1 spectrum but was not recognized during acquisition. For some, this might be due to the peak being of low intensity and discarded as noise, or because of other interfering peaks (Figure 4C). However, there are also cases where the cluster is well resolved (Figure 1C). Without details of how the instrument software determines the monoisotopic peak, a full evaluation is difficult. For a complete list of precursor m/z for Xi-MPA identifications and corresponding m/z of uncorrected, MaxQuant and OpenMS data, see SI Table S1−S3.

Note that in many acquisition methods, the machine only fragments peaks where it can successfully identify a full isotope cluster. Therefore, there might be instances of cross-linked peptides not being fragmented because of insufficient isotopic cluster quality, leading to lost identifications.

CONCLUSION

The size and low abundance of cross-linked peptides leads to frequent misassignment of the monoisotopic mass by instrument software, which in some instances even escapes correction by sophisticated correction approaches employed by MaxQuant and OpenMS. Considering multiple monoisotopic masses during search increases the number of cross-link PSMs 1.8−4.2-fold, without compromising search accuracy as judged by multiple assessment strategies including comparison of the gains against solved protein structures. The problem of wrongly assigned monoisotopic peaks will have an impact on most cross-link search engines since these all rely in some part on the precursor mass. The extent of the misassignment will however be sample and software-dependent. Even with improved acquisition or correction software, there will remain instances where the monoisotopic peak cannot be determined correctly before searching due to low intensity. Our search-assisted monoisotopic peak assignment provides a general solution to this problem by relying on MS2 identification in addition to precursor information.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.8b00600.

Table S4: Xi-MPA mass range reduction. Figure S1: OpenMS preprocessing workflow. Figure S2: Performance of Xi-MPA on EThcD, CID, and ETciD acquisitions of the pseudocomplex data set. Table S5: Summary for conducted significance tests for Figure 3B in the main text. Figure S3: Performance of Xi-MPA on HCD data of the HSA data set. Figure S4: Performance of Xi-MPA on the test fractions of the *C. thermophilum*data set. Figure S5: Dependency of the monoisotopic mass correction on precursor intensity. Figure S6: Dependency of identifications in Xi-MPA on mass (PDF)

Table S1: Precursor m/z of different processing methods of data set 1. Table S2: Precursor m/z of different processing methods of data set 2. Table S3: Precursor m/z of different processing methods of data set 3. Supporting Information: MS1 based mass range reduction (XLS)

DOI: 10.1021/acs.jproteome.8b00600
J. Proteome Res. 2018, 17, 3923−3931

AUTHOR INFORMATION

Corresponding Author

*E-mail: juri.rappsilber@tu-berlin.de.

ORCID

Swantje Lenz: 0000-0002-8839-5371
Sven H. Giese: 0000-0002-9886-2447
Lutz Fischer: 0000-0003-4978-0864
Juri Rappsilber: 0000-0001-5999-1310

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Dr Francis O'Reilly for comments and helpful discussions. This work was supported by the Einstein Foundation, the DFG [RA 2365/4-1], and the Wellcome Trust through a Senior Research Fellowship to JR [103139] and a multiuser equipment grant [108504]. The Wellcome Centre for Cell Biology is supported by core funding from the Wellcome Trust [203149].

REFERENCES

(1) Leitner, A.; Reischl, R.; Walzthoeni, T.; Herzog, F.; Bohn, S.; Förster, F.; Aebersold, R. Expanding the chemical cross-linking toolbox by the use of multiple proteases and enrichment by size exclusion chromatography. *Mol. Cell. Proteomics* **2012**, *11*, M111.014126.

(2) Chen, Z. A.; Jawhari, A.; Fischer, L.; Buchen, C.; Tahir, S.; Kamenski, T.; Rasmussen, M.; Lariviere, L.; Bukowski-Wills, J.-C.; Nilges, M.; et al. Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry. *EMBO J.* **2010**, *29*, 717−726.

(3) Tan, D.; Li, Q.; Zhang, M.-J.; Liu, C.; Ma, C.; Zhang, P.; Ding, Y.-H.; Fan, S.-B.; Tao, L.; Yang, B. et al. Trifunctional cross-linker for mapping protein-protein interaction networks and comparing protein conformational states. *Elife* **2016**, *5*, DOI: DOI: 10.7554/eLife.12509.

(4) Rampler, E.; Stranzl, T.; Orban-Nemeth, Z.; Hollenstein, D. M.; Hudecz, O.; Schlögelhofer, P.; Mechtler, K. Comprehensive Cross-Linking Mass Spectrometry Reveals Parallel Orientation and Flexible Conformations of Plant HOP2-MND1. *J. Proteome Res.* **2015**, *14*, 5048−5062.

(5) Mendes, M. L.; Fischer, L.; Chen, Z. A.; Barbon, M.; O'Reilly, F. J.; Bohlke-Schneider, M.; Belsom, A.; Dau, T.; Combe, C. W.; Graham, M. et al. An integrated workflow for cross-linking/mass spectrometry. 2018.

(6) Belsom, A.; Schneider, M.; Fischer, L.; Mabrouk, M.; Stahl, K.; Brock, O.; Rappsilber, J. Blind testing cross-linking/mass spectrometry under the auspices of the 11thcritical assessment of methods of protein structure prediction (CASP11). *Wellcome open research* **2016**, *1*, 24.

(7) Kolbowski, L.; Mendes, M. L.; Rappsilber, J. Optimizing the Parameters Governing the Fragmentation of Cross-Linked Peptides in a Tribrid Mass Spectrometer. *Anal. Chem.* **2017**, *89*, 5311−5318.

(8) 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**, *8*, 15473.

(9) Orbán-Németh, Z.; Beveridge, R.; Hollenstein, D. M.; Rampler, E.; Stranzl, T.; Hudecz, O.; Doblmann, J.; Schlögelhofer, P.; Mechtler, K. Structural prediction of protein models using distance restraints derived from cross-linking mass spectrometry data. *Nat. Protoc.* **2018**, *13*, 478−494.

(10) Schneider, M.; Belsom, A.; Rappsilber, J. Protein Tertiary Structure by Crosslinking/Mass Spectrometry. *Trends Biochem. Sci.* **2018**, *43*, 157−169.

(11) Hoopmann, M. R.; Zelter, A.; Johnson, R. S.; Riffle, M.; MacCoss, M. J.; Davis, T. N.; Moritz, R. L. Kojak: efficient analysis of chemically cross-linked protein complexes. *J. Proteome Res.* **2015**, *14*, 2190−2198.

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

(13) Yang, B.; Wu, Y.-J.; Zhu, M.; Fan, S.-B.; Lin, J.; Zhang, K.; Li, S.; Chi, H.; Li, Y.-X.; Chen, H.-F.; et al. Identification of cross-linked peptides from complex samples. *Nat. Methods* **2012**, *9*, 904−906.

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

(15) Renard, B. Y.; Kirchner, M.; Monigatti, F.; Ivanov, A. R.; Rappsilber, J.; Winter, D.; Steen, J. A. J.; Hamprecht, F. A.; Steen, H. When less can yield more - Computational preprocessing of MS/MS spectra for peptide identification. *Proteomics* **2009**, *9*, 4978−4984.

(16) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **2007**, *4*, 923−925.

(17) Fischer, L.; Rappsilber, J. Quirks of Error Estimation in Cross-Linking/Mass Spectrometry. *Anal. Chem.* **2017**, *89*, 3829−3833.

(18) Walzthoeni, T.; Claassen, M.; Leitner, A.; Herzog, F.; Bohn, S.; Förster, F.; Beck, M.; Aebersold, R. False discovery rate estimation for cross-linked peptides identified by mass spectrometry. *Nat. Methods* **2012**, *9*, 901−903.

(19) Maiolica, A.; Cittaro, D.; Borsotti, D.; Sennels, L.; Ciferri, C.; Tarricone, C.; Musacchio, A.; Rappsilber, J. Structural analysis of multiprotein complexes by cross-linking, mass spectrometry, and database searching. *Mol. Cell. Proteomics* **2007**, *6*, 2200−2211.

(20) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26*, 1367−1372.

(21) Tyanova, S.; Temu, T.; Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **2016**, *11*, 2301−2319.

(22) Sturm, M.; Bertsch, A.; Gröpl, C.; Hildebrandt, A.; Hussong, R.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Zerck, A.; Reinert, K.; et al. OpenMS - an open-source software framework for mass spectrometry. *BMC Bioinf.* **2008**, *9*, 163.

(23) Röst, H. L.; Sachsenberg, T.; Aiche, S.; Bielow, C.; Weisser, H.; Aicheler, F.; Andreotti, S.; Ehrlich, H.-C.; Gutenbrunner, P.; Kenar, E.; et al. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat. Methods* **2016**, *13*, 741−748.

(24) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20*, 1466−1467.

(25) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **2013**, *13*, 22−24.

(26) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. Open mass spectrometry search algorithm. *J. Proteome Res.* **2004**, *3*, 958−964.

(27) Yuan, Z.-F.; Liu, C.; Wang, H.-P.; Sun, R.-X.; Fu, Y.; Zhang, J.-F.; Wang, L.-H.; Chi, H.; Li, Y.; Xiu, L.-Y.; et al. pParse: a method for accurate determination of monoisotopic peaks in high-resolution mass spectra. *Proteomics* **2012**, *12*, 226−235.

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

(29) Giese, S. H.; Belsom, A.; Rappsilber, J. Optimized Fragmentation Regime for Diazirine Photo-Cross-Linked Peptides. *Anal. Chem.* **2016**, *88*, 8239−8247.

(30) Kastritis, P. L.; O'Reilly, F. J.; Bock, T.; Li, Y.; Rogon, M. Z.; Buczak, K.; Romanov, N.; Betts, M. J.; Bui, K. H.; Hagen, W. J.; et al. Capturing protein communities by structural proteomics in a thermophilic eukaryote. *Mol. Syst. Biol.* **2017**, *13*, 936.

DOI: 10.1021/acs.jproteome.8b00600
J. Proteome Res. 2018, 17, 3923−3931

(31) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **2008**, *24*, 2534−2536.
(32) Weisser, H.; Nahnsen, S.; Grossmann, J.; Nilse, L.; Quandt, A.; Brauer, H.; Sturm, M.; Kenar, E.; Kohlbacher, O.; Aebersold, R.; et al. An automated pipeline for high-throughput label-free quantitative proteomics. *J. Proteome Res.* **2013**, *12*, 1628−1644.
(33) Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Slagel, J.; Sun, Z.; Moritz, R. L. Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. *Proteomics: Clin. Appl.* **2015**, *9*, 745−754.
(34) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. *Anal. Chem.* **2002**, *74*, 5383−5392.
(35) Vizcaíno, J. A.; Csordas, A.; del-Toro, N.; Dianes, J. A.; Griss, J.; Lavidas, I.; Mayer, G.; Perez-Riverol, Y.; Reisinger, F.; Ternent, T.; et al. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **2016**, *44*, D447−56.

DOI: 10.1021/acs.jproteome.8b00600
J. Proteome Res. 2018, 17, 3923−3931

Manuscript 2:

Improved peptide backbone fragmentation is the primary advantage of MS-cleavable crosslinkers.

Included here as preprint, manuscript now available online
DOI: 10.1021/acs.analchem.1c05266

Improved peptide backbone fragmentation is the primary advantage of MS-cleavable crosslinkers

Lars Kolbowski[*,1], Swantje Lenz[*,1], Lutz Fischer[1], Ludwig R Sinn[1], Francis J O'Reilly[1], Juri Rappsilber[1,2]

[*] These authors contributed equally to this work

[1] Technische Universität Berlin, Chair of Bioanalytics, 10623 Berlin, Germany

[2] University of Edinburgh, Wellcome Centre for Cell Biology, Edinburgh EH9 3BF, UK

Proteome-wide crosslinking mass spectrometry studies have coincided with the advent of MS-cleavable crosslinkers that can reveal the individual masses of the two crosslinked peptides. However, recently such studies have also been published with non-cleavable crosslinkers suggesting that MS-cleavability is not essential. We therefore examined in detail the advantages and disadvantages of using the most popular MS-cleavable crosslinker, DSSO. Indeed, DSSO gave rise to signature peptide fragments with a distinct mass difference (doublet) for nearly all identified crosslinked peptides. Surprisingly, we could show that it was not these peptide masses that proved the main advantage of MS-cleavability of the crosslinker, but improved peptide backbone fragmentation which reduces ambiguity of peptide identifications. We also show that the more intricate MS3-based data acquisition approaches lack sensitivity and specificity, causing them to be outperformed by the simpler and faster stepped HCD method. This understanding will guide future developments and applications of proteome-wide crosslinking mass spectrometry.

INTRODUCTION

Crosslinking combined with mass spectrometry (Crosslinking MS) is a powerful tool for detecting protein-protein interactions and the structural characterization of proteins. Many key advances have been made in recent years to expand the complexity of the samples that can be analysed with this technology. These include the database search software[1–3], FDR estimation[4] and the enrichment of crosslinked peptides[5–7]. One of the key problems when identifying crosslinked peptides is that one must in principle identify two peptides from the same MS1 signal. The search space is therefore initially very large, comprising every pairwise combination of the peptides that are in the database, i.e. (n2+n)/2 crosslinked peptides. This large search space can be reduced experimentally by separating the crosslinked peptides during the measurement by help of an MS-cleavable crosslinker such as disuccinimidyl sulfoxide (DSSO)[8] or any of its alternatives[9].

The conceptual advantage of MS-cleavable crosslinkers is evident. The crosslinker readily cleaves upon activation in the mass spectrometer, releasing the individual peptides and thereby enabling the measurement of their individual masses. In the case of the most popular MS-cleavable crosslinker DSSO, the crosslinker cleaves preferentially at two different sites, leading to different crosslinker remnants (also called stubs) for each peptide (**Fig. 1A**). The asymmetric cleavage of this crosslinker produces a pair of alkene (A) and sulfenic acid (S) stub fragments[8]. The S stub fragment commonly loses water forming the unsaturated thiol (T). The two most frequently observed stub peaks per peptide, the A and the T fragment, form a signature doublet signal with a distinct mass difference, allowing their detection and subsequent calculation of peptide masses[10].

Knowing the individual peptide masses simplifies the database search, as it reduces the search space to pairwise combinations of peptides with these masses. With the individual peptides released in the mass spectrometer, one can also design more intricate data acquisition approaches. The two peptides can be fragmented individually using MS3, which provides separate fragment information of the two - now linear - peptides. For this, generally the crosslinked peptide is fragmented with a low-energy CID fragmentation first, to preferentially cleave the crosslinker instead of the peptide backbone. Then signature doublets are selected for MS3. This approach is routinely employed by studies that use the PIR crosslinker[7] and DSSO, while some others using DSSO supplement this with a complementary ETD MS2 spectrum[11].

In an alternative acquisition method, stepped HCD (sHCD), only a single MS2 spectrum is recorded for each crosslinked peptide pair. The peptide is subjected to multiple different collision energies and the fragments are recorded in a single MS2 spectrum. This spectrum should contain the signature doublet (from lower fragmentation energies) as well as additional backbone fragments (from higher fragmentation energies). These spectra can be searched in most crosslinking search tools, with optional filtering for spectra containing cleaved signature peaks during[2] or after[12] search.

Despite the clean crosslinker cleavage producing dominant signature peaks in proof-of-concept data of either approach, there is a lack of statistical data of how often this happens in general. It is unclear how many crosslinked peptides give rise to doublets, how prominent these doublets are, and how successful doublet selection is at covering the peptides. It is therefore unknown how many crosslinked spectra are left unidentified when relying on these doublets. sHCD compared

Table 1. Overview of analysed datasets.

sample	crosslinker	acquisition method	variable modifications used in re-analysis	ref
E. coli lysate	DSSO	stepped HCD (sHCD)	oxidation (M), methylation (D, E), deamidation (N, Q), BS3/DSSO -OH; -NH2 (K, nterm)	4
	BS3			
M. musculus synaptosomes	DSSO	CID-MS2-MS3+ETD-MS2	DSSO -OH; -NH2 (K, nterm)	15
E. coli 70S ribosome	DSSO	stepped HCD (sHCD)	DSSO -OH; -NH2 (K, nterm)	13
		CID-MS2-HCD-MS3		

favourably to CID methods in the number of crosslinks identified[13], but a methodical analysis comparing the information contained in their fragmentation spectra is missing and yet is crucial for future design of crosslinkers and acquisition methods.

MS-cleavable crosslinkers have been the tool of choice in many proteome-wide crosslinking MS studies, and it has been suggested that large-scale crosslinking MS depends on MS-cleavable crosslinkers[14]. While conceptually appealing, these advantages and potential limitations of MS-cleavable crosslinkers have yet to be analysed in detail in 'real world' scenarios - some comparisons exist, but usually only comparing a few crosslink spectrum matches (CSMs). We systematically investigated the influence of the popular MS-cleavable crosslinker DSSO on the fragmentation of crosslinked peptides. We achieve this by using crosslinker search software that does not rely on the cleaved stubs for identification. This allowed us to clarify how wide-spread the cleavage of DSSO actually is, and to probe the gain of knowing the individual peptide masses for identifying crosslinks.

METHODS

Database search and FDR filtering

Mass spectrometry raw data were processed using MSconvert[23] (v3.0.11729) to convert to mgf-file format. A linear peptide search was employed to determine median precursor and fragment mass errors. Peak list files were then recalibrated to account for mass shifts during measurement prior to analysis using xiSEARCH[3] 1.7.6.1 with the following settings: MS1 error tolerances of 3 ppm; MS2 error tolerance of 5 ppm for the *E. coli* lysate dataset and 15 ppm for the others; up to two missing precursor isotope peaks; tryptic digestion specificity with up to two missed cleavages; modifications: carbamidomethylation (Cys, +57.021464 Da) as fixed and oxidation (Met, +15.994915 Da), deamidation (Asn and Gln, +0.984016 Da), methylation (Glu and Asp, +14.015650 Da), amidated crosslinker (Lys and protein N-terminus, DSSO-NH2: +175.03031 Da; BS3-NH2: 155.09463 Da) and hydrolysed crosslinker (Lys and protein N-terminus, DSSO-OH: +176.01433 Da; BS3-OH: +156.07864 Da) as variable modifications; Maximum number of variable modifications per peptide: 1; losses: –CH3SOH, –H2O, –NH3 and additionally masses for crosslinker-containing ions were defined accounting for its cleavability (A: 54.01056 Da, S: 103.99320 Da, T: 85.98264). Crosslink sites for both reagents were allowed for side chains of Lys, Tyr, Ser, Thr and the protein N-terminus. Note that we included a "non-covalent" crosslinker with a mass of zero to flag spectra potentially arising from gas-phase associated peptides[24]. These spectra were removed prior to false-discovery-rate (FDR) estimation. Results were filtered prior to FDR to matches having a minimum of three matched fragments per peptide, a delta score of > 15% of the match score and a peptide length of at least six amino acids. Additionally, identifications of peptide sequences that are found in two or more proteins were removed. FDR was estimated using xiFDR[16] (v2.1.2) on a unique CSM level to 5% grouped by self- and heteromeric matches.

Data evaluation

CSMs passing FDR were re-annotated with pyXiAnnotator (https://github.com/Rappsilber-Laboratory/pyXiAnnotator/) with peptide, b-, and y-type ions using MS2 tolerances as described above. The resulting matched fragments were used to check for the occurrence of DSSO A-T doublets and to calculate fragment sequence coverages. We calculated the sequence coverage for our CSMs conservatively, as the ratio of matched N-terminal and C-terminal sequence fragments to the number of theoretically possible sequence fragments (i.e., 100% sequence coverage would mean the detection of at least one fragment from the N-terminal and one from the C-terminal series between all amino acid residues of a peptide). To evaluate the MS3 triggering behaviour the MS3 precursor m/z was extracted from the scan header and compared with the fragment annotation result of the corresponding MS2 CSM. If the MS3 precursor matched a crosslinked peptide stub fragment with 20 ppm error tolerance it was counted as correctly triggered.

RESULTS AND DISCUSSION

Prevalence of peptide doublets in fragmentation spectra of DSSO crosslinked peptides

We analysed three publicly available datasets of DSSO crosslinking experiments coming from three different labs, differing in acquisition method and sample complexity (**Table 1**). The dataset of crosslinked *E. coli* lysate was acquired using sHCD with a low, medium, and high normalized collision energy for each MS2[4]. sHCD is also one of two acquisition methods used to record a dataset of crosslinked, purified 70S ribosomes[13]. In addition to this, Stieger et. al also employed a CID-MS2-HCD-MS3 approach. For this, first a low-energy CID-MS2 was acquired. Then MS3 was triggered when doublets of the correct mass difference (32 Da for A-T) were detected (**Fig. 1A**). Finally, the third dataset called here "Synapse dataset" covered crosslinked mouse synaptosomes and was acquired with a CID-MS2-MS3+ETD-MS2 approach[15]. As in the Ribosome dataset, a low-energy CID-MS2 was acquired for doublet detection. Then, MS3 was acquired as described above, supplemented by an additional ETD-MS2 on the same MS1 precursor.

To assess the prevalence of doublets in the fragmentation spectra of crosslinked peptides, we re-searched the datasets using a search algorithm that does not rely on peptide doublets

for crosslink identification. After database search and filtering to 5% heteromeric (inter protein) CSM-level FDR[16], we looked for signature A and T stub fragment doublet peaks of the identified peptides and the intensity rank of these doublets in each spectrum (**Fig. 1A**).

Even though we did not require doublets to identify crosslinked peptides, they were very common features in our CSMs. We found doublets frequently for at least one peptide, independent of dataset and acquisition method (90 - 98%) (**Fig. 1B**). The CID acquisitions displayed a higher proportion of CSMs with both peptide doublets detected compared to the sHCD datasets. If one looks at only the common identifications of CID and HCD to make up for the difference in number of identifications, the amount of doublets detected for both peptides increases noticeably for sHCD (71%), making the difference to CID (81%) less pronounced (**Fig. 1B**) as does considering only single stub peaks (**Fig. S1**).

We next looked at the intensity of the doublet peaks across these datasets, as this is important for their use during acquisition and data analysis (**Fig. 1C**). In the majority of the spectra, the more abundant doublet is among the most intense peaks, independent of the fragmentation method used. In fact, a doublet peak is frequently the most abundant peak (34 - 53% of the doublet containing spectra). Almost all (94 - 98%) doublet-containing spectra have a peak of the more intense peptide doublet among the 20 most intense peaks.

Spectra typically displayed in publication figures suggest that also the less intense doublet is seen prominently in CID spectra. However, this was only the case for 10% (Ribosome) or 20% (Synapse) of the doublet containing CID spectra of our investigated data. Nevertheless, it is seen among the top 20 peaks in 78% (Ribosome) or 91% (Synapse) of the doublet containing CID spectra. For the sHCD data, the doublet ranks are lower, yet still approximately 70% of spectra have them among the 20 most intense peaks (**Fig. 1C**).

In conclusion, the first doublet is among the most intense peaks for the majority of CSMs independent of the fragmentation method. While the second doublet increases confidence in doublet calling, only one peptide doublet is necessary for deriving both peptide masses given that we know the precursor mass. The visibility of the second peptide doublet is crucial, however, for the successful selection of both peptides for MS3. We therefore investigated how successful selecting doublets from CID-MS2 spectra for MS3 was at covering one or both crosslinked peptides, and if this more complex approach produces more confident identifications than HCD-MS2.

Speed of HCD outperforms higher sequence coverage of CID+MS3

The ratio of identified doublets and their intensity ranks are important criteria for selecting peptides for MS3 fragmentation. However, absolute numbers of crosslink identifications may also be influenced by other aspects, such as backbone fragmentation and acquisition speed. We used the Ribosome dataset to compare these aspects, as it uses both methods on the same sample. Here, sHCD leads to 1.4 times more residue pairs identified than CID-MS3[17].

When comparing the common CSMs between CID and sHCD, the overall sequence coverage in sHCD is higher compared to low-energy CID (**Fig. 2A**). This comes as no surprise, as low-energy CID is primarily applied to separate the crosslinked peptides and not for peptide backbone fragmentation. It is intentionally combined with MS3 scans and ETD fragmentation to provide additional sequence information. When we include the corresponding MS3 scans, the sequence coverage increases noticeably compared to that of low-energy CID alone. The overall coverage from combining fragments from CID and MS3 surpasses the sHCD coverage. Therefore, the backbone fragmentation does not explain the higher number of CSMs for sHCD.

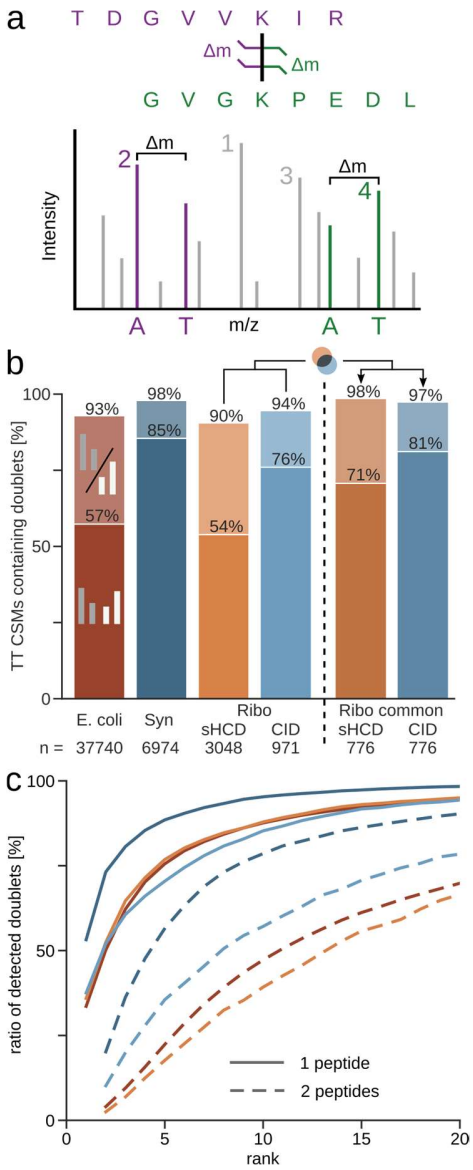


Figure 1. Statistics on frequency and intensity of peptide doublet peaks. (a) Illustration of DSSO cleavage and the resulting signature peptide doublets with the distinct mass difference Δm. Numbers annotate the intensity rank of the peaks, with the rank of the more intense of the doublet peaks being the rank of the whole doublet. (b) Ratio of identified target-target (TT) CSMs that contain one (lighter colour) or both (darker colour) peptide doublets in each dataset (5% CSM-level FDR). Datasets using sHCD are shown in orange-red while CID-MS3 based methods are in blue colours. (c) Percentage of detected doublets passing each intensity rank cut-off. Shown is the cumulative proportion of CSMs containing doublets. Datasets are coloured as in (b). Synapse (Syn); Ribosome (Ribo).

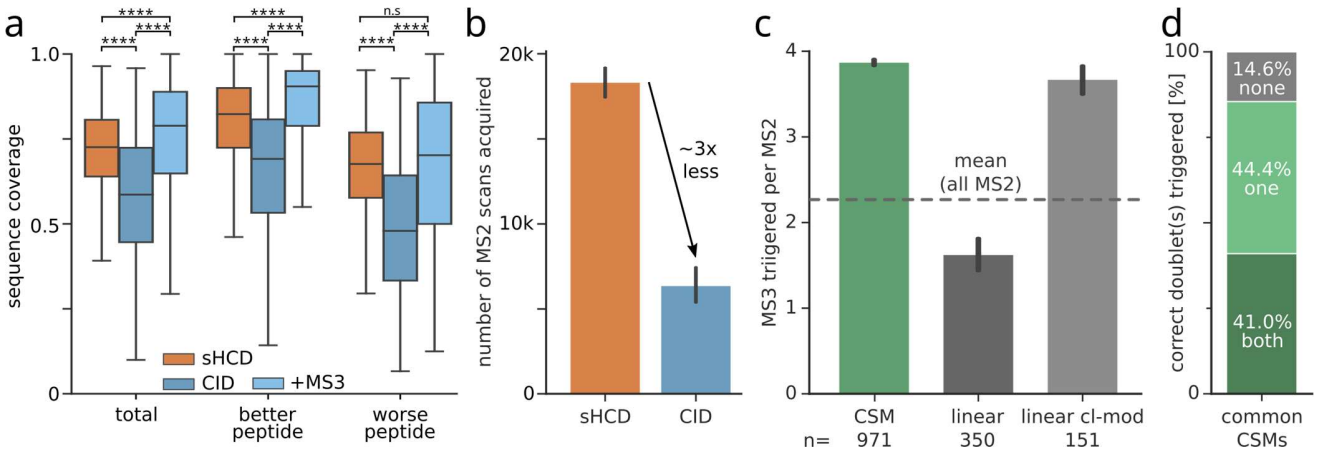


Figure 2. Speed of HCD outperforms higher sequence coverage of CID+MS3 in the Ribosome dataset. (a) Sequence coverage of common CSMs (n=776) identified in both sHCD and CID. Additionally, sequence coverage of CID spectra combined with their respective MS3 scans is shown. Boxplots depict the median (middle line), upper and lower quartiles (boxes), and 1.5 times the interquartile range (whiskers). Asterisks indicate significance calculated by a two-sided Wilcoxon signed-rank (p-value > 0.05: n.s., p-value < 0.0001: ****). (b) Number of acquired MS scans per fragmentation method. Error bars show the 0.95 confidence interval (n=7). (c) Number of triggered MS3 scans per MS2 scan, for CSMs, linear peptide spectrum matches, and crosslinker modified linear peptide spectrum matches, respectively. (d) Proportion of common CID CSMs having no doublets, only one, or both peptide doublets correctly triggered for MS3.

MS3 acquisition schemes require multiple scan and fragmentation events, while sHCD only acquires a single MS2 scan. This difference in complexity and, more importantly, acquisition speed is reflected in the number of total MS2 scans acquired, which on average is almost 3 times lower for the CID-MS3 method, because a lot of acquisition time is spent on acquiring the additional MS3 scans (**Fig. 2B**). The drastically lower sampling of precursors for fragmentation will consequently lead to the reduced detection of crosslinked peptides, which subsequently results in a lower number of crosslink identifications. This is exacerbated by many MS3 spectra being acquired for crosslinker-modified and even for unmodified linear peptides (**Fig. 2C**). Despite this excessive MS3 triggering, for only 41% of the CSMs, MS3 was triggered correctly on both peptide doublets (**Fig. 2D**). This is also reflected in the wider spread of sequence coverage for the worse fragmented peptide, which is crucial for unambiguous identification of both linked peptides (**Fig. 2A**). Note also that for this peptide the sequence coverage is not significantly increased in CID+MS3 over sHCD.

In this dataset, the speed of sHCD compensates for its slightly lower sequence coverage. sHCD also shows a more symmetric fragmentation of both peptides, as the MS3 approach is limited by its dependency on triggering on the correct doublets. Further development of MS3 approaches should focus on a more sensitive and selective MS3 selection, which in part is governed by the yield of the crosslinker cleavage.

Peptide doublets for quality control

While some database search algorithms have been built around peptide masses from doublets, others have been built without relying on them. Unarguably, peptide masses are useful information. In an attempt to quantify their value, we investigated the target-decoy CSMs (as representation of the random matches) for the occurrence of peptide doublets. Because heteromeric CSMs are the focus of most biological research questions and are also more challenging to identify, we focused on those for the analysis.

A substantial fraction of random matches have matching peptide doublets (>47% of heteromeric target-decoy CSMs, **Fig. 3A**). However, their extent varies considerably between the datasets. The highest proportion of doublets among target-decoy CSMs is found in the Ribosome dataset (88% or 92% for sHCD and CID, respectively). The *E. coli* dataset contains at least one doublet in 66% of the target-decoy CSMs, while this proportion decreases to 47% for the Synapse dataset. The amount of identified doublets present in target-decoy matches seems less dependent on the acquisition method, and more on the sample and database.

Although heteromeric target-decoy CSMs contain peptide doublets, they do so less often than the heteromeric target-target matches (**Fig. S2**). Based on this difference, we investigated the effect of using this metric as a quality filter. We prefiltered the search results to those spectra that contain at least one peptide with a detected doublet and then re-estimated 5% CSM-level FDR. The gains using this approach are very much dependent on the complexity of the dataset (**Fig. 3B**). Unsurprisingly, the Synapse dataset, which had the least target-decoys containing a matching doublet, shows the largest gains using this approach (19%). However, the *E. coli* dataset only gains 5% in heteromeric CSMs, even though there is a large difference in the proportion of peptide doublets between target and false matches (97% vs. 66%; **Fig. S2, 3A**). This led us to investigate the score distribution of doublet containing matches in more detail (**Fig. 3C**).

The vast majority of high-scoring target-target CSMs contain at least one doublet and are therefore not removed, while targets without a matched peptide doublet tend to have lower scores. In this lower scoring region, there is a steep increase in target-decoy matches, which is only slightly reduced by prefiltering for a doublet. The effect becomes more apparent when looking at the FDR at different score thresholds. While the increase in error is not as steep for the filtered matches as for the unfiltered, it still grows exponentially (**Fig. 3D**). This holds true also for the Ribosome datasets and to a lesser extent for the Synapse dataset (**Fig. S3-5**).

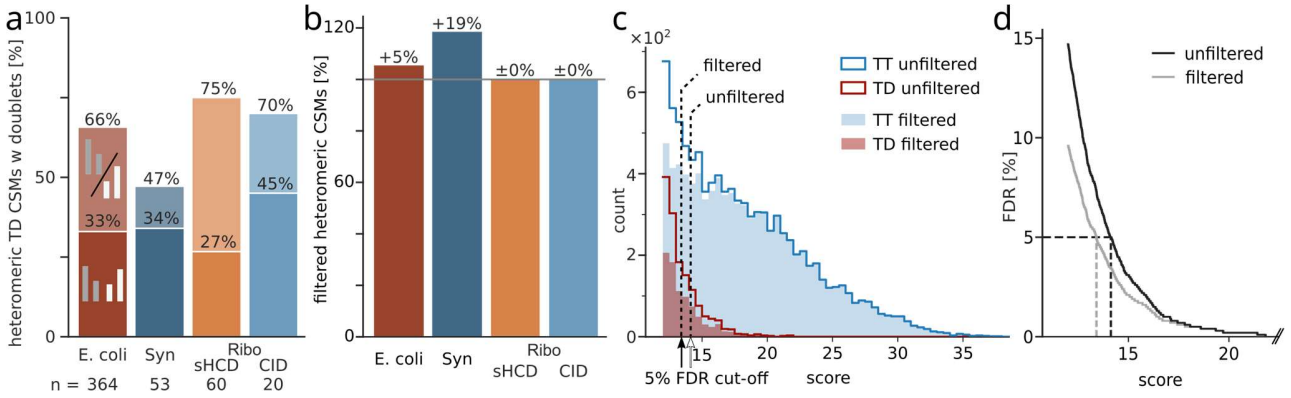


Figure 3. Peptide doublets as a quality control metric for heteromeric identifications. (a) Percentage of heteromeric target-decoy CSMs that contain one or two peptide doublets across datasets at 5% CSM-level FDR. (b) Proportion of heteromeric CSMs at 5% CSM-level FDR when filtering spectra to contain a peptide doublet compared to unfiltered data. (c) Score distribution of heteromeric matches in the *E. coli* dataset. Shown is the distribution of targets and target-decoy matches with and without filtering for peptide doublets. Dashed lines show the resulting score cutoffs at 5% FDR. (d) FDR (interpolated values for visualization) of unfiltered and peptide doublet filtered *E. coli* data. Synapse (Syn); Ribosome (Ribo).

The moderate gains of using doublets for post-search filtering also suggests that using them during search will offer only moderate gains. Presumably, spectra of high quality, which contain doublets, also tend to contain sufficient peptide fragment peaks so that identification is possible without relying on peptide mass information.

Comparison of a cleavable to a non-cleavable crosslinker

Non-cleavable crosslinkers are widely believed to be unsuitable for complex samples[14,18,19]. This bases on the assumption that not knowing the individual peptide masses before the search results in the need for exhaustive combination of all peptides in the database and thus an explosion of the search space. However, there are multiple large-scale studies that have successfully employed a non-cleavable crosslinker despite these assumptions[4,12,20,21]. These are based on a detailed understanding of how crosslinked peptides fragment[22], that offered a computational solution to knowing the individual peptide masses which was then implemented in the search algorithm xiSEARCH[3]. In light of successful usages of both types of crosslinkers, we decided to compare their spectral information to understand any costs and benefits. In addition to DSSO, the published *E. coli* dataset also contains data from the non-cleavable crosslinker BS3. As the data for both crosslinkers were prepared and acquired in a very comparable manner, this dataset offers an opportunity to directly compare the effects of BS3 to DSSO on a complex mixture analysis. Importantly, because of its size and the high number of CSMs identified, the dataset is well suited for statistical evaluation.

A manual side-by-side comparison of CSMs identified in both datasets suggests DSSO to have richer spectra with more fragments. Especially, fragments containing the crosslinking site appear to be more present, mostly as fragments containing an A/S/T stub of DSSO (**Fig. 4A**). We then performed a statistical evaluation of this observation over common CSMs of the two crosslinkers (**Fig. 4B**). This confirmed that DSSO led indeed to a significantly higher sequence coverage than BS3. While the coverage of linear fragments is very similar between the two crosslinkers, the coverage of link site-containing fragments is significantly higher for DSSO. Link site-containing fragments contain the full second peptide (+P) or, additionally for cleavable crosslinkers, just a cleaved crosslinker stub. Indeed, A/S/T stub fragments are the major source of link site-containing fragments for DSSO, while +P coverage is lower than that of BS3. This means that the increased sequence coverage for DSSO stems exclusively from cleaved crosslinker fragments. Crosslinker cleavage appears to promote the cleavage of peptide backbone sites and/or their detection.

The better sequence coverage of DSSO-linked peptides improves the separation of true from false CSMs (**Fig. 4C**). For heteromeric matches, DSSO has a larger area under the curve, and especially more high scoring targets, effectively leading to an increase in heteromeric CSMs. While for BS3 3308 heteromeric CSMs were identified, the DSSO dataset resulted in more than twice as many (7316, +121%) (**Fig. 4D**). For self-CSMs, only 29% more CSMs were identified with DSSO than with BS3 (**Fig. S6**), indicating that self-CSMs are approaching exhaustive coverage at the given experimental detection limit. Similar results were seen when including retention time data of heteromeric and self-CSMs[20].

To investigate the effect of the cleaved crosslinker fragments on the overall crosslink search performance, we performed another search in which the DSSO crosslinker was treated as non-cleavable. In this search, only 1866 heteromeric CSMs were identified (-74%). Filtering these results for doublet containing results, as described before, increased identifications to 3064. This is, however, still a loss of 58% of CSMs compared to the search considering DSSO as cleavable. Collectively, these observations demonstrate that A/S/T stub fragments play a central role in the success of DSSO for crosslinking mass spectrometry, especially for more complex samples.

CONCLUSIONS

Our work finds a surprisingly limited value of doublet information stemming from crosslinker cleavage for the identification of crosslinks. Nonetheless, we find cleavable crosslinkers to lead to the identification of substantially more heteromeric CSMs. We pinpoint improved sequence coverage as the major contributor to this. This has implications for how to conduct crosslinking studies and the future development of the methodology. Firstly, as many suspected but possibly not for the right reasons, cleavable crosslinkers are preferable for crosslink mixture analyses.

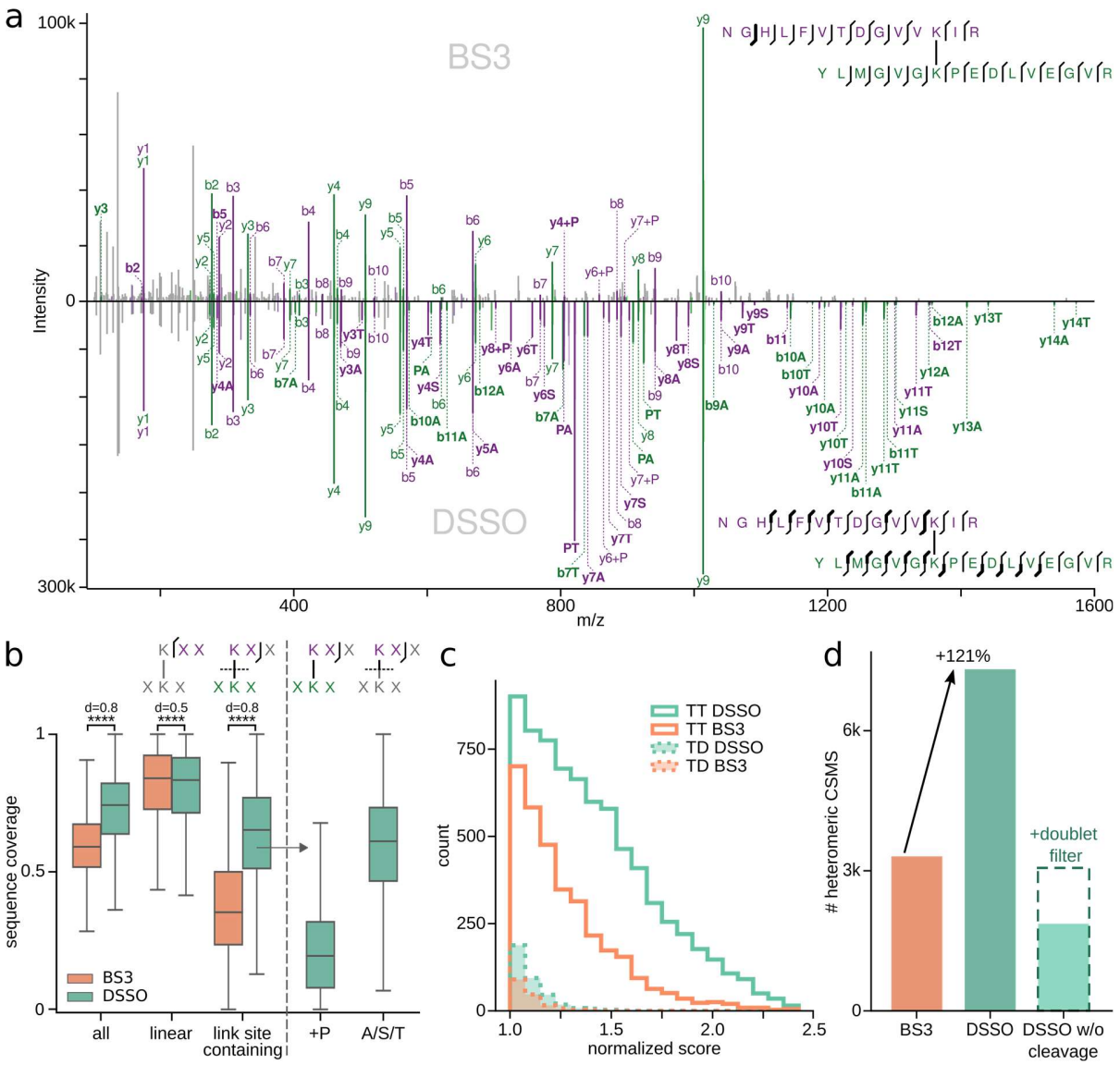


Figure 4. Comparison of non-cleavable crosslinker BS3 to the MS-cleavable crosslinker DSSO. (a) Example MS2 spectrum of a high scoring CSM identified in both datasets. Upper panel shows the CSM from the BS3 dataset. Lower panel shows the same peptide m/z-species identified in the DSSO dataset. Unique fragments are highlighted in bold. (b) Sequence coverage of all, linear and link site-containing fragments (CSMs: n=1437). For DSSO, link site-containing fragments are additionally separated into fragments containing the full second peptide (+P) or only the cleaved crosslinker stub (A/S/T). Boxplots depict the median (middle line), upper and lower quartiles (boxes), and 1.5 times the interquartile range (whiskers). Asterisks indicate significance calculated by a two-sided Wilcoxon signed-rank (p-value < 0.0001: ****). (c) Target-target and target-decoy score distributions of heteromeric CSMs for BS3 and DSSO. Scores were normalized to their respective score cut-off at 10% FDR. (d) Number of heteromeric CSMs passing 5% CSM-level FDR for BS3 and DSSO. As a control, DSSO was additionally searched as a non-cleavable crosslinker and also filtered for the presence of peptide doublets.

Secondly, sHCD is the recommended acquisition method as it achieves almost the same sequence coverage as CID-MS3, but is much faster. CID-MS3 currently lacks speed, specificity and sensitivity. Consequently, future developments of crosslinkers and acquisition methods should focus primarily on sequence information, without compromising acquisition speed. Current choices governing acquisition schemes rely on experimental comparisons, to which we add a methodological understanding of the key parameters that govern crosslink identification. With this, we hope to pave the way for simplified, cost-effective, and standardised workflows that a wider number of labs can use.

ASSOCIATED CONTENT

AUTHOR INFORMATION

Corresponding Author

Juri Rappsilber. Email: juri.rappsilber@tu-berlin.de

Author Contributions

S.L., L.K., L.S., F.J.O., and J.R. designed the experiments. S.L., L.K., and L.F. processed Crosslinking MS data; S.L., L.K., F.J.O., and J.R. prepared figures and wrote the manuscript with input from all authors.

ACKNOWLEDGMENT

The work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2008 – 390540038 – UniSysCat and grant no. 426290502 and by the Wellcome Trust through a Senior Research Fellowship to JR (103139). The Wellcome Centre for Cell Biology is supported by core funding from the Wellcome Trust (203149).

REFERENCES

1. 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).
2. 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).
3. Mendes, M. L. et al. An integrated workflow for crosslinking mass spectrometry. Mol. Syst. Biol. 15, e8994 (2019).
4. Lenz, S. et al. Reliable identification of protein-protein interactions by crosslinking mass spectrometry. Nat. Commun. 12, 3564 (2021).
5. Belsom, A. & Rappsilber, J. Anatomy of a crosslinker. Curr. Opin. Chem. Biol. 60, 39–46 (2021).
6. Steigenberger, B., Pieters, R. J., Heck, A. J. R. & Scheltema, R. A. PhoX: An IMAC-Enrichable Cross-Linking Reagent. ACS Cent Sci 5, 1514–1522 (2019).
7. Zhang, H. et al. Identification of protein-protein interactions and topologies in living cells with chemical cross-linking and mass spectrometry. Mol. Cell. Proteomics 8, 409–420 (2009).
8. Randall, A., Baldi, P., Rychnovsky, S. D. & Huang, L. Development of a novel cross-linking strategy for fast and accurate identification of cross-linked peptides of protein complexes. Molecular & Cellular (2011).
9. Matzinger, M. & Mechtler, K. Cleavable Cross-Linkers and Mass Spectrometry for the Ultimate Task of Profiling Protein-Protein Interaction Networks in Vivo. J. Proteome Res. 20, 78–93 (2021).
10. Sinz, A. Divide and conquer: cleavable cross-linkers to study protein conformation and protein-protein interactions. Anal. Bioanal. Chem. 409, 33–44 (2017).
11. Liu, F., Rijkers, D. T. S., Post, H. & Heck, A. J. R. Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. Nat. Methods 12, 1179–1184 (2015).
12. O'Reilly, F. J. et al. In-cell architecture of an actively transcribing-translating expressome. Science 369, 554–557 (2020).
13. Stieger, C. E., Doppler, P. & Mechtler, K. Optimized Fragmentation Improves the Identification of Peptides Cross-Linked by MS-Cleavable Reagents. J. Proteome Res. 18, 1363–1370 (2019).
14. Piersimoni, L. & Sinz, A. Cross-linking/mass spectrometry at the crossroads. Anal. Bioanal. Chem. 412, 5981–5987 (2020).
15. Gonzalez-Lozano, M. A. et al. Stitching the synapse: Cross-linking mass spectrometry into resolving synaptic protein interactions. Sci Adv 6, eaax5783 (2020).
16. Fischer, L. & Rappsilber, J. Quirks of Error Estimation in Cross-Linking/Mass Spectrometry. Anal. Chem. 89, 3829–3833 (2017).
17. Stieger, C. E., Doppler, P. & Mechtler, K. Optimized Fragmentation Improves the Identification of Peptides Cross-Linked by MS-Cleavable Reagents. J. Proteome Res. 18, 1363–1370 (2019).
18. Steigenberger, B., Albanese, P., Heck, A. J. R. & Scheltema, R. A. To Cleave or Not To Cleave in XL-MS? J. Am. Soc. Mass Spectrom. 31, 196–206 (2020).
19. Yu, C. & Huang, L. Cross-Linking Mass Spectrometry: An Emerging Technology for Interactomics and Structural Biology. Anal. Chem. 90, 144–165 (2018).
20. 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).
21. Linden, A. et al. A cross-linking mass spectrometry approach defines protein interactions in yeast mitochondria. Mol. Cell. Proteomics (2020) doi:10.1074/mcp.RA120.002028.
22. Giese, S. H., Fischer, L. & Rappsilber, J. A Study into the Collision-induced Dissociation (CID) Behavior of Cross-Linked Peptides. Mol. Cell. Proteomics 15, 1094–1104 (2016).
23. Holman, J. D., Tabb, D. L. & Mallick, P. Employing ProteoWizard to Convert Raw Mass Spectrometry Data. Curr. Protoc. Bioinformatics 46, 13.24.1–9 (2014).
24. Giese, S. H., Belsom, A., Sinn, L., Fischer, L. & Rappsilber, J. Noncovalently Associated Peptides Observed during Liquid Chromatography-Mass Spectrometry and Their Effect on Cross-Link Analyses. Anal. Chem. 91, 2678–2685 (2019).

Manuscript 3:

Reliable identification of protein-protein interactions by crosslinking mass spectrometry.

Manuscript available online
DOI: 10.1038/s41467-021-23666-z

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-23666-z OPEN

Reliable identification of protein-protein interactions by crosslinking mass spectrometry

Swantje Lenz[1,3], Ludwig R. Sinn[1,3], Francis J. O'Reilly[1,3], Lutz Fischer[1], Fritz Wegner[1] & Juri Rappsilber[1,2]✉

Protein-protein interactions govern most cellular pathways and processes, and multiple technologies have emerged to systematically map them. Assessing the error of interaction networks has been a challenge. Crosslinking mass spectrometry is currently widening its scope from structural analyses of purified multi-protein complexes towards systems-wide analyses of protein-protein interactions (PPIs). Using a carefully controlled large-scale analysis of *Escherichia coli* cell lysate, we demonstrate that false-discovery rates (FDR) for PPIs identified by crosslinking mass spectrometry can be reliably estimated. We present an interaction network comprising 590 PPIs at 1% decoy-based PPI-FDR. The structural information included in this network localises the binding site of the hitherto uncharacterised protein YacL to near the DNA exit tunnel on the RNA polymerase.

[1] Bioanalytics, Institute of Biotechnology, Technische Universität Berlin, Berlin, Germany. [2] Wellcome Centre for Cell Biology, University of Edinburgh, Edinburgh, UK. [3] These authors contributed equally: Swantje Lenz, Ludwig R. Sinn, Francis J. O'Reilly. ✉email: Juri.Rappsilber@tu-berlin.de

Crosslinking mass spectrometry (Crosslinking MS) has become a key technology for understanding the architecture of multi-protein complexes by providing distance restraints between protein residues[1]. These studies are typically performed on purified complexes, but in recent years pioneering studies have used Crosslinking MS to study the topology of PPIs in more complex systems, such as cell lysates, organelles or whole cells[2–13]. Crosslinking MS is therefore emerging as a technique for mapping PPIs alongside existing tools, such as two-hybrid screens, affinity purification, proximity labelling techniques and co-fractionation studies. Importantly, Crosslinking MS studies do not require tagging of proteins and can fixate interactions inside cells prior to cell disruption. Crosslinking MS can therefore detect otherwise difficult to observe PPIs, including weak or transient interactions and interactions involving proteins that are not easily solubilised. Unlike other large-scale PPI mapping technologies, the interactions are detected between individual residues and therefore also provide information on protein complex topology.

As for any technology for mapping PPIs, the reported interactions must be reliable to be useful. Large numbers of spurious PPIs are avoided by correctly estimating FDRs and then trimming the list of reported PPIs to the desired error rate. The standard method for error estimation in classical LC-MS-based proteomics is the target-decoy approach, where a decoy database of spurious peptide sequences is included to model random identifications. This approach assumes that the rate of matches to the decoy database is an estimator of false positives (type I error rate). This target-decoy approach has been adapted for Crosslinking MS[14–18]. Recently, however, concerns have emerged regarding current FDR methods[12,19,20] and the need for improvements is recognised widely across the Crosslinking MS field[21].

Matches in Crosslinking MS are different from those in classical proteomics because two peptides are combined to make one match. This leaves two potential opportunities for a false match, which requires additional considerations when applying the target-decoy approach, such as a crosslink-specific equation for calculating FDR[15,16]. Two additional considerations have been suggested for correctly estimating errors in crosslinking-based PPI screens. The first, whether to consider crosslinks between peptides within one protein sequence (self-links, including homomeric crosslinks) separately from crosslinks between distinct protein sequences (heteromeric crosslink)[4,15]. The second, how to handle propagation of error between the different levels of information, i.e. from crosslinked spectrum matches (CSMs), to peptide pairs, to residue pairs and finally to PPIs[16]. However, both considerations have not been systematically tested and therefore they have remained controversial with no consensus emerging for if and how they should be implemented (Supplementary Table 1).

In this work, we tested different approaches for FDR estimation and demonstrated how incorrect handling of the error estimation can have huge effects on the reliability of the reported PPIs. For this, we designed a carefully controlled large-scale crosslinking study of the model organism *E. coli* by fractionating lysate via size exclusion chromatography (SEC), crosslinking within the individual fractions, and then pooling all fractions. Proteins that did not share the same SEC fraction could not be crosslinked and therefore reveal false PPIs, without needing to rely on decoys. We used this sample to demonstrate that self-links and heteromeric crosslinks must always be separated for FDR and that data must be merged into PPIs before correct estimation of error in crosslinking-based PPI investigations.

Results

Theoretical considerations on FDR estimation in crosslinking MS. Naively, FDR is estimated based on a score distribution of CSMs to the target and decoy databases, using the decoy matches as a model of random and hence false target matches. However, the size of the search space, and therefore the chance of random matching, is inherently different for heteromeric crosslinks and self-links (Fig. 1a). In our database of 4350 proteins, the chance of matching a decoy crosslink (random) within the heteromeric crosslinks is 10.6 times higher than within the self-links (Fig. 1b). Controlling FDR in the total set of CSMs, and then selecting only heteromeric matches thus enriches for false positives. This leads to a large underestimation of the error within heteromeric CSMs, which describe PPIs (Supplementary Fig. 1). Consequently, heteromeric crosslinks must be considered separately from self-links during FDR estimation.

A second consideration is that a naïve FDR for CSMs may not reflect the error among reported PPIs. When merging data from CSMs to PPIs false and true matches may behave differently and thus the relative error will change. CSMs merge into peptide pairs, peptide pairs into residue pairs and residue pairs into PPIs (see Methods). False PPIs are the result of random CSMs and thus less likely than true PPIs to be supported by multiple CSMs. Multiple true CSMs are therefore much more likely to merge into a single PPI. This leads to a change in ratio between true and false matches as one merges CSMs into PPIs. Consequently, CSMs must be merged into PPIs before FDR estimation of PPIs (Fig. 1c).

These considerations apply universally, as they are independent of crosslinker chemistry and data analysis workflow.

Construction of a test system to investigate methods of FDR estimation in crosslinking MS. To test different approaches for FDR estimation we produced a sample for which we experimentally know a large number of the potential false PPIs. We prepared simplified cellular fractions enriched in protein complexes by separating *E. coli* lysate by size exclusion chromatography (Fig. 1d). The resulting 44 fractions span the molecular weight range from ~3 MDa to 150 kDa. A portion of each fraction was analysed by label-free quantitative proteomics to generate elution profiles of each protein across all 44 fractions. We identified 1926 *E. coli* proteins in these fractions combined. Consequently, the complexity of our sample approximates that of whole *E. coli* cells[22]. The abundance of the detected proteins spans six orders of magnitude, producing a challenging sample for detecting crosslinks.

The remainder of each fraction was split equally and crosslinked with BS3 or DSSO, respectively. For each crosslinker, the crosslinked fractions were then pooled and digested. The crosslinked peptides were first enriched by strong cation exchange chromatography (SCX) to enrich crosslinked peptides in nine high-salt fractions. Each high-salt SCX fraction was subsequently fractionated in a second chromatographic dimension by hydrophilic strong anion exchange chromatography (hSAX) into ten fractions. Following this extensive fractionation, the crosslinked peptides were acquired by LC-MS (2×90 fractions, 32.5 days of mass-spectrometric acquisition per crosslinker) to generate a substantial dataset for testing FDR methods.

Proteins eluting in the same size exclusion fraction may be crosslinked in this analysis. In contrast, proteins that were not in the same fraction cannot be crosslinked together, i.e. are 'non-crosslinkable' pairs (either because they were not identified at all or below an abundance threshold (Supplementary Fig. 1). If such a non-crosslinkable protein pair is identified during data analysis, it is a false match. This experimental assessment of PPI error is independent of the target-decoy approaches and therefore offers an opportunity to benchmark target-decoy-based PPI-FDR methods.

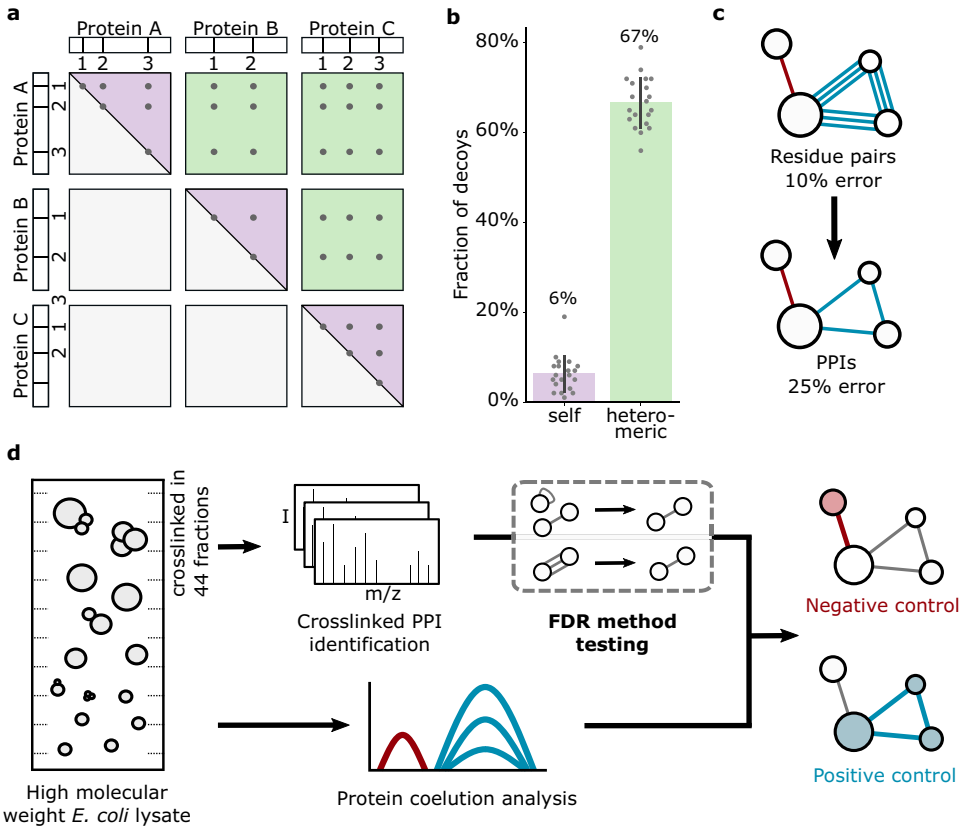


Fig. 1 Considerations for crosslinked PPI-FDR and experimental workflow. a For matches within the same protein sequence (with a non-directional crosslinker[17]), a crosslink from A1 to A2 is indistinguishable from A2 to A1 (theoretically possible search space shown as purple triangles). In contrast, heteromeric matches are not symmetrical, and therefore occupy a larger random space (green squares). **b** Fraction of decoys in random picks of 100 self and 100 heteromeric CSMs from the search output before any FDR filtering (random picks, $n = 20$, i.e. ten per crosslinker dataset). Error bars show standard deviation from the mean. Source data are provided as a Source Data file. **c** Schematic showing error increase when merging crosslinked residue pairs to PPIs. Proteins are indicated as circles; blue and red lines represent true and false linkages, respectively. **d** Experimental workflow. *E. coli* lysate was separated and crosslinked in individual high molecular weight fractions, pooled again to simulate a complex mixture and analyzed by mass spectrometry. Quantitative proteomics of uncrosslinked fractions provided protein coelution data.

Impact of CSM-FDR estimation on the reliability of identified PPIs. We first searched against a database comprising all (4350) *E. coli* proteins, including those not detected in our sample. Crosslinks of protein pairs defined as non-crosslinkable above were defined as false. At a naïve 5% decoy-based CSM-level FDR (not distinguishing self and heteromeric crosslinks), we identified 20,833 (5655 heteromeric) unique CSMs for BS3 and 22,296 (6923 heteromeric) unique CSMs for DSSO. We chose 5% FDR to have sufficient false identifications for precise FDR estimation at all information levels. In close agreement, our experimental control revealed that 4% of these CSMs are false (Fig. 2a, b). Note that CSMs in this manuscript refer to unique CSMs; using redundant CSMs will produce spurious FDR estimations (Supplementary Fig. 2).

However, naïve CSM-level FDR leads to many false PPIs, as our experimental control reveals. For this, the heteromeric CSMs of naïve 5% decoy-based CSM-FDR were merged into PPIs. Counting our non-crosslinkable PPIs then revealed that 36% of the reported PPIs were false in this DSSO dataset (Fig. 2a). In the BS3 dataset the results were very similar with naïve 5% decoy-based CSM-FDR leading to 35% false PPIs (Fig. 2b).

Given this large deviation between naïve decoy-based CSM-FDR and experimentally determined PPI error we sought further controls at the level of data analysis. As additional independent controls of decoy-based FDR we therefore used three entrapment database searches. First, we searched our spectra against *E. coli* sequences supplemented with the same number of human protein sequences. Second, we added the full *S. cerevisiae* proteome to the *E. coli* sequences and, finally, both databases were combined into an even larger entrapment database. Here we know any identified PPI that includes a human or yeast protein is false. According to these entrapment controls, at a naïve 5% decoy-based CSM-FDR, the PPI error reached an average of 45% (Supplementary Fig. 2). Although this corroborated the notion that naïve decoy-based CSM-FDR leads to a gross underestimation of the PPI error we devised two additional controls of decoy-based FDRs. For one, we performed searches using a fictional (wrong mass) crosslinker in addition to BS3 or DSSO, respectively. Any CSM involving this fictional crosslinker is a known false positive. While one of the two matched peptides in such a CSM might be correct, the other must be false to compensate for the false crosslinker mass when making up to the precursor mass. As a last control we searched previously high-confidence matched scans with shifted precursor masses to generate a set of false crosslinked peptides. One of the peptides constituting the original precursor could still be matched correctly. However, as the precursor mass was shifted, again, the second peptide cannot be matched correctly. So, any peptide pair match to these spectra constitutes a false positive. These controls reported a PPI error of 50 and 60%, respectively, at naïve 5% decoy-based CSM-FDR (Supplementary Fig. 2). In summary, not only the experimental but also the three entrapment and the two wrong mass controls revealed naïve decoy-based CSM-FDR to be inadequate for estimating PPI error.

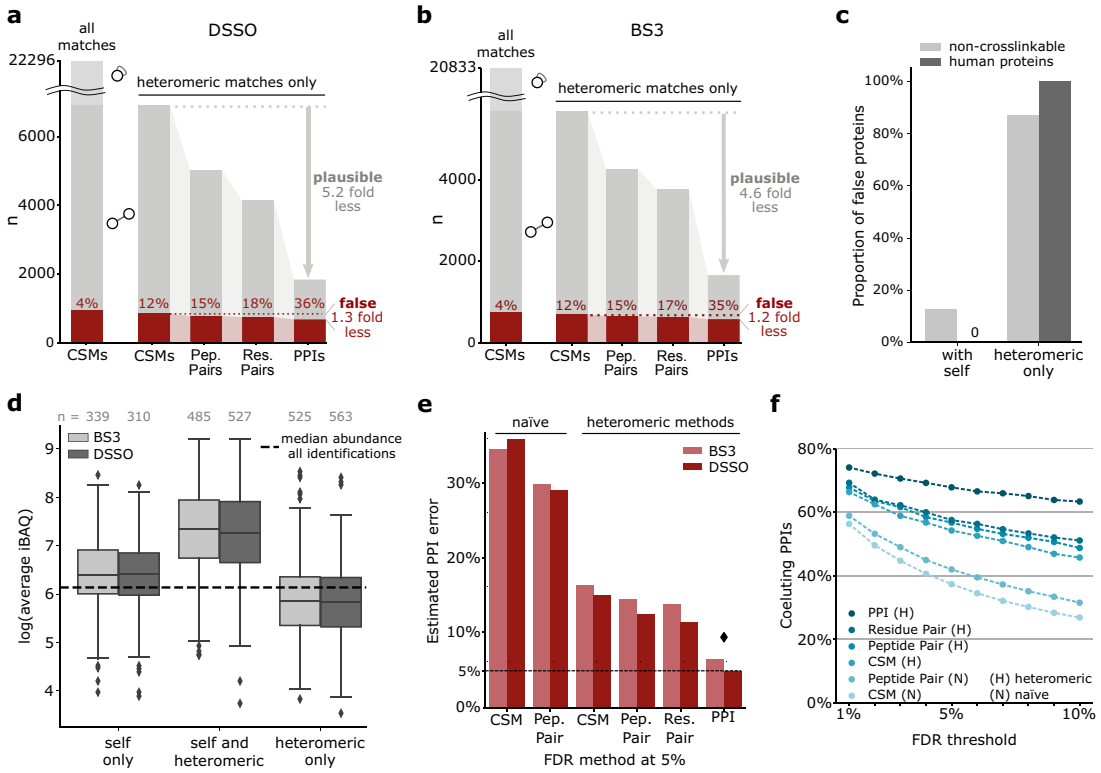


Fig. 2 Comparative analysis of different methods of FDR estimation in crosslinking MS. a, **b** False identifications as a function of merging heteromeric CSMs passing a naïve decoy-based CSM-FDR of 5% for **a** DSSO and **b** BS3, respectively. When merging crosslink data from CSMs to PPIs, the number of identifications decreases and the fraction of false identifications increases. CSMs rarely corroborated each other in false PPIs while plausible PPIs were supported by multiple CSMs. Heteromeric CSMs are indicted by circles connected with a straight line; self-CSMs by a circle with curved line. **c** Proportion of proteins involved in false PPIs with self-links or with only heteromeric crosslinks, of non-crosslinkable *E. coli* proteins and the entrapment database. **d** Proteins found exclusively in heteromeric PPIs had a lower abundance than all identified proteins and thus a low chance to be detected. Boxplots depict the median (middle line), upper and lower quartiles (boxes), 1.5 times of the interquartile range (whiskers) as well as outliers (single points). **e** PPI error resulting from a 5% FDR threshold of FDR approaches performed in other studies (Supplementary Table 1). Each bar is from a separate FDR calculation. Diamond denotes the method leading to the PPI error closest to 5%. **f** Fraction of protein pairs with similar elution profiles (correlation coefficient > 0.5) among the PPIs passing a given FDR threshold, applying different published FDR methods (Supplementary Table 1). Averages of BS3 and DSSO data are shown (Also presented sepearately in Supplementary Fig. 4). Source data for panels **a**, **b** and **d** are provided as a Source Data file.

Of note, in our experimental control 87% of false PPIs involved proteins that were seen only with heteromeric crosslinks, i.e. that lacked self-links (Fig. 2c). In the entrapment control this number increased to 100% (Fig. 2c). If observed at all, heteromeric-only proteins had a lower median abundance than all proteins in the sample suggesting that they are enriched in random matches (Fig. 2d). In contrast, the median abundance of proteins detected with both, self and heteromeric crosslinks, was 14.8-fold higher than the median of all identified proteins (significantly higher abundance than all identified proteins, $p < 0.0001$ using a one-sided Kolmogorov–Smirnov test) (Fig. 2d). The proportion of PPIs involving heteromeric-only proteins may thus be an indicator of reliability when evaluating published Crosslinking MS data.

Comparative analysis of different methods of FDR estimation in crosslinking MS. To address this inflated error of the naïve decoy-based CSM-FDR we returned to our initial theoretical considerations. Indeed, assessing heteromeric matches separately from self-matches decreased false PPIs substantially (35 to 16% and 36 to 15%, for BS3 and DSSO, respectively) (Fig. 2e). However, the error remained three times higher than the targeted 5%. We therefore also considered error propagation between information levels. As predicted, CSMs rarely corroborated each other in false PPIs while plausible PPIs were supported by multiple CSMs (1.2 versus 4.6 for BS3, 1.3 versus 5.2 for DSSO), irrespective of the crosslinker (Fig. 2a, b). This effect was most pronounced when merging unique residue pairs into PPIs. Error control at lower information levels therefore leads to large proportions of reported PPIs being false (Fig. 2e). This also holds true for all other reporting levels (i.e. CSMs, peptide pairs and residue pairs) (Supplementary Fig. 3).

In contrast, first merging CSMs for each PPI and then assessing the FDR gave more reliable results: 6.6% and 4.9% false PPIs when applying 5% decoy-based PPI-FDR (Fig. 2e) for BS3 and DSSO, respectively. This is also supported by the other controls, which indicated an actual error close to 5% (4.8% for BS3 and 4.9% for DSSO) when applying 5% decoy-based PPI-FDR (Supplementary Fig. 2).

As a positive control, we evaluated the proportion of PPIs that were supported by correlation of protein coelution profiles (Fig. 2f, Supplementary Fig. 4). The fraction of supported PPIs was highest when using heteromeric PPI-FDR and the proportion decreased when raising the FDR threshold, as expected. The same trends are true for the alternative positive control of using interaction evidence from the STRING database (Supplementary Fig. 4).

High-quality PPIs in *E. coli* lysate. An FDR threshold should be chosen to meet the stringency required by the study (Supplementary Fig. 5). At a heteromeric decoy-based PPI-FDR of 5%

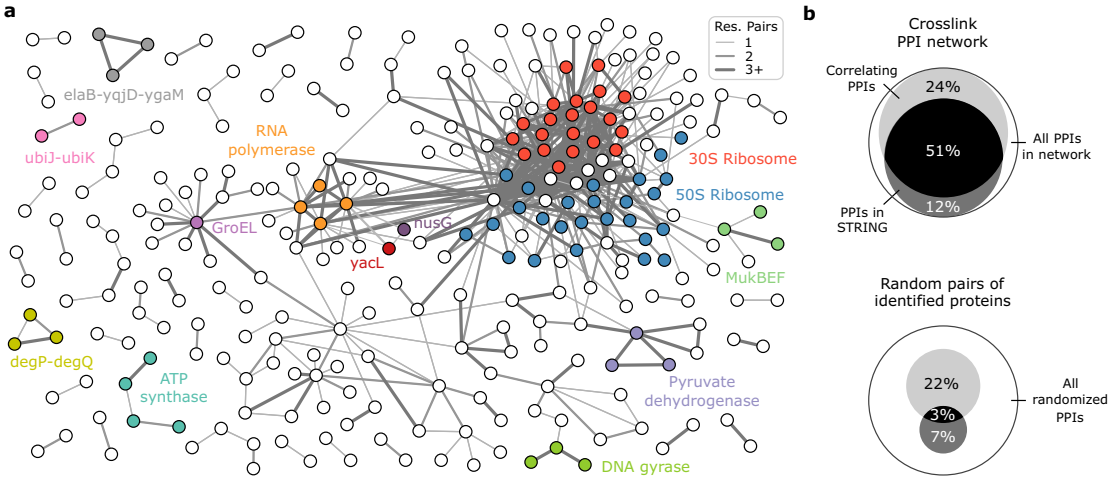


Fig. 3 Heteromeric PPI-FDR leads to high fidelity PPI network in *E. coli* lysate. a Crosslinking MS-derived PPI network of soluble high molecular weight *E. coli* proteome. Selected proteins (circles) and protein complexes are highlighted. The proteins AceA and TnaA were removed for clarity. **b** Characterization of the obtained PPI network in comparison to random PPIs from proteins identified in coelution data. Shown are the overlaps with STRING database and coelution data (correlation coefficient > 0.5).

applied on our data, 756 PPIs are reported, with 38 expected to be false. To focus on a high-quality subset of PPIs in the *E. coli* lysate, we applied a 1% heteromeric PPI-FDR cut-off, yielding 590 PPIs involving 308 proteins (Fig. 3a, Supplementary Data 1), connected with a total of 2539 residues pairs.

Three hundred sixty-six (62%) of these PPIs are connected by more than one residue pair (Supplementary Fig. 5). Eleven percent of the proteins found in PPIs had no self-links, but most had abundances higher than the sample median. These proteins tend to be small and thus produce few peptides, so can be difficult to observe by mass spectrometry (Supplementary Fig. 5). We found 63% (370) of the detected PPIs in the STRING database (Fig. 3b). Ninety eight percent (576) were found to be eluting in a fraction together and 68% had similar elution profiles (correlation coefficient > 0.5), suggesting that they form stable complexes (Supplementary Data 2). Ribosomal proteins had a complex elution pattern, presumably due to the presence of assembly intermediates, although many of the proteins that were found crosslinked to the ribosome are known interactors (26 of 53).

The crosslink-based PPI network included 289 protein pairs with highly similar coelution (correlation coefficient > 0.8). The majority of these were known interactions including complexes like ATP synthase, pyruvate dehydrogenase, MukBEF or DNA gyrase. The data confirmed binding of acyl carrier protein to MukBEF, and of YacG to DNA gyrase (Supplementary Fig. 6 and 7). In addition, 130 PPIs with highly similar coelution were not yet experimentally confirmed for *E. coli* K12, though 55 of these had a STRING entry based on other evidence. Novel interactions included those between the small ribosomal regulators ElaB, YgaM and YqjD, the periplasmic endoproteases DegP and DegQ, the ubiquinone biosynthesis accessory factors UbiK and UbiJ, as well as GroEL and potential substrates (Supplementary Fig. 8).

RNA polymerase (RNAP) crosslinked to 23 proteins (Fig. 4a). Previous interaction evidence was available for 20 of these, including the transcription factors RpoD and GreB, and the transcriptional regulators NusG, NusA and RapA; all crosslinks are in agreement with previously suggested binding sites (Supplementary Fig. 9). YacL, a protein of unknown function that was found to be associated with RNAP in pull-down experiments[23], crosslinked to the beta and beta' subunit of RNAP (four residue pairs), as well as to NusG (two residue pairs). It also coeluted with RNAP (correlation of 0.988) with an abundance comparable to NusG (Fig. 4b).

To confirm the interaction, we performed pulldowns using K12 strains with endogenously tagged ORFs of YacL, NusG and RpoB to carry an affinity-tag. YacL affinity-enriched RNAP and NusG (Fig. 4c) and, conversely, NusG and RpoB enriched YacL, thus confirming the association of YacL with NusG and RNAP (Supplementary Fig. 10). To further constrain the binding site of YacL on the RNAP and confirm the interaction site by use of a different crosslinker we crosslinked these affinity-enriched complexes using the photoactivatable crosslinker sulfo-SDA, which can provide a higher density of crosslinks than DSSO or BS3[24]. Sulfo-SDA crosslinking either of the three affinity-enriched proteins confirmed the direct binding of YacL to the RNAP and NusG with a combined 14 unique residue pairs (Supplementary Fig. 10 and Supplementary Data 3-5). The total of 20 residue pairs from DSSO, BS3 and sulfo-SDA thus constrain the binding site of YacL to RNAP and NusG. Sixteen of these residue pairs were between our I-TASSER model of YacL and regions of RNAP-NusG included in the solved structure (PDB 6C6U [https://doi.org/10.2210/pdb6c6u/pdb]). These were used in DisVis to calculate the accessible interaction space and localize YacL on RNAP next to NusG at the DNA exit site (Fig. 4d).

Discussion

There have been several recent advances in enrichment and detection of crosslinked peptides by Crosslinking MS that suggest it will soon be able to map large portions of the cellular interactome in a single experiment[3,13,25]. This will open the door to detecting changes in these interactomes in different cellular states by quantification of the abundances of the detected crosslinks[2,26]. All of these advances require correctly controlled FDR to produce results that can be relied upon.

In previous studies, the quality of identified crosslinks was assessed by measuring inter-residue distances in known protein structures. However, for proteome-wide crosslinking studies, this approach is inherently biased towards true interactions as they are likely to be enriched in known complexes[27]. The majority of random PPIs are neglected by this FDR evaluation method, making this approach completely inadequate for reliable PPI error estimation.

In this work, we experimentally demonstrated that Crosslinking MS can reliably identify PPIs using the target-decoy approach as a quantitative error metric. Decoys are only a model of false positives with a number of underlying

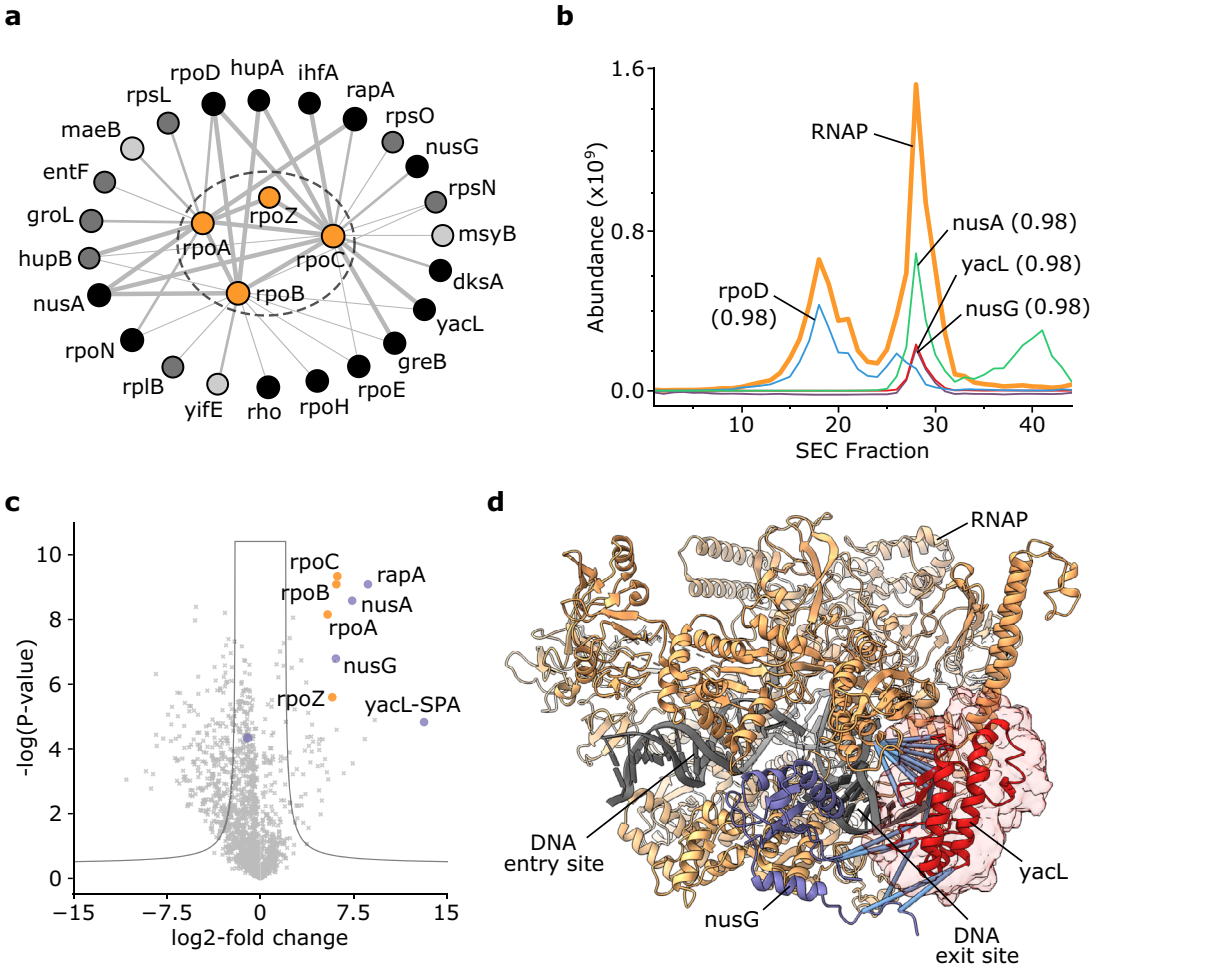


Fig. 4 The uncharacterised protein YacL binds to RNA polymerase. a PPI subnetwork of RNAP. Line thickness between proteins (circles) increases with frequency of observed crosslinks (i.e. one, two, three and more crosslinked residue pairs). Colour scheme for RNAP binders: light grey = coelution correlation > 0.5, dark grey = STRING database combined score ≥ 150, black = both of the previous categories. **b** Elution traces of RNAP (average abundance of its subunits) and selected RNAP binders with their minimal elution correlation coefficient to any RNAP constituent. **c** Volcano plot showing the affinity enrichment of SPA-tagged YacL, which co-enriches RNAP (orange) and a number of proteins found crosslinked to the RNAP in the YacL affinity-enrichment experiment (violet). **d** RNAP with bound NusG (PDB 6C6U [10.2210/pdb6c6u/pdb]) and the region of Crosslinking MS-defined accessible interaction space with 14 satisfied restraints for YacL (I-TASSER model, placed for visualisation purposes) highlighted. Crosslinks between YacL and NusG or RNAP are highlighted in blue.

assumptions[16] and they cannot model false positives that do not arise from spectral matching, such as peptides non-covalently associating during LC-MS[28]. Considering these caveats, it is reassuring that our four different controls closely agree with the outcome of the target-decoy approach. This negates the need for any additional heuristics suggested by others[27]. We showed that the target-decoy approach requires separating self and heteromeric crosslinks and that error should be estimated for the information level that is being reported. For example, when reporting residue pairs for structural analyses of individual protein complexes, residue pair-level FDR should be applied. However, when reporting PPIs, CSMs need to be merged to PPI level prior to FDR estimation. Other ways of merging CSM scores into PPI scores from the one we use here are possible. However, for accurate PPI error estimation, these methods would need to adhere to the two fundamental considerations. These concepts were implemented in our open-source FDR estimation software tool, xiFDR v2.0, which is crosslink search software independent. The large dataset presented here, with its internal controls, will allow testing of other aspects of the Crosslinking MS workflow in the future.

Correctly controlled error is an important element of any discovery-based technology. This remains a challenge even in well-established PPI mapping technologies including two-hybrid and affinity purification studies. Crosslinking MS for mapping PPIs now has a reliable FDR estimation procedure. This is an essential prerequisite for this technology to bridge the gap between structural studies and systems biology by reliably revealing topologies of PPIs in their native environments.

Methods

Materials. Unless otherwise stated, reagents were purchased in the highest quality available from Sigma (now Merck, Darmstadt, Germany). Empore 3M C18-Material for LC-MS sample cleanup was from Sigma (St. Louis, MO, USA), glycerol from Carl Roth (Karlsruhe, Germany). The BS3 (bis (sulfosuccinimidyl) suberate) and sulfo-SDA (sulfosuccinimidyl 4,4'-azipentanoate) crosslinkers were supplied by Pierce Biotechnology (Thermo Fisher Scientific, Waltham, MA, USA) and the DSSO (disuccinimidyl sulfoxide) crosslinker from Cayman Chemical (Ann Arbor, MI, USA).

Biomass production. A single clone of *Escherichia coli* K12 strain (BW25113 purchased from DSMZ, Germany; https://www.dsmz.de/) grown on Agar plates was selected for inoculation of lysogeny broth (LB)-media. A preculture aliquot was used to start fermentation in a Biostat A plus bioreactor (Sartorius, Göttingen,

Germany) in LB medium with 0.5% (w/v) glucose and at 37 °C. The pH and dissolved oxygen were monitored and adjusted by the addition of sodium hydroxide/phosphoric acid or stir speed control, respectively. Overall growth was monitored by optical density measurements at 600 nm. When the culture reached an optical density of 10, the fermentation was stopped and the culture rapidly cooled in stirred ice water followed by harvesting the biomass by centrifugation at 5000 × *g*, 4 °C for 15 min. Cell pellets were stored at −80 °C after washing with PBS and snap-freezing in liquid nitrogen.

For pull-down experiments, *Escherichia coli* K12 strains with endogenously C-terminal SPA-tagged rpoB, nusG and yacL (purchased from Horizon, Cambridge, United Kingdom, https://horizondiscovery.com/) were plated according to distributor's instructions. A single clone of each strain was selected for genetic validation and subsequent starter cultures. Gene sequences were validated by PCR using primers hybridizing upstream of each open reading frame of interest and within the SPA-tag sequence (Supplementary Data 6). With the exception of the yacL-ORF which had a non-silent point mutation (Q118L), all protein- and tag-coding sequences were correct. Production cultures were inoculated into terrific broth medium and cultivated at 32 °C in baffled flasks until late log-phase. Biomass was harvested and stored after snap-freezing as described above.

Cell lysis and high-molecular-weight proteome fractionation by size exclusion chromatography (SEC). Cell pellets were suspended at 0.2 g wet-mass per ml in ice-cold lysis buffer (50 mM Hepes pH 7.2 at RT, 50 mM KCl, 10 mM NaCl, 1.5 mM $MgCl_2$, 5% (v/v) glycerol, 1 mM dithiothreitol (DTT), spatula tip of chicken egg white lysozyme (Sigma, St. Louis, MO, USA)). Cells were lysed by sonication on ice. Prior to sonication cOmplete EDTA-free protease-inhibitor cocktail (Roche, Basel, Switzerland) was added according to the manufacturer's instructions. After sonication, 125 units of Benzonase (Merck, Darmstadt, Germany) were added. Subsequently, the lysate was cleared of cellular debris by centrifugation for 15 min at 4 °C and 15,000 × *g*. DTT was added again to 2 mM. This cleared lysate was subjected to ultracentrifugation using a 70 Ti fixed-angle rotor for 1 h at 106,000 × *g* and 4 °C. Then, the supernatant was concentrated using ultrafiltration with Amicon spin filters (15 kDa molecular weight cut-off) to reach a total protein concentration of 10 mg/ml, as judged by microBCA assay (Thermo Fisher Scientific, Waltham, MA, USA). Aggregates were removed by centrifugation for 5 min at 16,900 × *g* and 4 °C. Two milligrams of soluble high molecular weight proteome was loaded onto a BioSep SEC-S4000 column (600 × 7.8 mm, pore size 500 Å, particle size 5 µm, Phenomenex, CA, USA) and fractionated at 200 µl/min flow rate and 4 °C while collecting fractions of 200 µl over the separation range from ~3 MDa to 150 kDa (as judged by Gel filtration calibration kit (HMW), GE Healthcare).

Affinity-pulldowns of RNA polymerase constituents and binders. Cells were lysed by sonication identically to the protocol described above. The supernatants from centrifugation for 1 h at 4 °C and 20,000 × *g* were incubated with washed Anti-FLAG M2 agarose beads (Sigma, St. Louis, MO, USA) on a vertical rotator for 2 h at 4 °C, according to the specifications of the manufacturer. Supernatants after incubation were discarded and beads washed twice with wash buffer (10 mM Tris*HCl pH 7.4 at RT, 100 mM NaCl, 10% (v/v) glycerol) and once with modified lysis buffer (50 mM Hepes pH 7.2 at RT, 50 mM KCl, 10 mM NaCl, 1.5 mM $MgCl_2$, 5% (v/v) glycerol). M2 beads from replica pulldowns were pooled in a single tube and again resuspended in modified lysis buffer. TEV protease (Sigma, St. Louis, MO, USA) was added >0.5 U/µl M2 beads and the protein complexes of interest eluted over 1 h at 16 °C with gentle agitation. Aliquots of cleared supernatants and eluates from TEV cleavage were collected and processed as described below.

Sample preparation for LC-MS protein identification with non-crosslinked samples. For protein identification from SEC fractionation, aliquots (40 µl) of each fraction were precipitated by adding four volumes of cold acetone followed by an incubation at −20 °C overnight. Pellets were collected by centrifugation and supernatants discarded. Protein pellets were air-dried and subsequently solubilized using 6 M urea, 2 M thiourea, 100 mM ABC (ammonium bicarbonate). Derivatization was accomplished by incubating for 30 min at RT with 10 mM DTT followed by 20 mM IAA (iodoacetamide) for 30 min in the dark at RT, respectively. Proteases were added to the samples: LysC (1:100 (m/m)) for 4.5 h at 37 °C, followed by diluting 1:5 with 100 mM ABC and continued with trypsin (1:25 (m/m)) at 37 °C for 16 h. The reactions were stopped by adding TFA (trifluoroacetic acid) to a pH of 2–3. Subsequently, sample cleanup following the Stage-tip protocol was performed and samples were stored at −20 °C until LC-MS acquisition. Samples from pulldowns were processed similarly with the following changes: use of 8 M urea, 100 mM ammonium bicarbonate for solubilization; blocking of reduced cysteines with 30 mM IAA; digestion with trypsin (ca. 1:50 (m/m)).

LC-MS protein identification with non-crosslinked samples. Protein identifications in SEC fractions and from pull-down experiments via LC-MS were conducted using a Q Exactive HF mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) coupled to an Ultimate 3000 RSLC nano system (Dionex, Thermo Fisher Scientific, Sunnyvale, USA), operated under Tune 2.9, SII for Xcalibur 1.4 and Xcalibur 4.1. 0.1% (v/v) formic acid and 80% (v/v) acetonitrile, 0.1% (v/v) formic acid served as mobile phases A and B, respectively. Samples were loaded in 1.6% acetonitrile, 0.1% formic acid on an Easy-Spray column (C18, 50 cm, 75 µm ID, 2 µm particle size, 100 Å pore size) operated at 45 °C and running with 300 nl/min flow. Peptides were eluted with the following gradient: 2 to 6% buffer B in 1 min, 6 to 10% B in 2 min, 10 to 30%B in 37 min, 30 to 35% in 5 min followed by 35 to 45%B in 2 min. Then, the column was set to washing conditions within 1.5 min to 90% buffer B and flushed for another 5 min. For the mass spectrometer the following settings were used: MS1 scans resolution 120,000, AGC (automatic gain control) target 3×10^6, maximum injection time 50 ms, scan range from 350 to 1600 *m/z*. The ten most abundant precursor ions with z = 2–6, passing the peptide match filter ("preferred") were selected for HCD (higher-energy collisional dissociation) fragmentation employing stepped normalized collision energies (29 ± 2). The quadrupole isolation window was set to 1.6 *m/z*. Minimum AGC target was 2.5×10^4, maximum injection time was 80 ms. Fragment ion scans were recorded with a resolution of 15,000, AGC target set to 1×10^5, scanning with a fixed first mass of 100 *m/z*. Dynamic exclusion was enabled for 30 s after a single count and included isotopes. Each LC-MS acquisition took 75 min.

Quantitative proteomics database search. Raw data from bottoms-up proteomics experiments were processed using MaxQuant[29] version 1.6.0.16 operated under default settings (fully tryptic digestion with two missed cleavages maximum; up to five variable modifications per peptide (oxidised methionine and acetylated protein N-termini), MS1 match tolerance 20 ppm (first search)/4.5 ppm (main search), MS/MS match tolerance 20 ppm); carbamidomethylation of cysteine set as fixed modification; 1% PSM and protein group FDR). Each SEC fraction or pull-down replica injection was treated as an individual experiment. Quantitation by iBAQ[30] requiring a minimum of two peptides (unique + razor) and matching between runs were enabled. For data from pull-down experiments, label-free quantitation was enabled with default settings (LFQ minimum ratio count of 2, Fast LFQ enabled, minimum number/average of neighbour 2/6, stabilize large LFQ ratios and requirement for MS2 for LFQ comparisons enabled). Supernatant samples from cell lysis were included to increase absolute protein identifications via the matching between runs feature. The database used was the Uniprot curated reference proteome UP000000625 with two unreviewed entries removed summing to a total 4350 proteins (retrieved on 04/08/2019).

Protein enrichment from pull-down experiments was assessed using Perseus[31] version 1.5.6.0. Proteins identified by site only, reverse hits and contaminants were filtered out. LFQ protein quantitation data was log2-transformed and filtered to contain three valid values in at least one experiment (e.g. in any TEV eluate). Missing values were imputed on the total matrix with default settings (width: 0.3, downshift 1.8). Volcano plots comparing TEV eluates of targeted affinity enrichment with K12 wildtype mock enrichment were created using a two-sided, two-sample *t*-test with 1% FDR and an artificial variance S0 of 2. For high-resolution figures, the matrix and the cut-off curve were exported to reproduce the plots in python 3.7 with pandas 0.24.2 using the seaborn 0.9.0 package.

Protein crosslinking, digestion and sample cleanup of SEC fractions. The remaining parts of the SEC fractions (160 µl, see above) were split into two 75 µl aliquots, for the two crosslinking reactions, and adjusted to 97.5 µl with 1× SEC-Running buffer. Crosslinker stock solutions were prepared freshly at 30 mM in water free DMF. Crosslinking of the fractions was initiated by quickly mixing each sample with 2.5 µl crosslinker stock to a final concentration of 0.75 mM crosslinker. The crosslinking reaction was incubated for 2 h on ice before quenching with ABC at 50 mM and further incubation for 30 min on ice. The crosslinked samples were acetone-precipitated at −20 °C overnight (see above). Protein was solubilized in 6 M Urea, 2 M Thiourea, 100 mM ABC. For sample reduction and alkylation, 10 mM DTT for 30 min at RT and 20 mM IAA for 30 min at RT in the dark were employed. For sample proteolysis, LysC was added at 1:100 (m/m) ratio and incubated for 4 h at 37 °C. Upon 1:5 dilution with 100 mM ABC, Trypsin was added to the sample (1:25 (m/m)) and digestion continued for 16 h at 37 °C followed by stopping via addition of TFA to a pH of 2–3. The digests were desalted using SPE cartridges following the manufacturer's instruction and eluates dried, aliquoted and stored at −20 °C until further use.

Multidimensional offline fractionation of crosslinked peptide samples. All crosslinked peptide pools were fractionated using an Äkta pure system (GE Healthcare, Chicago, IL, USA) employing a PolySulfoethyl A SCX column (100 × 2.1 mm, 300 Å, 3 µm) equipped with a guard column of identical stationary phase (10 × 2.0 mm) (PolyLC, Columbia, MD, USA) running at 0.2 ml/min for the first separation dimension. Here, mobile phase A consisted of 10 mM KH_2PO_4 pH 3.0, 30% ACN while mobile phase B contained 1 M KCl in addition. The system was kept at 21 °C throughout the fractionation. Dried digestion aliquots of 400 ug peptides were dissolved in mobile phase A. Upon injection, peptides were eluted isocratically for 2 min followed by an exponential gradient up to 700 mM KCl with following steps: 12 min to 12.7%, followed by 1-min steps to 14.5, 16.3, 18.8, 23.0, 30.0, 40.0, 70.0% B. Fractions of 200 µl size were collected over the elution range. The same nine high-salt fractions from five replica SCX runs were pooled for desalting using Stage-tips. Dried Stage-tip eluates of each individual SCX fraction were then subjected to the second dimension offline fractionation by hSAX

chromatography. Here, a Dionex IonPac AS-24 hSAX column (250 × 2.0 mm) with an AG-24 guard column (Thermo Fisher Scientific, Dreieich, Germany) were used on Äkta pure system (see above). Mobile phase A consisted of 20 mM Tris*HCl pH 8.0 with mobile phase B containing 1 M NaCl in addition. The system was kept at 15 °C for these experiments. Samples were eluted from Stage-tips, dried and resuspended in mobile phase A. Again, peptides were loaded under isocratic conditions for 3 min, and then eluted by an exponential gradient with the following steps 1.8, 3.5, 5.3, 7.1, 9.1, 11.2, 13.5, 16.3, 19.7, 24.1, 30.2, 38.8, 51.5, 70.6, 100% B lasting for one minute each. Fractions of 150 µl size were collected throughout the elution phase. Adjacent fractions were pooled to give ten pools in total (fractions 3–6/7–14/15–17/18–19/20–21/22–23/24–25/26–27/29–29/30–35), that were desalted using Stage-tips.

Protein crosslinking, digestion and crosslink enrichment of pull-down eluates. The remaining pull-down eluate fractions (minus aliquots for protein identification, see above) were split into five fractions. The heterobifunctional photoactivatable crosslinker sulfo-SDA was dissolved in modified lysis buffer (50 mM Hepes pH 7.2 at RT, 50 mM KCl, 10 mM NaCl, 1.5 mM $MgCl_2$, 5% (v/v) glycerol) and immediately added to the samples at 50, 100, 250, 500 and 1000 µM. The crosslinking reaction proceeded in the dark for 2 h on ice. UV-crosslinking was achieved by irradiation with a high-power UV-A LED laser (LuxiGenTM) at 365 nm for 15 s at one Ampere[32]. Samples were frozen and stored at −20 °C. Next, the samples were denatured by adding solid urea to give an 8 M solution, reduced using DTT at 10 mM following incubation at RT for 30 min and derivatized at 30 mM IAA over 20 min at RT and in the dark. LysC protease was added (protease:protein ratio ca. 1:100 (m/m)) and the samples digested for 4 h at 37 °C. Then, the samples were diluted 1:5 with 100 mM ABC and trypsin was added at a ratio of ~1:50 (m/m). Digestion progressed for 16 h at 37 °C until stopping with TFA. Digests were cleaned up using C18 StageTips.

Eluted peptides were fractionated using a Superdex Peptide 3.2/300 column (GE Healthcare, Chicago, IL, USA) at a flow rate of 10 µl min−1 using 30% (v/v) acetonitrile and 0.1 % (v/v) trifluoroacetic acid as mobile phase[33]. Early 50-µl fractions were collected, dried and stored at −20 °C prior to LC-MS analysis.

LC-MS for crosslink identification. LC-MS analysis of crosslinked peptides derived from the SEC-separated *E. coli* proteome and multidimensional fractionation was performed using a Q Exactive HF mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) coupled to an Ultimate 3000 RSLC nano system (Dionex, Thermo Fisher Scientific, Sunnyvale, USA), operated under Tune 2.11, SII for Xcalibur 1.5 and Xcalibur 4.2. Mobile phases A and B consisted of 0.1% (v/v) formic acid and 80% (v/v) acetonitrile, 0.1% (v/v) formic acid, respectively. Samples were loaded in 1.6% acetonitrile, 0.1% formic acid on an Easy-Spray column (C18, 50 cm, 75 µm ID, 2 µm particle size, 100 Å pore size) running at 300 nl/min flow and kept at 45 °C. Analytes were eluted with the following gradient: 2 to 7.5% buffer B in 5 min, followed by a linear 80-min gradient of 7.5 to 42.5% and an increase to 50% B over 2.5 min. Then, the column was set to washing conditions within 2.5 min to 95% buffer B and flushed for another 5 min. The mass-spectrometric settings for MS1 scans used were: resolution set to 120,000, AGC of 3×10^6, maximum injection time of 50 ms, scanning from 400–1450 *m/z* in profile mode. The ten most intense precursor ions that passed the peptide match filter ("preferred") and with z = 3–6 were isolated using a 1.4 *m/z* window and fragmented by HCD using in-house optimized stepped normalized collision energies (BS3: 30 ± 6; DSSO: 24 ± 6). Fragment ion scans were acquired at a resolution of 60,000, AGC of 5×10^4, maximum injection time of 120 ms scanning from 200–2000 *m/z*, underfill ratio set to 1%. Dynamic exclusion was enabled for 30 s (including isotopes). In-source-CID was enabled at 15 eV to minimize gas-phase associated peptides[28]. Each LC-MS run took 120 min.

For LC-MS/MS analysis of sulfo-SDA crosslinked samples, we used an Orbitrap Fusion Lumos Tribrid mass spectrometer (Thermo Fisher Scientific, Germany) connected to an Ultimate 3000 RSLCnano system (Dionex, Thermo Fisher Scientific, Germany), which were operated under Tune 3.4, SII for Xcalibur 1.6 and Xcalibur 4.4. Fractions from SEC were resuspended in 1.6% acetonitrile 0.1% formic acid and loaded onto an EASY-Spray column of 50 cm length (Thermo Scientific) running at 300 nl/min. Gradient elution using water with 0.1% formic acid and 80% acetonitrile with 0.1% formic acid was accomplished using optimised gradients for each SEC fraction (from 2–18% mobile phase B to 37.5-46.5% over 90 min, followed by a linear increase to 45–55 and 95% over 2.5 min each). Each fraction was analysed in duplicate. The settings of the mass spectrometer were as follows: Data-dependent mode with 2.5s-Top-speed setting; MS1 scan in the Orbitrap at 120,000 resolution over 400 to 1500 *m/z* with 250% normalized AGC target; MS2 scan trigger only on precursors with z = 3–7+, AGC target set to "standard", maximum injection time set to "dynamic"; fragmentation by HCD employing a decision tree logic with optimised collision energies[34,35]; MS2 scan in the Orbitrap at resolution of 60,000; dynamic exclusion was enabled upon a single observation for 60 s.

Crosslink database search for BS3 and DSSO. Raw data from mass spectrometry were processed using msConvert (version 3.0.11729)[36] including denoising (top 20 peaks in 100 *m/z* bins) and conversion to mgf-file format. Precursor masses were re-calibrated to account for mass shifts during measurement. Obtained peak files were analysed using xiSEARCH 1.6.746[5] with the following settings: MS1/MS2 error tolerances 3 and 5 ppm, allowing up to two missing isotope peaks[37], tryptic digestion specificity with up to two missed cleavages, carbamidomethylation on cysteine as fixed and oxidation on methionine as variable modification, losses: $-CH_3SOH/-H_2O/-NH_3$, crosslinker BS3 (138.06807 Da linkage mass) or DSSO (158.0037648 Da linkage mass) with variable crosslinker modifications on linear peptides ("BS3-NH2" 155.09463 Da, "BS3-OH" 156.07864 Da, "DSSO-NH2" 175.03031 Da, "DSSO-OH" 176.01433 Da). xiSEARCH algorithms are identical for both crosslinkers (BS3 and DSSO). For samples crosslinked with DSSO, additional loss masses for crosslinker-containing ions were defined accounting for its cleavability ("A" 54.01056 Da, "S" 103.99320 Da, "T" 85.98264). Matches were not filtered for having DSSO-specific signature peaks. Crosslink sites for both reagents were allowed for side chains of Lys, Tyr, Ser, Thr and the protein N-terminus. Note that we included a non-covalent crosslinker with a mass of zero to flag spectra potentially arising from gas-phase associated peptides. These spectra were removed prior to false-discovery-rate (FDR) estimation[28].

As for the non-crosslinked samples, the full *E. coli* proteome of 4350 proteins was used. For the entrapment database control the database was extended by three different entrapment databases (see below). For the final PPI network, the search database was reduced to only proteins identified in our 44 SEC fractions, to reduce noise in the database. Decoys were generated for all searches, including the entrapment database. For this, protein sequences were reversed and for each decoy protein the enzyme specific amino acids were swapped with their preceding amino acid[29].

FDR calculation for BS3 and DSSO datasets. Results were filtered prior to FDR to crosslinked peptide matches having a minimum of three matched fragments per peptide, a delta score of 15% of the match score and a peptide length of at least six amino acids. Additionally, identifications ambiguously matching to two proteins or more were removed. FDR was calculated based on decoy matches by xiFDR (version 2.0dev) using Eq. (1):[16]

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

Depending on the experiment, FDR was employed on different result levels (CSM, peptide pair, residue pair or protein pair) with defined thresholds. Scores of higher levels were calculated as described by Fischer and Rappsilber[16] using Eq. (2):

$$\mathrm{Score}_{\mathrm{higher\ level}} = \sqrt{\Sigma\ (\mathrm{Score}_{\mathrm{lower\ level}})^2} \quad (2)$$

FDR was solely calculated based on that score, no further improvement by other information was done at this point. To account for the improvement of the identification by prefiltering on lower levels[16], the same threshold was employed on each of the lower levels. Self- and heteromeric crosslinks were handled together or separately by enabling/disabling the grouping option.

For the final PPI network, BS3 and DSSO PPIs were separately filtered to 1% heteromeric PPI-FDR. As the score cut-offs differed between the two datasets, the scores from each dataset were first normalized (i.e. the local FDR was used as a normalized score) to range between 0 and 1. Subsequently, the two tables were concatenated and the FDR calculated again as described above and filtered to 1% FDR.

Non-crosslinkable control. Due to the high sensitivity of mass spectrometry we identified a long list of proteins in each SEC fraction. Theoretically, all of the proteins in a fraction could be crosslinked. In practice, however, even if this were the case, we could not detect all these crosslinks because many would be below our detection limit. We therefore set out to heuristically determine the detection limit in our analysis. For this, the iBAQ values were determined across all SEC fractions of all proteins that were part of an identified heteromeric peptide pair, at a generous 10% heteromeric peptide-pair FDR. For each identified pair of proteins, the respective iBAQ pairs were determined across all SEC fractions (note that iBAQ could be zero if a protein was not identified in a given fraction). Looking into each fraction the lower of the two iBAQ values was kept. The maximum of this distribution over all fractions was then taken, called "best lower iBAQ" of a protein pair. We assumed this to be the appropriate abundance estimate for a protein pair and therefore the best estimate of the chance for this pair to be observed in our experiment as crosslinked. The question now is what abundance is sufficient. As a heuristic, we removed the lower 5% iBAQ values (iBAQ of 4.3E6). This removed very few (169, 7%) of our identified protein pairs (n.b. at a very loose FDR threshold), i.e. did not change much the outcome of our identification data by generating false negatives.

For any identified protein pair to be considered plausible, both proteins had to be found in at least one SEC fraction together with individual iBAQ values above our iBAQ threshold (iBAQ of 4.3E6). Otherwise, they were defined as non-crosslinkable. 544,274 (6% of all theoretically possible PPIs in the *E. coli* proteome of 4350 proteins) PPIs are defined as plausible (Supplementary Data 7), while 8,914,801 (94%) PPIs are non-crosslinkable. Note that unlike an error control using an entrapment database during search, only proteins that could make up the sample were considered.

Additionally, the difference in the sizes of the false and plausible search spaces needs to be taken into account for error estimation, i.e. the different number of possible tryptic peptides. While all matches in the false search space are false by

definition, some matches in the plausible search space will also be random. To account for these, the Lysine/Arginine content of proteins in the respective groups was used as an estimate for the number of possible peptides and the observed error is calculated with Eq. (3):

$$\mathrm{Error}_{\mathrm{PPI,non-crosslinkable}} = \frac{n_{\mathrm{false}}}{n_{\mathrm{plausible}} + n_{\mathrm{false}}} \cdot \frac{\mathrm{KR}_{\mathrm{plausible}} + \mathrm{KR}_{\mathrm{false}}}{\mathrm{KR}_{\mathrm{false}}} \quad (3)$$

$$= \frac{n_{\mathrm{false}}}{n_{\mathrm{plausible}} + n_{\mathrm{false}}} \cdot 1.09$$

where n_{false} is the number of PPIs defined as "non-crosslinkable", $n_{\mathrm{plausible}}$ the number of PPIs that are plausible, both after passing the respective FDR calculation. $\mathrm{KR}_{\mathrm{plausible}}$ and $\mathrm{KR}_{\mathrm{false}}$ are the sums of Lysines and Arginines in the proteins of the plausible or false interactions, respectively. Therefore, the Lysine and Arginine normalisation factor, here 1.09, is database specific.

Entrapment database control calculation. As a second control, the error of matched PPIs was estimated based on known wrong matches to three entrapment databases of different sizes. For one, the same number (4350) of human proteins of similar size was added by sampling a human protein similar in Lysine and Arginine content for each *E. coli* protein. As a second entrapment database, the full *S. cerevisiae* proteome was added. Finally, both databases were combined for a third entrapment database.

PPIs were defined as false if one or more proteins in the PPI was a human or yeast protein. Additionally, the difference in entrapment and possible search space has to be taken into account, similar to the approach for the non-crosslinkable control, following Eq. (4):

$$\mathrm{Error}_{\mathrm{PPI,entrapment}} = \frac{n_{\mathrm{entrapment}}}{n_{\mathrm{E.coli}} + n_{\mathrm{entrapment}}} \cdot \frac{KR_{\mathrm{E.coli}} + KR_{\mathrm{entrapment}}}{KR_{\mathrm{entrapment}}} \quad (4)$$

As expected from doubling the original database size by adding an entrapment database of equal size, the search space normalisation approximates to 2 (1.998) for the human entrapment database. For the yeast proteome and human proteins and yeast proteome databases the Arginine and Lysine normalization factors are 1.19 and 1.16, respectively.

Wrong crosslinker control. As a third control we performed searches using a wrong mass crosslinker in addition to BS3 or DSSO, respectively. Both crosslinker masses were reduced by 28.031 Dalton. Note that for these searches, DSSO was treated as non-cleavable. Wrong mass matches were treated separately, i.e. the same PPI matched to correct and wrong crosslinker appeared twice in the results. The error was normalized by a factor of 2 (see above).

Wrong precursor mass control. Spectra passing a 1% heteromeric CSM-FDR were extracted. For these spectra, the precursor mass was downshifted by 28.031 and 42.047 Dalton (corresponding to the mass of two or three methylations). Correct and shifted spectra were searched together, every match to a known wrong spectrum counted as wrong. Here, the error was not corrected as we assume the unknown wrong matches in the correct mass spectra to be only at 1% based on the FDR employed before.

Correlation of protein elution profiles. Proteins were quantified in each SEC fraction as described above. iBAQ values for each protein were normalized by the maximum of the respective protein over the course of fractionation, leading to normalized abundance values between 1 and 0. For each combination of proteins, elution peaks were detected via the scipy python package (1.4.1). In an elution window of 7 or more fractions, the abundances of the proteins were correlated (Pearson). PPIs with elution profiles with a correlation coefficient >0.5 were counted as having similar elution profiles. Code was written in python 3.7.

PPI network comparison with STRING database. For all *E. coli* K12 proteins identified in quantitative proteomics experiments, interaction evidence from the STRING database v10.5[38] was used (scores ranging from 0 to 1000, retrieved from https://string-db.org on 12/19/18). PPIs were accounted as known if the STRING combined score was equal or higher than 150. PPIs were defined as lacking experimental evidence when the STRING experimental score was lower than 150. Note that STRING defines 150 as the lowest cut-off in favour of an interaction.

Plotting protein elution profiles. For the creation of protein elution profiles, fraction-wise iBAQ intensities for each protein from the MaxQuant search were used (see above), hereinafter referred to as abundance. Individual proteins are represented by their gene names while protein complexes, when shown in the figures, are labelled with their respective complex name. Abundance values for protein complexes were averaged for all components as listed in EcoCyc[39], (retrieved from https://ecocyc.org/ on 9/25/18). Plots were created in python 3.7 with pandas 0.24.2 using the seaborn 0.9.0 package.

Protein structural models. Models of protein complexes with mapped residue pairs (Supplementary Data 8) were prepared with xiVIEW[40], python 3.7 with pandas 0.24.2 and ChimeraX 0.92[41].

All structural PPI models were downloaded from the protein data bank (https://www.rcsb.org/): PDB 5t4O [https://doi.org/10.2210/pdb5T4O/pdb] (ATP synthase[42]), PDB 6RKW [https://doi.org/10.2210/pdb6RKW/pdb] (DNA gyrase[43]), PDB 4PKO [https://doi.org/10.2210/pdb4PKO/pdb] (GroEL[44]), PDB 4S20 [https://doi.org/10.2210/pdb4S20/pdb] (RapA[45]), PDB 6RIN [https://doi.org/10.2210/pdb6RIN/pdb] (GreB[46]), PDB 5MS0 [https://doi.org/10.2210/pdb5MS0/pdb] (NusG[47]), PDB 6FLQ [https://doi.org/10.2210/pdb6FLQ/pdb] (NusA[48]) and PDB 4ZH3 [https://doi.org/10.2210/pdb4ZH3/pdb] (RpoD[49]).

Database search and FDR calculation for sulfo-SDA crosslinked pulldowns. A recalibration of the precursor *m/z* was conducted based on high-confidence linear peptide identifications[37]. The re-calibrated peak lists were searched against the sequences of proteins identified in a given pull-down and with an iBAQ ≥ 5e6 along with their reversed sequences (as decoys) using xiSEARCH (v.1.7.6.2) for identification. MS-cleavability of the sulfo-SDA crosslinker was considered[50]. Final crosslink lists were compiled using the identified candidates filtered to 2% FDR on residue pair-level and 5% on PPI level with xiFDR v.2.1.5[17].

RNA polymerase binding site of YacL. An I-TASSER41 (v.5.1)[51] model for YacL was generated with default settings based on the Uniprot sequence (see above). DisVis (v.2.0)[52] ran under default settings, with YacL as scanning model and fixed model PDB 6C6U [https://doi.org/10.2210/pdb6c6u/pdb][53] with residue 118–127 of NusG modelled using the Modeller[54] plug-in in Chimera[55]. Residue pairs of YacL to RNAP and NusG were used as restraints with a minimal distance of 2 Å, and a maximal distance of 30 or 20 Å for DSSO/BS3 and sulfo-SDA, respectively. The density displayed in Fig. 4d corresponds to the accessible interaction space with 14 satisfied restraints. The I-TASSER model was placed for visualisation purposes only.

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

Data availability

Raw data and MaxQuant outputs from quantitative proteomics SEC-MS experiments were deposited with the ProteomeXchange Consortium partner repository jPOSTrepo under the accession codes JPST000843[56] and PXD019004. Raw data and MaxQuant outputs from quantitative proteomics AP-MS experiments were deposited with the ProteomeXchange Consortium partner repository jPOSTrepo under the accession codes JPST001090[56] and PXD024146. All raw data, peak lists and search result files from BS3/DSSO crosslinking experiments in the SEC fractions and after multidimensional fractionation were deposited with the ProteomeXchange Consortium partner repository jPOSTrepo under the accession codes JPST000845[56] and PXD019120. All raw data, peak lists and search result files from affinity-enrichment and crosslinking experiments were deposited with the ProteomeXchange Consortium partner repository jPOSTrepo under the accession JPST001091[56] and PXD024148. We accessed the STRING database (v10.5) via https://string-db.org/. The new link for this version is https://version-10-5.string-db.org/. The used resource can be downloaded using the following link: https://version-10-5.string-db.org/download/protein.links.detailed.v10.5/511145.protein.links.detailed.v10.5.txt.gz. Models from the protein data bank (PDB) can be found under the following links: PDB 5t4O [https://doi.org/10.2210/pdb5T4O/pdb] (ATP synthase[42]), PDB 6RKW [https://doi.org/10.2210/pdb6RKW/pdb] (DNA gyrase[43]), PDB 4PKO [https://doi.org/10.2210/pdb4PKO/pdb] (GroEL[44]), PDB 4S20 [https://doi.org/10.2210/pdb4S20/pdb] (RapA[45]), PDB 6RIN [https://doi.org/10.2210/pdb6RIN/pdb] (GreB[46]), PDB 5MS0 [https://doi.org/10.2210/pdb5MS0/pdb] (NusG[47]), PDB 6FLQ [https://doi.org/10.2210/pdb6FLQ/pdb] (NusA[48]), PDB 4ZH3 [https://doi.org/10.2210/pdb4ZH3/pdb] (RpoD[49]), PDB 6C6U [https://doi.org/10.2210/pdb6c6u/pdb] (RNAP-NusG[53]). Source data are provided with this paper.

Code availability

The xiFDR version[57] used in this manuscript (v2.0.dev) is available via Zenodo at https://doi.org/10.5281/zenodo.4682917. More recent xiFDR versions can be downloaded from https://github.com/Rappsilber-Laboratory/xiFDR or https://www.rappsilberlab.org/software/xifdr/.

Received: 21 August 2020; Accepted: 28 April 2021;
Published online: 11 June 2021

References

1. O'Reilly, F. J. & Rappsilber, J. Cross-linking mass spectrometry: methods and applications in structural, molecular and systems biology. *Nat. Struct. Mol. Biol.* **25**, 1000–1008 (2018).

2. Chavez, J. D., Keller, A., Zhou, B., Tian, R. & Bruce, J. E. Cellular interactome dynamics during paclitaxel treatment. *Cell Rep.* **29**, 2371–2383.e5 (2019).
3. Steigenberger, B., Pieters, R. J., Heck, A. J. R. & Scheltema, R. A. PhoX: an IMAC-enrichable cross-linking reagent. *ACS Cent. Sci.* **5**, 1514–1522 (2019).
4. 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).
5. Mendes, M. L. et al. An integrated workflow for crosslinking mass spectrometry. *Mol. Syst. Biol.* **15**, e8994 (2019).
6. Ryl, P. S. J. et al. In situ structural restraints from crosslinking mass spectrometry in human mitochondria. *J. Proteome Res.* https://doi.org/10.1021/acs.jproteome.9b00541 (2019).
7. Gonzalez-Lozano, M. A. et al. Stitching the synapse: cross-linking mass spectrometry into resolving synaptic protein interactions. *Sci. Adv.* **6**, eaax5783 (2020).
8. Makepeace, K. A. T. et al. Improving identification of in-organello protein-protein interactions using an affinity-enrichable, isotopically coded, and mass spectrometry-cleavable chemical crosslinker. *Mol. Cell. Proteomics* **19**, 624–639 (2020).
9. Bartolec, T. K. et al. Cross-linking mass spectrometry analysis of the yeast nucleus reveals extensive protein-protein interactions not detected by systematic two-hybrid or affinity purification-mass spectrometry. *Anal. Chem.* **92**, 1874–1882 (2020).
10. Linden, A. et al. A cross-linking mass spectrometry approach defines protein interactions in yeast mitochondria. *Mol. Cell. Proteomics* https://doi.org/10.1074/mcp.RA120.002028 (2020).
11. 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).
12. Yugandhar, K. et al. MaXLinker: proteome-wide cross-link identifications with high specificity and sensitivity. *Mol. Cell. Proteomics* **19**, 554–568 (2020).
13. O'Reilly, F. J. et al. In-cell architecture of an actively transcribing-translating expressome. *Science* **369**, 554–557 (2020).
14. Maiolica, A. et al. Structural analysis of multiprotein complexes by cross-linking, mass spectrometry, and database searching. *Mol. Cell. Proteomics* **6**, 2200–2211 (2007).
15. Walzthoeni, T. et al. False discovery rate estimation for cross-linked peptides identified by mass spectrometry. *Nat. Methods* **9**, 901–903 (2012).
16. Fischer, L. & Rappsilber, J. Quirks of error estimation in cross-linking/mass spectrometry. *Anal. Chem.* **89**, 3829–3833 (2017).
17. Fischer, L. & Rappsilber, J. False discovery rate estimation and heterobifunctional cross-linkers. *PLoS ONE* **13**, e0196672 (2018).
18. Yang, B. et al. Identification of cross-linked peptides from complex samples. *Nat. Methods* **9**, 904–906 (2012).
19. Keller, A., Chavez, J. D., Felt, K. C. & Bruce, J. E. Prediction of an upper limit for the fraction of interprotein cross-links in large-scale in vivo cross-linking studies. *J. Proteome Res.* **18**, 3077–3085 (2019).
20. Beveridge, R., Stadlmann, J., Penninger, J. M. & Mechtler, K. A synthetic peptide library for benchmarking crosslinking-mass spectrometry search engines for proteins and protein complexes. *Nat. Commun.* **11**, 742 (2020).
21. Leitner, A. et al. Towards increased reliability, transparency and accessibility in cross-linking mass spectrometry. *Structure* **28**, 1259–1268 (2020).
22. Schmidt, A. et al. The quantitative and condition-dependent Escherichia coli proteome. *Nat. Biotechnol.* **34**, 104–110 (2016).
23. Butland, G. et al. Interaction network containing conserved and essential protein complexes in Escherichia coli. *Nature* **433**, 531–537 (2005).
24. Belsom, A., Schneider, M., Fischer, L., Brock, O. & Rappsilber, J. Serum albumin domain structures in human blood serum by mass spectrometry and computational biology. *Mol. Cell. Proteomics* **15**, 1105–1116 (2016).
25. Matzinger, M., Kandioller, W., Doppler, P., Heiss, E. H. & Mechtler, K. Fast and highly efficient affinity enrichment of Azide-A-DSBSO cross-linked peptides. *J. Proteome Res.* **19**, 2071–2079 (2020).
26. Chen, Z. A. & Rappsilber, J. Quantitative cross-linking/mass spectrometry to elucidate structural changes in proteins and their complexes. *Nat. Protoc.* **14**, 171–201 (2019).
27. Yugandhar, K., Wang, T.-Y., Wierbowski, S. D., Shayhidin, E. E. & Yu, H. Structure-based validation can drastically underestimate error rate in proteome-wide cross-linking mass spectrometry studies. *Nat. Methods* **17**, 985–988 (2020).
28. Giese, S. H., Belsom, A., Sinn, L., Fischer, L. & Rappsilber, J. Noncovalently associated peptides observed during liquid chromatography-mass spectrometry and their effect on cross-link analyses. *Anal. Chem.* **91**, 2678–2685 (2019).
29. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
30. Schwanhäusser, B. et al. Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
31. Tyanova, S. et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731–740 (2016).
32. Horne, J. E. et al. Rapid mapping of protein interactions using tag-transfer photocrosslinkers. *Angew. Chem. Int. Ed.* **57**, 16688–16692 (2018).
33. Leitner, A. et al. Expanding the chemical cross-linking toolbox by the use of multiple proteases and enrichment by size exclusion chromatography. *Mol. Cell. Proteomics* **11**, M111.014126 (2012).
34. Giese, S. H., Belsom, A. & Rappsilber, J. Optimized fragmentation regime for diazirine photo-cross-linked peptides. *Anal. Chem.* **88**, 8239–8247 (2016).
35. Kolbowski, L., Mendes, M. L. & Rappsilber, J. Optimizing the parameters governing the fragmentation of cross-linked peptides in a tribrid mass spectrometer. *Anal. Chem.* **89**, 5311–5318 (2017).
36. Chambers, M. C. et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).
37. Lenz, S., Giese, S. H., Fischer, L. & Rappsilber, J. In-search assignment of monoisotopic peaks improves the identification of cross-linked peptides. *J. Proteome Res.* **17**, 3923–3931 (2018).
38. Szklarczyk, D. et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
39. Keseler, I. M. et al. The EcoCyc database: reflecting new knowledge about Escherichia coli K-12. *Nucleic Acids Res.* **45**, D543–D550 (2017).
40. Combe, C. W., Fischer, L. & Rappsilber, J. xiNET: cross-link network maps with residue resolution. *Mol. Cell. Proteomics* **14**, 1137–1147 (2015).
41. Goddard, T. D. et al. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci.* **27**, 14–25 (2018).
42. Sobti, M. et al. Cryo-EM structures of the autoinhibited E. coli ATP synthase in three rotational states. *elife* **5**, e21598 (2016).
43. Vanden Broeck, A., Lotz, C., Ortiz, J. & Lamour, V. Cryo-EM structure of the complete E. coli DNA gyrase nucleoprotein complex. *Nat. Commun.* **10**, 4935 (2019).
44. Fei, X., Ye, X., LaRonde, N. A. & Lorimer, G. H. Formation and structures of GroEL:GroES2 chaperonin footballs, the protein-folding functional form. *Proc. Natl Acad. Sci. USA* **111**, 12775–12780 (2014).
45. Liu, B., Zuo, Y. & Steitz, T. A. Structural basis for transcription reactivation by RapA. *Proc. Natl Acad. Sci. USA* **112**, 2006–2010 (2015).
46. Abdelkareem, M. 'men et al. Structural basis of transcription: RNA polymerase backtracking and its reactivation. *Mol. Cell* **75**, 298–309.e4 (2019).
47. Said, N. et al. Structural basis for λN-dependent processive transcription antitermination. *Nat. Microbiol* **2**, 17062 (2017).
48. Guo, X. et al. Structural basis for NusA stabilized transcriptional pausing. *Mol. Cell* **69**, 816–827.e4 (2018).
49. Feng, Y. et al. Structural basis of transcription inhibition by CBR hydroxamidines and CBR pyrazoles. *Structure* **23**, 1470–1481 (2015).
50. Iacobucci, C. et al. Carboxyl-photo-reactive MS-cleavable cross-linkers: unveiling a hidden aspect of diazirine-based reagents. *Anal. Chem.* **90**, 2805–2809 (2018).
51. Yang, J. & Zhang, Y. I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res.* **43**, W174–W181 (2015).
52. van Zundert, G. C. P. & Bonvin, A. M. J. J. DisVis: quantifying and visualizing accessible interaction space of distance-restrained biomolecular complexes. *Bioinformatics* **31**, 3222–3224 (2015).
53. Kang, J. Y. et al. Structural basis for transcript elongation control by NusG family universal regulators. *Cell* **173**, 1650–1662.e14 (2018).
54. Sali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).
55. Pettersen, E. F. et al. UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
56. Okuda, S. et al. jPOSTrepo: an international standard data repository for proteomes. *Nucleic Acids Res.* **45**, D1107–D1111 (2017).
57. Lenz, S. et al. Reliable identification of protein-protein interactions by crosslinking mass spectrometry. *Zenodo* https://doi.org/10.5281/zenodo.4682917 (2021).

Acknowledgements

We would like to thank Richard Scheltema, Alexander Leitner, Andrea Sinz, Michael Hoopmann, Marc Wilkins, Fan Liu, Henning Urlaub, James Bruce, Si-Min He, Meng-Qiu Dong and Dermot Harnett for comments on the manuscript. We thank Tabea Schütze for fermenting *E. coli*. The work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 2008—390540038—UniSysCat, by grant no. 392923329/GRK2473, grant no. 426290502 and by the Wellcome Trust through a Senior Research Fellowship to J.R. (103139). The Wellcome Centre for Cell Biology is supported by core funding from the Wellcome Trust (203149).

Author contributions

F.O., L.S., S.L. and J.R. designed the experiments; L.S., F.W. and F.O. prepared the samples. L.S. did the affinity-enrichment experiments. L.S., S.L. and L.F. collected and

processed Crosslinking MS data; L.F. designed and implemented xiFDR software; S.L., L.S., F.O. and J.R. prepared figures and wrote the manuscript with input from all authors.

Funding
Open Access funding enabled and organized by Projekt DEAL.

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/s41467-021-23666-z.

Correspondence and requests for materials should be addressed to J.R.

Peer review information *Nature Communications* thanks Si-Min He, Nir Kalisman and the other, anonymous, reviewer for their contribution to the peer review of this work. Peer reviewer reports are available.

Outlook

Crosslinking MS is a powerful tool producing valuable information to aid structural biology. Its application to purified complexes can by now be seen as a standardised and easy workflow and is routinely applied in *in vitro* structural studies. As a complementary method to high-resolution structural biology techniques like electron microscopy in integrative structural biology approaches, it aids in solving lower resolution data or regions (O'Reilly et al., 2020). In addition, crosslinking MS is becoming a useful tool for systems biology studies. A set of successful large-scale, and *in situ*, studies has been published, delivering PPIs and structural information at the same time (Chavez et al., 2019; Gonzalez-Lozano et al., 2020; Linden et al., 2020). However, the experimental effort required for large-scale crosslinking is still considerable and often only covers the most abundant parts of the proteome. More developments will be needed to increase depth and feasibility of in-cell crosslinking MS.

In this thesis I have demonstrated developments in the crosslinking MS pipeline that yielded large improvements in data acquisition and reliability. Any further optimization, wet or dry lab, needs thorough analysis of the data. For example, to further improve database search results, implementation of machine learning in the crosslinking search engines seems like an obvious next step as it has been shown to be beneficial in standard proteomics search engines (The et al., 2016), although it will require careful implementation to avoid overtraining.

One sometimes overlooked optimization aspect is the spectral quality. Many people have tried to optimise for fast acquisitions to gain as many spectra as possible. However, as spectra are the basis for any identification, to further improve crosslinking analyses it is required that their quality improves further. While it is possible some gains will be made by invention of more sensitive mass spectrometers, another way to improve spectral quality is by increasing the abundance of the crosslinked peptides. Besides chromatographic enrichment, enrichment of the crosslinker itself has been developed, but is yet to be widely employed. Recently, mass spectrometer vendors have made the instrument application programming interface available to users. This allows real-time analysis of the data and on the fly decision making during acquisition and has been used in linear proteomics already (Wichmann et al., 2019). Adapting this to crosslinking MS workflows offers multiple routes to increase spectral quality and depth, e.g. by optimising precursor selection (currently hindered by the dynamic range problem on MS1) or choice of fragmentation schemas.

Another focus should also be the reliability of the crosslinking MS results. This requires an established error control, which we demonstrated in this work for large-scale crosslinking data and their PPIs. We provide a dataset to test search and FDR approaches and demonstrate multiple controls transferable to other samples, which allows adaptation of these validations by other crosslinking MS labs. As crosslinking PPIs are starting to be fed into public databases used by others (Schweppe et al., 2016), their reliability is getting more important. Improved quality of crosslinking data will also lead to growing credibility outside of the field.

While the work of this thesis has focused on NHS-ester crosslinkers, photo-crosslinking with the UV-activatable crosslinker SDA on a larger scale is more difficult. Because it can link to any amino acid on one side, the crosslinked peptide mixture is more complex and each

crosslink less abundant. Besides experimental optimisation, large scale photo-crosslinking will require more computational efforts, i.e. further optimisation of the database search.

In the next years, proteome-wide crosslinking will become more feasible, enabling its application with less effort and by a wider user community. Increased depth of the analysis will lead to more detailed interaction networks that also include lower abundant proteins. Taken together the field is racing towards a future where the protein interaction maps can be generated in cells and in tissues without gene-tagging or cell disruption. The quantitative approaches that are possible with proteomics will allow comparison of these cellular interaction maps with disease states.

Acknowledgements

A big thank you goes to my family and friends, who have accompanied me to where I am today. To my parents, who I knew I could always turn to if I needed something. To my sisters, who would always be just a phone call or message away if there was something to talk about. And to my brother, who was always happy to be here and who I hope will keep being curious.
Theresa, you have been a big emotional support throughout the years and I'm happy to have such a good and long lasting friend. And to Sevgi - we have made the best out of a difficult start and I want to thank you for the open ear and support over the past months.

Doing this thesis required a bit more extra help than for others. Anne, Ralf, Petra and many others, you often have done more for me than what was required and had to share my ups and downs – thank you.

My time in the Rappsilber lab involved a lot of work, but at the same time a great deal of fun. First, I want to thank Juri for creating the environment which I very much enjoyed working in the past years and for giving me the opportunity to do my PhD there. I have learned many things besides crosslinking from you and could always ask for valuable advice if needed. I have appreciated the many scientific discussions we had, which I hope will be continued in the future in one way or the other.

Francis, besides handling my many outbursts gracefully, you have taught me your approach to science and gave me confidence in me and my work over the past years. Andrea, you had a lot of knowledge to share and I enjoyed learning from you so much. You both have been massive influences on me and I'm happy to call you my friends.
Sven, I am thankful to you for introducing me to crosslinking when I started out in the group. Lutz, you have always been there for any (technical) support or detailed discussions. Lars, you have made working on our fastest project as enjoyable as possible. Ludwig, you have been a great colleague - and friend - to share drinks, banter, or extensive discussions. Fabi, besides the almost never ending supply of coffee, it was a pleasure to share an office with you.

I don't think I would be where I am today without the influences the team has had on me. Everyone in the lab has supported me in one way or the other over the years. Thank you all!

References

Audain, E., Uszkoreit, J., Sachsenberg, T., Pfeuffer, J., Liang, X., Hermjakob, H., Sanchez, A., Eisenacher, M., Reinert, K., Tabb, D. L., Kohlbacher, O., & Perez-Riverol, Y. (2017). In-depth analysis of protein inference algorithms using multiple search engines and well-defined metrics. *Journal of Proteomics*, *150*, 170–182. https://doi.org/10.1016/j.jprot.2016.08.002

Bartolec, T. K., Smith, D.-L., Pang, C. N. I., Xu, Y. D., Hamey, J. J., & Wilkins, M. R. (2020). Cross-linking Mass Spectrometry Analysis of the Yeast Nucleus Reveals Extensive Protein-Protein Interactions Not Detected by Systematic Two-Hybrid or Affinity Purification-Mass Spectrometry. *Analytical Chemistry*, *92*(2), 1874–1882. https://doi.org/10.1021/acs.analchem.9b03975

Belsom, A., & Rappsilber, J. (2021). Anatomy of a crosslinker. *Current Opinion in Chemical Biology*, *60*, 39–46. https://doi.org/10.1016/j.cbpa.2020.07.008

Beveridge, R., Stadlmann, J., Penninger, J. M., & Mechtler, K. (2020). A synthetic peptide library for benchmarking crosslinking-mass spectrometry search engines for proteins and protein complexes. *Nature Communications*, *11*(1), 742. https://doi.org/10.1038/s41467-020-14608-2

Böhning, J., & Bharat, T. A. M. (2021). Towards high-throughput in situ structural biology using electron cryotomography. *Progress in Biophysics and Molecular Biology*, *160*, 97–103. https://doi.org/10.1016/j.pbiomolbio.2020.05.010

Chavez, J. D., Keller, A., Zhou, B., Tian, R., & Bruce, J. E. (2019). Cellular Interactome Dynamics during Paclitaxel Treatment. *Cell Reports*, *29*(8), 2371–2383.e5. https://doi.org/10.1016/j.celrep.2019.10.063

Chavez, J. D., Lee, C. F., Caudal, A., Keller, A., Tian, R., & Bruce, J. E. (2018). Chemical Crosslinking Mass Spectrometry Analysis of Protein Conformations and Supercomplexes in Heart Tissue. *Cell Systems*, *6*(1), 136–141.e5. https://doi.org/10.1016/j.cels.2017.10.017

Chen, Z. A., Jawhari, A., Fischer, L., Buchen, C., Tahir, S., Kamenski, T., Rasmussen, M., Lariviere, L., Bukowski-Wills, J.-C., Nilges, M., Cramer, P., & Rappsilber, J. (2010). Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry. *The EMBO Journal*, *29*(4), 717–726. https://doi.org/10.1038/emboj.2009.401

Chen, Z.-L., Meng, J.-M., Cao, Y., Yin, J.-L., Fang, R.-Q., Fan, S.-B., Liu, C., Zeng, W.-F., Ding, Y.-H., Tan, D., Wu, L., Zhou, W.-J., Chi, H., Sun, R.-X., Dong, M.-Q., & He, S.-M. (2019). A high-speed search engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides. *Nature Communications*, *10*(1), 3404. https://doi.org/10.1038/s41467-019-11337-z

Cox, J., & Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, *26*(12), 1367–1372. https://doi.org/10.1038/nbt.1511

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., & Mann, M. (2011). Andromeda: a peptide search engine integrated into the MaxQuant environment. *Journal of Proteome Research*, *10*(4), 1794–1805. https://doi.org/10.1021/pr101065j

Elias, J. E., & Gygi, S. P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*, *4*(3), 207–214. https://doi.org/10.1038/nmeth1019

Fischer, L., & Rappsilber, J. (2017). Quirks of Error Estimation in Cross-Linking/Mass Spectrometry. *Analytical Chemistry*, *89*(7), 3829–3833. https://doi.org/10.1021/acs.analchem.6b03745

Fritzsche, R., Ihling, C. H., Götze, M., & Sinz, A. (2012). Optimizing the enrichment of cross-linked products for mass spectrometric protein analysis. *Rapid Communications in Mass Spectrometry:*

RCM, *26*(6), 653–658. https://doi.org/10.1002/rcm.6150

Giese, S. H., Fischer, L., & Rappsilber, J. (2016). A Study into the Collision-induced Dissociation (CID) Behavior of Cross-Linked Peptides*. *Molecular & Cellular Proteomics: MCP*, *15*(3), 1094–1104. https://doi.org/10.1074/mcp.M115.049296

Giese, S. H., Sinn, L. R., Wegner, F., & Rappsilber, J. (2021). Retention time prediction using neural networks increases identifications in crosslinking mass spectrometry. *Nature Communications*, *12*(1), 3237. https://doi.org/10.1038/s41467-021-23441-0

Gonzalez-Lozano, M. A., Koopmans, F., Sullivan, P. F., Protze, J., Krause, G., Verhage, M., Li, K. W., Liu, F., & Smit, A. B. (2020). Stitching the synapse: Cross-linking mass spectrometry into resolving synaptic protein interactions. *Science Advances*, *6*(8), eaax5783. https://doi.org/10.1126/sciadv.aax5783

Götze, M., Iacobucci, C., Ihling, C. H., & Sinz, A. (2019). A Simple Cross-Linking/Mass Spectrometry Workflow for Studying System-wide Protein Interactions. *Analytical Chemistry*, *91*(15), 10236–10244. https://doi.org/10.1021/acs.analchem.9b02372

Hoopmann, M. R., Zelter, A., Johnson, R. S., Riffle, M., MacCoss, M. J., Davis, T. N., & Moritz, R. L. (2015). Kojak: efficient analysis of chemically cross-linked protein complexes. *Journal of Proteome Research*, *14*(5), 2190–2198. https://doi.org/10.1021/pr501321h

Houel, S., Abernathy, R., Renganathan, K., Meyer-Arendt, K., Ahn, N. G., & Old, W. M. (2010). Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies. *Journal of Proteome Research*, *9*(8), 4152–4160. https://doi.org/10.1021/pr1003856

Kao, A., Chiu, C.-L., Vellucci, D., Yang, Y., Patel, V. R., Guan, S., Randall, A., Baldi, P., Rychnovsky, S. D., & Huang, L. (2011). Development of a novel cross-linking strategy for fast and accurate identification of cross-linked peptides of protein complexes. *Molecular & Cellular Proteomics: MCP*, *10*(1), M110.002212. https://doi.org/10.1074/mcp.M110.002212

Klammer, A. A., Yi, X., MacCoss, M. J., & Noble, W. S. (2007). Improving tandem mass spectrum identification using peptide retention time prediction across diverse chromatography conditions. *Analytical Chemistry*, *79*(16), 6111–6118. https://doi.org/10.1021/ac070262k

Kolbowski, L., Mendes, M. L., & Rappsilber, J. (2017). Optimizing the Parameters Governing the Fragmentation of Cross-Linked Peptides in a Tribrid Mass Spectrometer. *Analytical Chemistry*, *89*(10), 5311–5318. https://doi.org/10.1021/acs.analchem.6b04935

Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D., & Nesvizhskii, A. I. (2017). MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nature Methods*, *14*(5), 513–520. https://doi.org/10.1038/nmeth.4256

Kristensen, A. R., Gsponer, J., & Foster, L. J. (2012). A high-throughput approach for measuring temporal changes in the interactome. *Nature Methods*, *9*(9), 907–909. https://doi.org/10.1038/nmeth.2131

Lasker, K., Förster, F., Bohn, S., Walzthoeni, T., Villa, E., Unverdorben, P., Beck, F., Aebersold, R., Sali, A., & Baumeister, W. (2012). Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(5), 1380–1387. https://doi.org/10.1073/pnas.1120559109

Lee, B.-G., Merkel, F., Allegretti, M., Hassler, M., Cawood, C., Lecomte, L., O'Reilly, F. J., Sinn, L. R., Gutierrez-Escribano, P., Kschonsak, M., Bravo, S., Nakane, T., Rappsilber, J., Aragon, L., Beck, M., Löwe, J., & Haering, C. H. (2020). Cryo-EM structures of holo condensin reveal a subunit flip-flop mechanism. *Nature Structural & Molecular Biology*, *27*(8), 743–751. https://doi.org/10.1038/s41594-020-0457-x

Leitner, A., Reischl, R., Walzthoeni, T., Herzog, F., Bohn, S., Förster, F., & Aebersold, R. (2012). Expanding the chemical cross-linking toolbox by the use of multiple proteases and enrichment

by size exclusion chromatography. *Molecular & Cellular Proteomics: MCP*, *11*(3), M111.014126. https://doi.org/10.1074/mcp.M111.014126

Lenz, S., Giese, S. H., Fischer, L., & Rappsilber, J. (2018). In-Search Assignment of Monoisotopic Peaks Improves the Identification of Cross-Linked Peptides. Journal of Proteome Research, 17(11), 3923–3931. https://doi.org/10.1021/acs.jproteome.8b00600

Lenz, S., Sinn, L. R., O'Reilly, F. J., Fischer, L., Wegner, F., & Rappsilber, J. (2021). Reliable identification of protein-protein interactions by crosslinking mass spectrometry. *Nature Communications*, *12*(1), 3564. https://doi.org/10.1038/s41467-021-23666-z

Linden, A., Deckers, M., Parfentev, I., Pflanz, R., Homberg, B., Neumann, P., Ficner, R., Rehling, P., & Urlaub, H. (2020). A cross-linking mass spectrometry approach defines protein interactions in yeast mitochondria. *Molecular & Cellular Proteomics: MCP*. https://doi.org/10.1074/mcp.RA120.002028

Liu, F., Lössl, P., Rabbitts, B. M., Balaban, R. S., & Heck, A. J. R. (2018). The interactome of intact mitochondria by cross-linking mass spectrometry provides evidence for coexisting respiratory supercomplexes. *Molecular & Cellular Proteomics: MCP*, *17*(2), 216–232. https://doi.org/10.1074/mcp.RA117.000470

Liu, F., Lössl, P., Scheltema, R., Viner, R., & Heck, A. J. R. (2017). Optimized fragmentation schemes and data analysis strategies for proteome-wide cross-link identification. *Nature Communications*, *8*, 15473. https://doi.org/10.1038/ncomms15473

Mendes, M. L., Fischer, L., Chen, Z. A., Barbon, M., O'Reilly, F. J., Giese, S. H., Bohlke-Schneider, M., Belsom, A., Dau, T., Combe, C. W., Graham, M., Eisele, M. R., Baumeister, W., Speck, C., & Rappsilber, J. (2019). An integrated workflow for crosslinking mass spectrometry. *Molecular Systems Biology*, *15*(9), e8994. https://doi.org/10.15252/msb.20198994

Müller, M. Q., Dreiocker, F., Ihling, C. H., Schäfer, M., & Sinz, A. (2010). Cleavable cross-linker for protein structure analysis: reliable identification of cross-linking products by tandem MS. *Analytical Chemistry*, *82*(16), 6958–6968. https://doi.org/10.1021/ac101241t

O'Reilly, F. J., & Rappsilber, J. (2018). Cross-linking mass spectrometry: methods and applications in structural, molecular and systems biology. *Nature Structural & Molecular Biology*, *25*(11), 1000–1008. https://doi.org/10.1038/s41594-018-0147-0

O'Reilly, F. J., Xue, L., Graziadei, A., Sinn, L., Lenz, S., Tegunov, D., Blötz, C., Singh, N., Hagen, W. J. H., Cramer, P., Stülke, J., Mahamid, J., & Rappsilber, J. (2020). In-cell architecture of an actively transcribing-translating expressome. *Science*, *369*(6503), 554–557. https://doi.org/10.1126/science.abb3758

Parrish, J. R., Gulyas, K. D., & Finley, R. L., Jr. (2006). Yeast two-hybrid contributions to interactome mapping. *Current Opinion in Biotechnology*, *17*(4), 387–393. https://doi.org/10.1016/j.copbio.2006.06.006

Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., & Séraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnology*, *17*(10), 1030–1032. https://doi.org/10.1038/13732

Robinson, P. J., Trnka, M. J., Pellarin, R., Greenberg, C. H., Bushnell, D. A., Davis, R., Burlingame, A. L., Sali, A., & Kornberg, R. D. (2015). Molecular architecture of the yeast Mediator complex. *eLife*, *4*. https://doi.org/10.7554/eLife.08719

Ryl, P. S. J., Bohlke-Schneider, M., Lenz, S., Fischer, L., Budzinski, L., Stuiver, M., Mendes, M. M. L., Sinn, L., O'Reilly, F. J., & Rappsilber, J. (2020). In Situ Structural Restraints from Cross-Linking Mass Spectrometry in Human Mitochondria. *Journal of Proteome Research*, *19*(1), 327–336. https://doi.org/10.1021/acs.jproteome.9b00541

Schäpe, P., Kwon, M. J., Baumann, B., Gutschmann, B., Jung, S., Lenz, S., Nitsche, B., Paege, N., Schütze, T., Cairns, T. C., & Meyer, V. (2019). Updating genome annotation for the microbial

cell factory Aspergillus niger using gene co-expression networks. *Nucleic Acids Research*, *47*(2), 559–569. https://doi.org/10.1093/nar/gky1183

Schweppe, D. K., Zheng, C., Chavez, J. D., Navare, A. T., Wu, X., Eng, J. K., & Bruce, J. E. (2016). XLinkDB 2.0: integrated, large-scale structural analysis of protein crosslinking data. *Bioinformatics* , *32*(17), 2716–2718. https://doi.org/10.1093/bioinformatics/btw232

Steigenberger, B., Pieters, R. J., Heck, A. J. R., & Scheltema, R. A. (2019). PhoX: An IMAC-Enrichable Cross-Linking Reagent. *ACS Central Science*, *5*(9), 1514–1522. https://doi.org/10.1021/acscentsci.9b00416

Tang, X., & Bruce, J. E. (2010). A new cross-linking strategy: protein interaction reporter (PIR) technology for protein-protein interaction studies. *Molecular bioSystems*, *6*(6), 939–947. https://doi.org/10.1039/b920876c

The, M., MacCoss, M. J., Noble, W. S., & Käll, L. (2016). Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. *Journal of the American Society for Mass Spectrometry*, *27*(11), 1719–1727. https://doi.org/10.1007/s13361-016-1460-7

Trnka, M. J., Baker, P. R., Robinson, P. J. J., Burlingame, A. L., & Chalkley, R. J. (2014). Matching Cross-linked Peptide Spectra: Only as Good as the Worse Identification *. *Molecular & Cellular Proteomics: MCP*, *13*(2), 420–434. https://doi.org/10.1074/mcp.M113.034009

von Appen, A., Kosinski, J., Sparks, L., Ori, A., DiGuilio, A. L., Vollmer, B., Mackmull, M.-T., Banterle, N., Parca, L., Kastritis, P., Buczak, K., Mosalaganti, S., Hagen, W., Andres-Pons, A., Lemke, E. A., Bork, P., Antonin, W., Glavy, J. S., Bui, K. H., & Beck, M. (2015). In situ structural analysis of the human nuclear pore complex. *Nature*, *526*(7571), 140–143. https://doi.org/10.1038/nature15381

Walzthoeni, T., Claassen, M., Leitner, A., Herzog, F., Bohn, S., Förster, F., Beck, M., & Aebersold, R. (2012). False discovery rate estimation for cross-linked peptides identified by mass spectrometry. *Nature Methods*, *9*(9), 901–903. https://doi.org/10.1038/nmeth.2103

Wang, G., Wu, W. W., Zhang, Z., Masilamani, S., & Shen, R.-F. (2009). Decoy methods for assessing false positives and false discovery rates in shotgun proteomics. *Analytical Chemistry*, *81*(1), 146–159. https://doi.org/10.1021/ac801664q

Wichmann, C., Meier, F., Virreira Winter, S., Brunner, A.-D., Cox, J., & Mann, M. (2019). MaxQuant.Live Enables Global Targeting of More Than 25,000 Peptides. *Molecular & Cellular Proteomics: MCP*, *18*(5), 982–994. https://doi.org/10.1074/mcp.TIR118.001131

Yugandhar, K., Wang, T.-Y., Leung, A. K.-Y., Lanz, M. C., Motorykin, I., Liang, J., Shayhidin, E. E., Smolka, M. B., Zhang, S., & Yu, H. (2020). MaXLinker: Proteome-wide Cross-link Identifications with High Specificity and Sensitivity. *Molecular & Cellular Proteomics: MCP*, *19*(3), 554–568. https://doi.org/10.1074/mcp.TIR119.001847

Yugandhar, K., Wang, T.-Y., Wierbowski, S. D., Shayhidin, E. E., & Yu, H. (2020). Structure-based validation can drastically underestimate error rate in proteome-wide cross-linking mass spectrometry studies. *Nature Methods*, *17*(10), 985–988. https://doi.org/10.1038/s41592-020-0959-9

Supplement

Supporting Information: In-Search Assignment of Monoisotopic Peaks Improves the Identification of Cross-Linked Peptides

Swantje Lenz[1], Sven H. Giese[1], Lutz Fischer[2], 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

List of Supporting Information

*juri.rappsilber@tu-berlin.de

MS1 based mass range reduction

Since considering multiple precursor masses increases search time, we developed an approach to reduce the range searched in Xi-MPA. For each MS2 spectrum, the precursor peak is identified in the corresponding MS1 spectrum. Then, the most abundant occurrence of this peak is searched in a retention time window of 20 seconds (±10 seconds) and the corresponding MS1 spectrum is extracted. Assuming that the assigned charge state was correct, the newly extracted MS1 spectrum is searched for the true monoisotopic peak, i.e. lighter isotope peaks than the one that was reported in the MS2 spectrum. According to Table S4 each MS2 spectrum gets assigned an individual range of possible precursor masses to be used during Xi-MPA.

Table S4: Xi-MPA mass range reduction.

Lighter Peaks Present	**Search Range**
none	mass range without lightest mass*
continuous (without gaps)	lightest two peaks found
single peaks (with gaps)	mass range up to lightest peak found

* In case the range this approach is done is only up to -2 Da, -1 Da will still be searched here.

This approach was evaluated on the first and last fractions of dataset 3 with a mass range of up to -4 Da. On average, the masses searched in Xi-MPA were reduced by 24% per file, while the number of within PSMs is 97% of the search without range reduction.
Note that this approach increases the time of the preprocessing before search to some extent. Therefore, it is only worthwhile undertaking for searches with a large database, for which the time of the search itself is long.
In previous approaches we tried to incorporate a mass and / or intensity cutoff or dependency when selecting the considered mass range. However, the applied heuristics resulted in significant losses in PSM numbers, presumably because a clear cut in the distributions is missing (Figure 4A and S5).
Python scripts were written using the pyopenms package [1] and are available under https://github.com/Rappsilber-Laboratory/Xi-MPA_scripts.

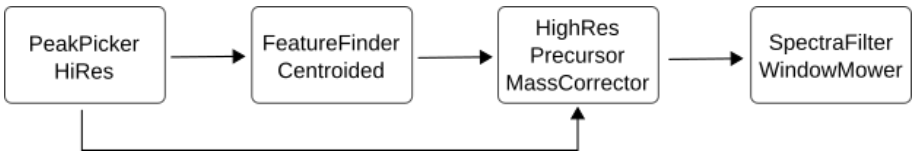


Figure S1: OpenMS preprocessing workflow. PeakPickerHiRes was used with the following settings: 'ms levels' was set to 1 and 'signal to noise' set to 0 (disabled). For the tool FeatureFinderCentroided the following settings were changed: 'feature:min score' - 0.6, 'mass trace:min spectra' - 7, 'isotopic pattern:charge low' - 3, and 'isotopic pattern:charge high' - 7. In HighResPrecursorMassCorrector 'feature:rt tolerance' was changed to 15. SpectraFilterWindowMower was used with 'movetype' - jump, 'windowsize' - 100, and 'peakcount' 20.

test statistic	p-value	H_1	Data	significant
35196.5	2.54E-29	less	0 vs. ref	True
36626.5	4.68E-27	less	-1 vs. ref	True
38792.5	6.30E-32	less	-2 vs. ref	True
39451.0	4.66E-25	less	-3 vs. ref	True
28239.5	4.78E-19	less	-4 vs. ref	True

Table S5: Summary for conducted significance tests for Fig. 3B in the main text. The significance level α was set to 0.05 before the statistical analysis. The Wilcoxon rank sum test with continuity correction was used in R. Abbreviations: ref - reference distance distribution derived from all cross-linkable residues. 0, -1, -2, -3 and -4 denote the subsets of PSMs with the corresponding mass shift of the precursor.

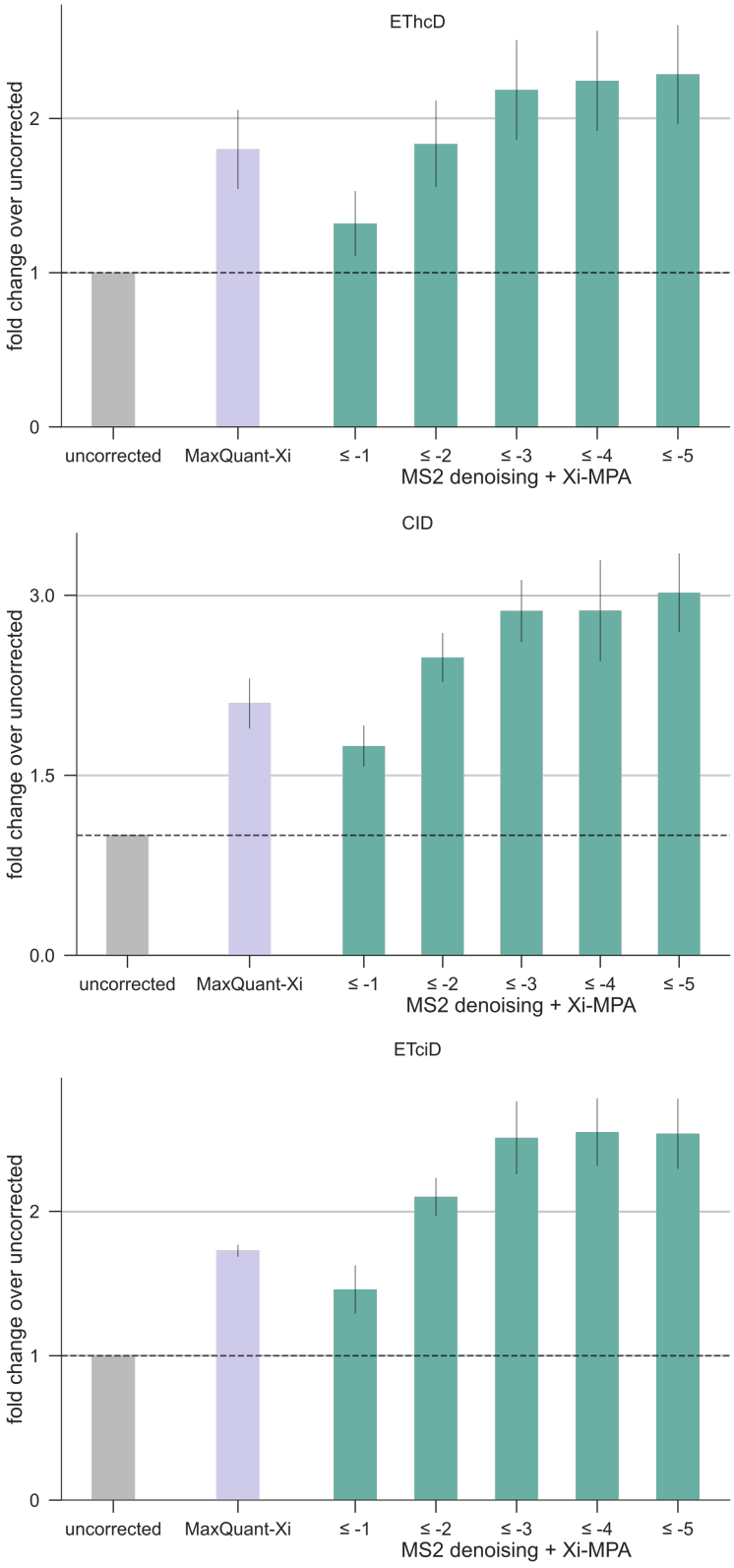


Figure S2: Performance of Xi-MPA on EThcd, CID, and ETciD acquired data of the pseudo-complex dataset. Different ranges in Xi-MPA were tested and evaluated on the number of PSMs. Shown is the mean fold change of the respective setting to the number of PSMs from unprocessed data. For all fragmentation methods, the number of identifications increases compared to the unprocessed data. While for EThcD 251 PSMs were identified for the unprocessed data, 434 PSMs resulted from MaxQuant-Xi and 542 PSMs from Xi-MPA with up to -4 Da. Numbers of identified PSMs for CID data are: 265 PSMs for unprocessed, 552 for MaxQuant-Xi, and 753 for Xi-MPA (-4 Da). Finally, 197 PSMs are identified in unprocessed data for ETciD, while 340 resulted from MaxQuant-Xi and 502 from Xi-MPA. Although the increase of Xi-MPA is smaller for EThcD and ETciD data than for CID and HCD data, it is the approach with the most identifications.

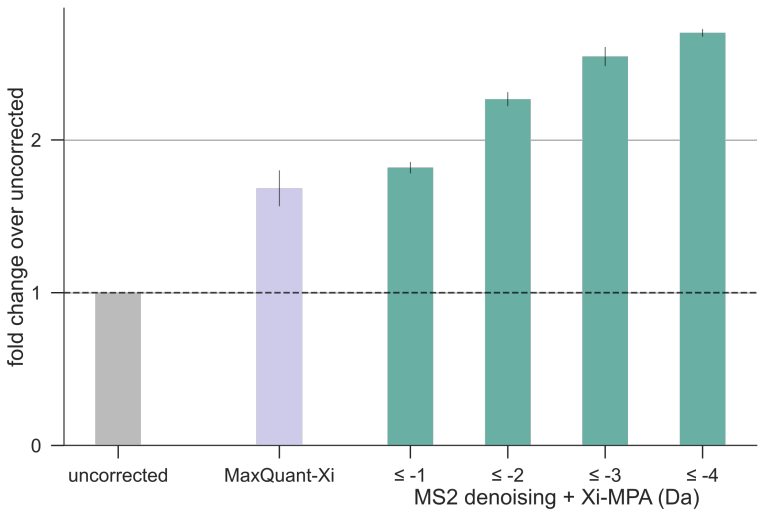


Figure S3: Performance of Xi-MPA on HCD acquisitions of the HSA dataset. The dashed line equals a fold change of 1, meaning the same number of PSMs as in the unprocessed data was identified. Different ranges of Xi-MPA were tested and compared to MaxQuant-Xi results. While the latter led to 1127 PSMs, Xi-MPA with up to -4 Da resulted in 1816 identifications.

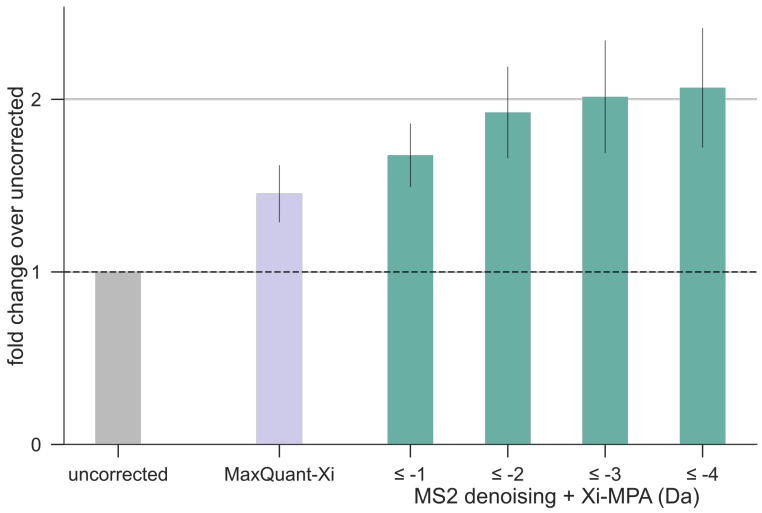


Figure S4: Performance of Xi-MPA on the first and last fraction of the *C. thermophilum* dataset. As for the other two datasets, different ranges of Xi-MPA were compared to MaxQuant-Xi results. MaxQuant led to 2966 identifications, while Xi-MPA with -4 Da identified 4013 PSMs. Considering masses up to -3 Da led to a similar number of PSMs (3945) than up to -4 Da.

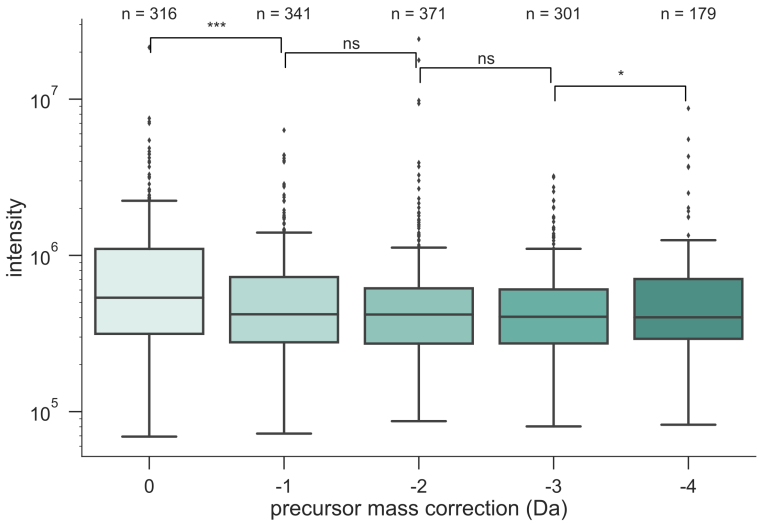


Figure S5: Dependency of the monoisotopic mass correction on precursor intensity. Scans of the pseudo-complex dataset identified in the Xi-MPA search were evaluated regarding their mass correction. Corrections to lighter masses occur more often for precursors with lower intensity. Significance is denoted by asterisks (ns: p-value>0.05, *: p-value<0.05, ***: p-value<0.001).

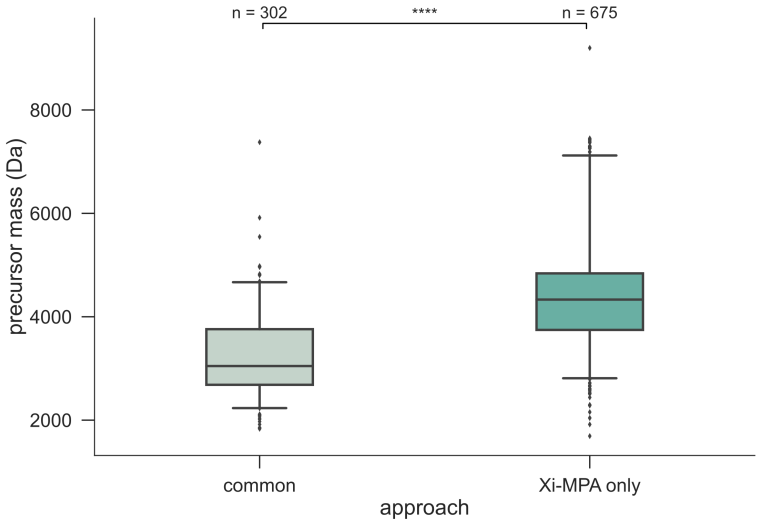


Figure S6: Correction of the monoisotopic mass is more successful for lighter peptides, while Xi-MPA identifies larger peptides more often. Precursor masses of scans identified in all approaches (preprocessing in MaxQuant and OpenMS and Xi-MPA) are compared to scans solely identified in Xi-MPA. (****: p-value<0.0001)

References

[1] Hannes L. Röst, Uwe Schmitt, Ruedi Aebersold, and Lars Malmström. pyopenms: A python-based interface to the openms mass-spectrometry algorithm library. *PROTEOMICS*, 14(1):74–77, 2014.

Supplementary Materials for

Improved peptide backbone fragmentation is the primary advantage of MS-cleavable crosslinkers

Lars Kolbowski[*,1], Swantje Lenz[*,1], Lutz Fischer[1], Ludwig R Sinn[1], Francis J O'Reilly[1], Juri Rappsilber[†,1,2]

[1] Bioanalytics, Institute of Biotechnology, Technische Universität Berlin, 13355 Berlin, Germany
[2] Wellcome Centre for Cell Biology, University of Edinburgh, Edinburgh EH9 3BF, UK

[*] These authors contributed equally to this work.

[†] Corresponding author email address: juri.rappsilber@tu-berlin.de

This file includes:

Supplementary Figures 1-6

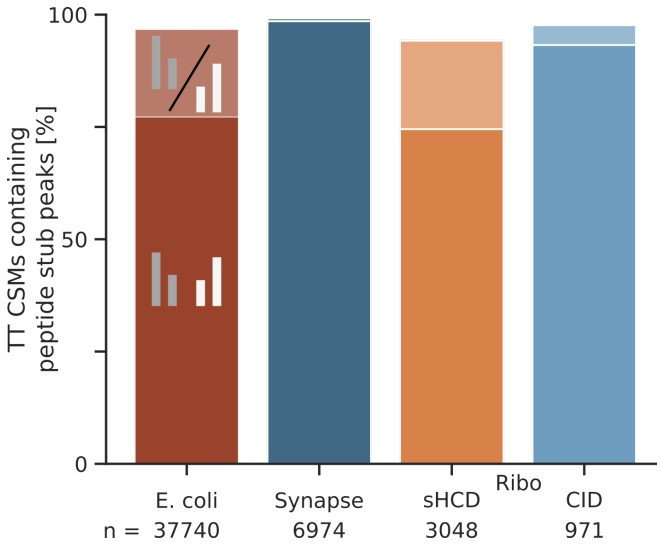
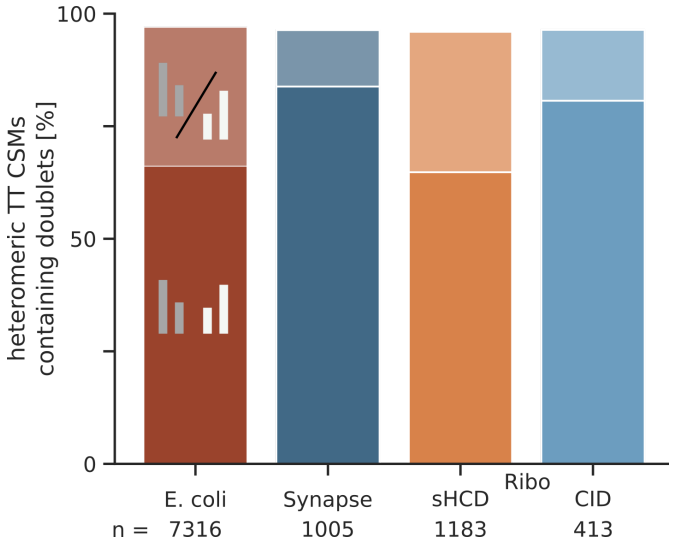


Figure S1. Ratio of identified target-target (TT) CSMs that contain at least one peptide stub peak for one (lighter colour) or both (darker colour) crosslinked peptides across datasets (5% CSM-level FDR).



Figures S2. Ratio of identified heteromeric target-target (TT) CSMs that contain one (lighter colour) or both (darker colour) peptide doublets across datasets (5% CSM-level FDR).

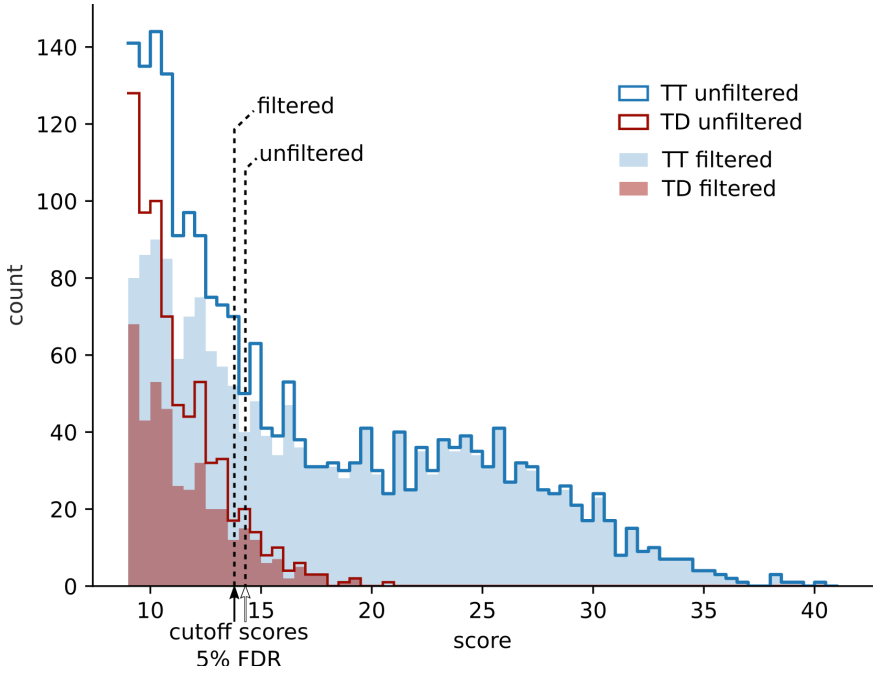


Figure S3. Score distribution of heteromeric matches in the Ribosome HCD dataset. Shown is the distribution of targets and target-decoy matches with and without filtering for peptide doublets. Arrows show the resulting score cutoffs at 5% FDR.

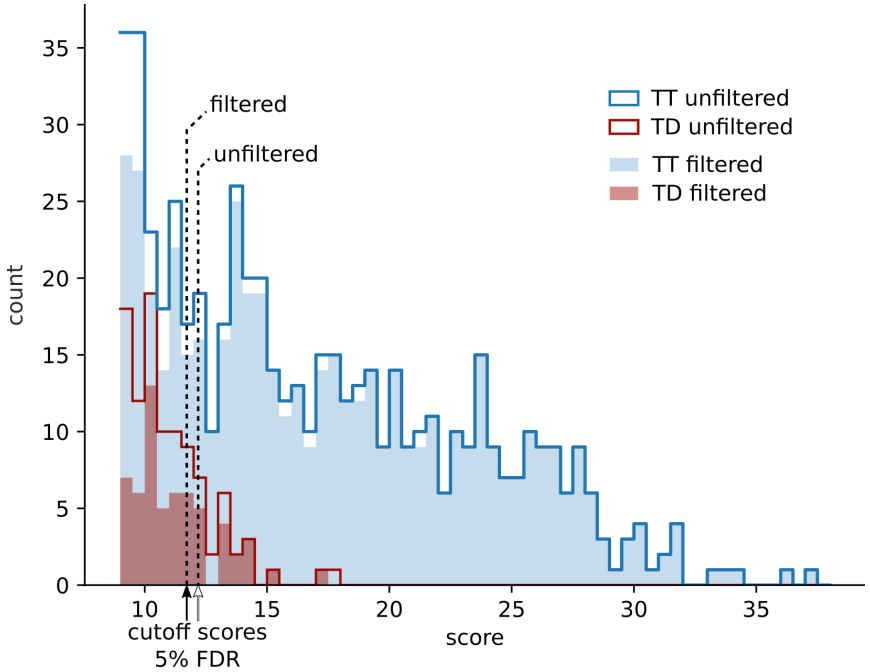


Figure S4. Score distribution of heteromeric matches in the Ribosome CID dataset. Shown is the distribution of targets and target-decoy matches with and without filtering for peptide doublets. Arrows show the resulting score cutoffs at 5% FDR.

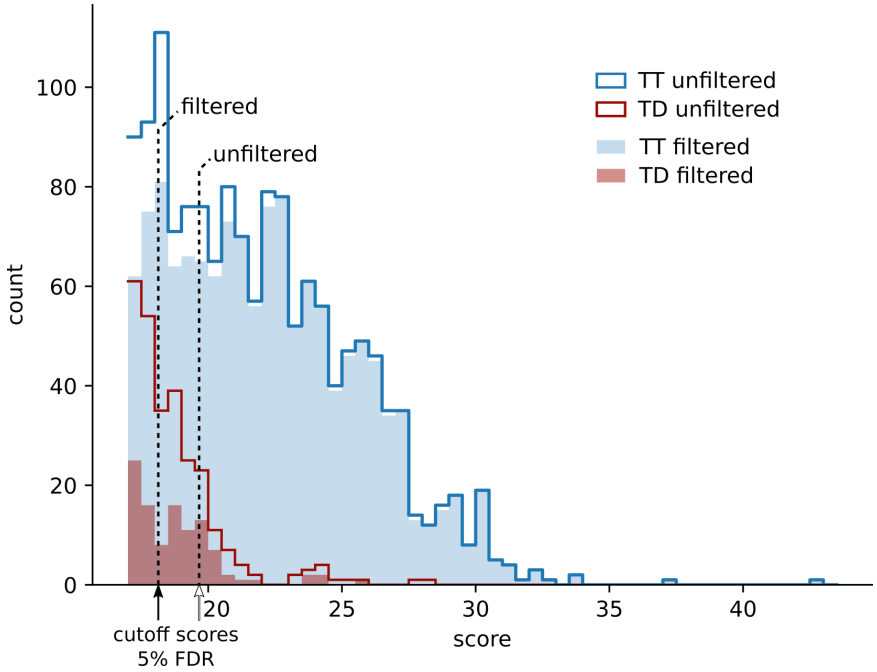


Figure S5. Score distribution of heteromeric matches in the Synapse dataset. Shown is the distribution of targets and target-decoy matches with and without filtering for peptide doublets. Arrows show the resulting score cutoffs at 5% FDR.

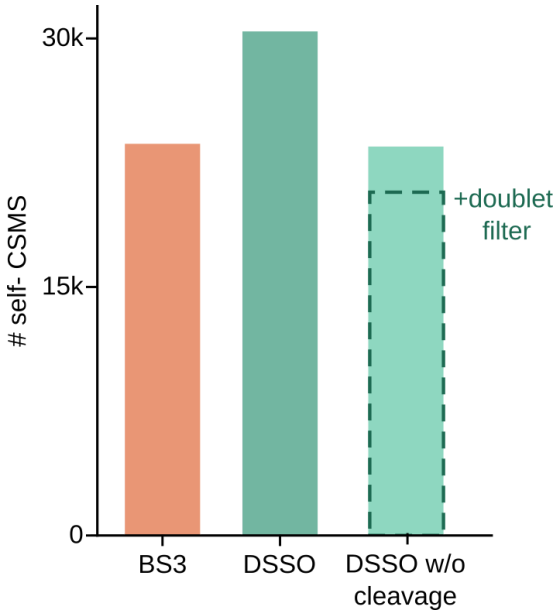


Figure S6. Number of self-CSMs passing 5% CSM-level FDR for BS3 and DSSO. DSSO was additionally searched as a non-cleavable crosslinker and filtered for the presence of peptide doublets.

Supplementary Materials for

Reliable identification of protein-protein interactions by crosslinking mass spectrometry

Swantje Lenz[1]*, Ludwig R. Sinn[1]*, Francis O'Reilly[1]*, Lutz Fischer[1], Fritz Wegner[1], Juri Rappsilber[1,2]

[1] Bioanalytics, Institute of Biotechnology, Technische Universität Berlin, 13355 Berlin, Germany
[2] Wellcome Centre for Cell Biology, University of Edinburgh, Edinburgh EH9 3BF, UK

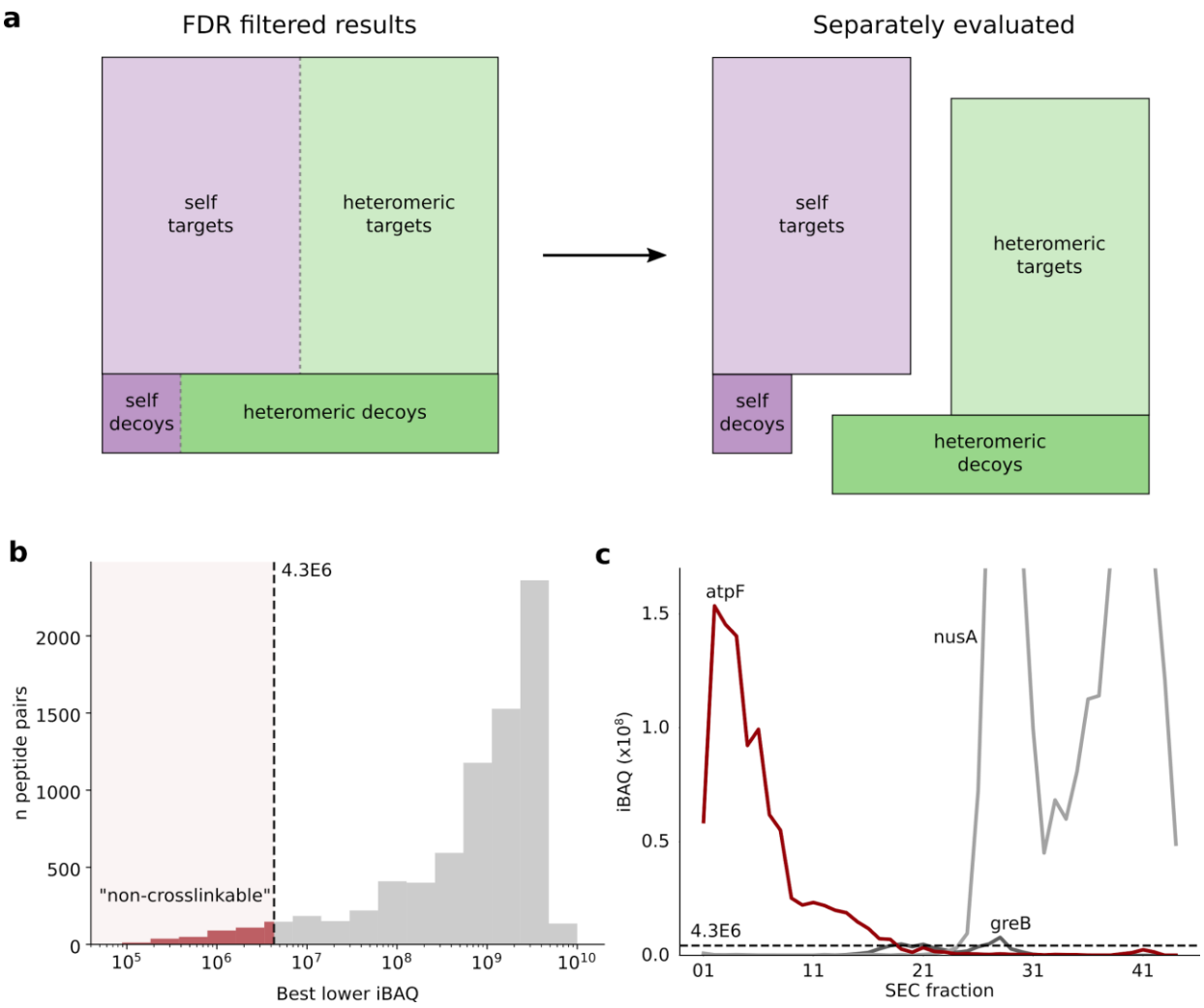
*These authors contributed equally.

Correspondence to: Juri.Rappsilber@tu-berlin.de

This file includes:

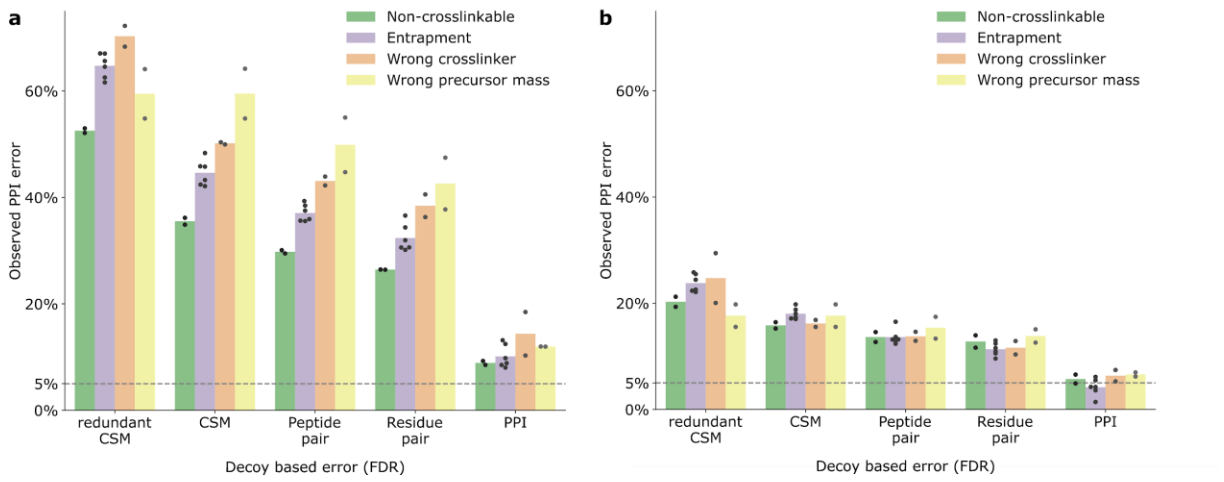
Supplementary Table 1: Recent large scale crosslinking mass spectrometry studies and aspects of their FDR method

Sample	Search / FDR Software	Self- / heteromeric crosslinks separated for FDR	FDR level	threshold	reference
Murine synaptosomes	XlinkX / PD	no	CSM	2%	1
Saccharomyces cerevisiae nucleus	XlinkX / PD	no	CSM	1%	2
Saccharomyces cerevisiae mitochondria	pLink1	no	CSM	1%	3
Several previously published datasets	pLink2	yes	CSM	5%	4
Drosophila melanogaster embryo lysate	MeroX	yes	CSM	1%	5
Human cells	Comet / XLinkProphet	yes	Peptide pair	1%	6
Human cell lysate	MaxLinker	no	Peptide pair	1%	7
Human cell lysate	XlinkX / PD	no	Peptide pair	1%	8
Saccharomyces cerevisiae mitochondria	Kojak	yes	Peptide pair	2%	9
Human cell lysate	xiSEARCH / xiFDR	yes	Residue pair	5%	10
Human mitochondria	xiSEARCH / xiFDR	yes	Residue pair	5%	11
Mycoplasma pneumoniae cells	xiSEARCH / xiFDR	yes	Residue pair & PPI	5%	12



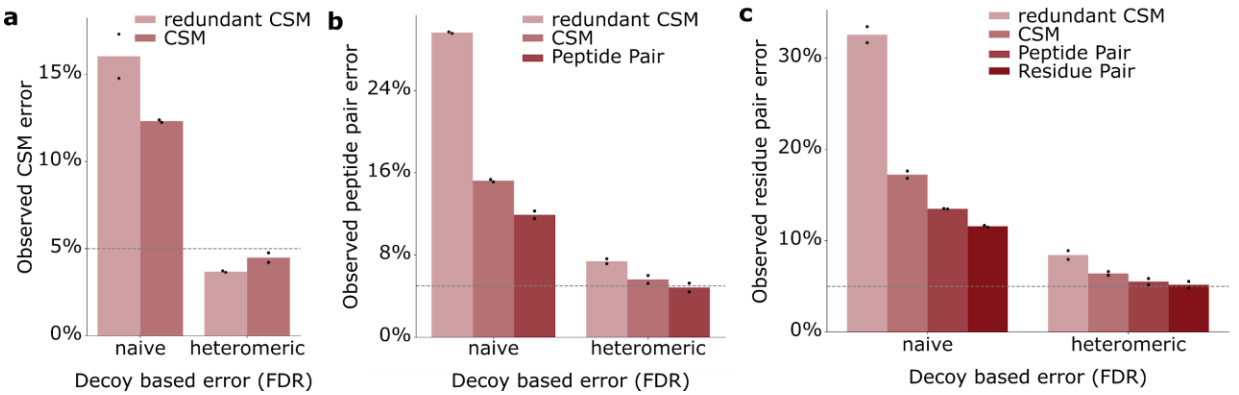
Supplementary Figure 1: Non-crosslinkable control definition.

a Illustration of the distribution of decoy matches for self and heteromeric crosslinks. If considered together for FDR calculation, heteromeric decoy matches will be matched more frequently and make up most of the summed decoy matches. If subsequently heteromeric matches are evaluated separately (e.g. for reporting PPIs), their error will be larger than the previously calculated FDR (which would only be correct for the data as a whole). **b** Abundance distribution of proteins identified with heteromeric crosslinks at a generous 10% heteromeric peptide pair FDR (best lower iBAQ, see Methods). Protein pairs are accepted as plausible if both proteins reach the 5th percentile in the same fraction, otherwise the pair is defined as 'non-crosslinkable'. 544,274 (6% of all possible PPIs in the *E. coli* proteome of 4350 proteins) PPIs are defined as plausible (Supplementary data 7), while 8,914,801 (94%) PPIs are non-crosslinkable. The dashed line represents the chosen 5th-percentile cutoff of 4.3E6. **c** Example of proteins defined either as plausible to crosslink to NusA (GreB, dark grey) or "non-crosslinkable" (AtpF, red). Although AtpF is present in some fractions together with NusA, it is too low abundant in those, therefore the proteins are defined as "non-crosslinkable". In contrast, GreB reaches the cutoff in the same fractions with NusA, therefore the two are crosslinkable. Source data are provided as a Source Data file.



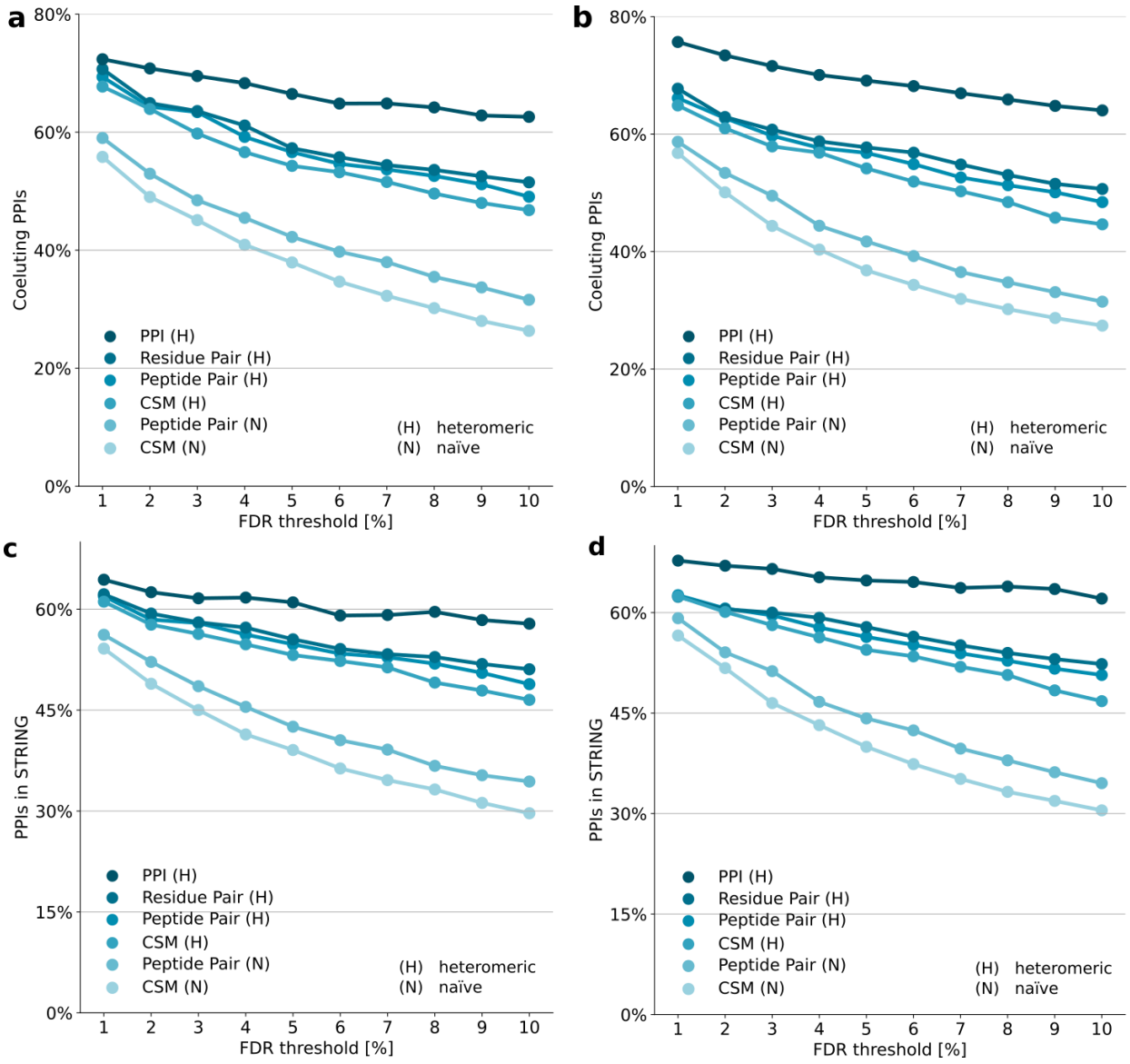
Supplementary Figure 2: Observed PPI-level error using a 5% FDR cutoff at different information levels applied to entrapment, wrong crosslinker, wrong precursor mass and non-crosslinkable controls.

Each bar represents a separate FDR calculation performed at different information levels on each of the controls. The y-axis displays the observed PPI-level errors of the respective controls employing a 5% **a** naive or **b** heteromeric FDR thresholds on different information levels. Bars represent the mean error, with individual data points shown on top (for each crosslinker dataset or entrapment database). Source data are provided as a Source Data file.



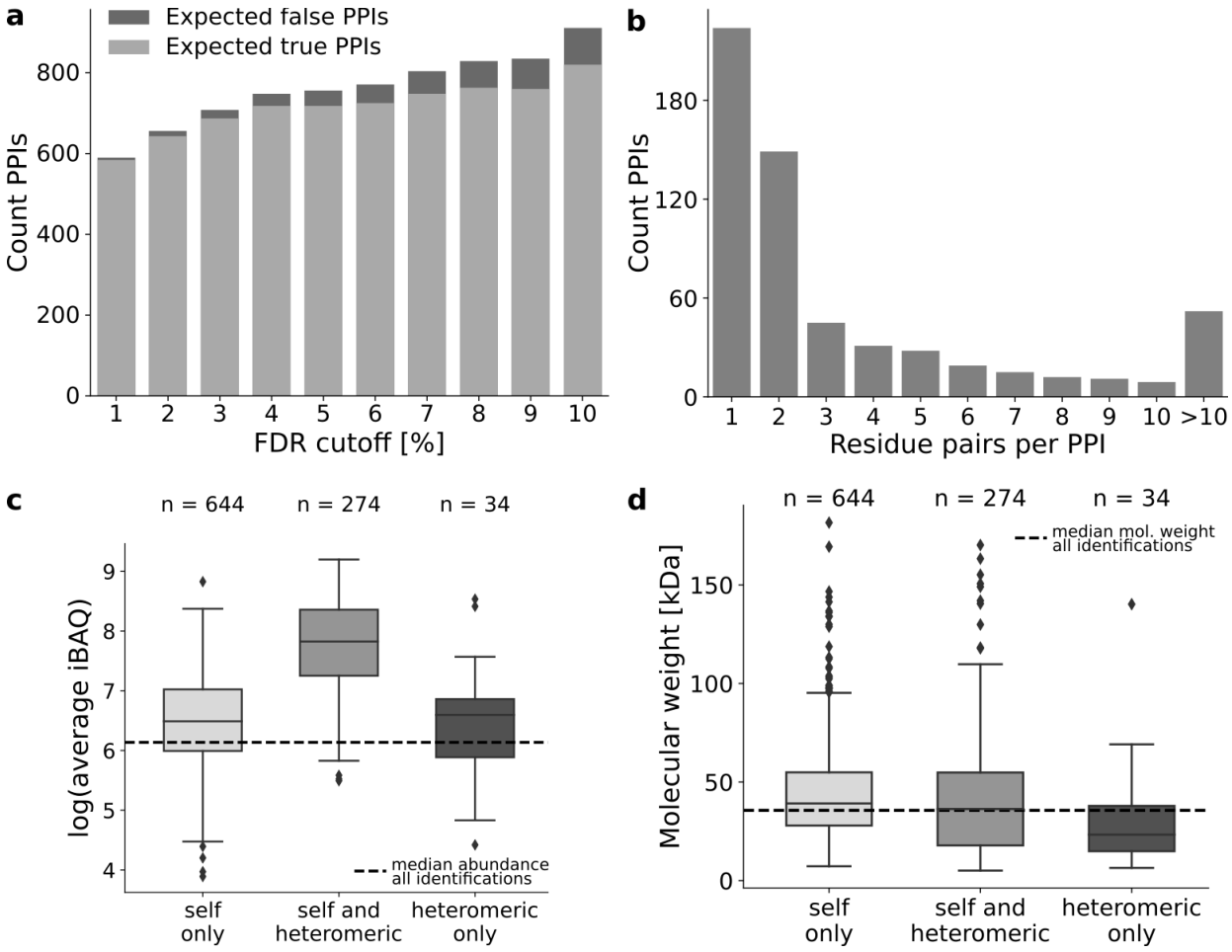
Supplementary Figure 3: **Observed errors using the non-crosslinkable control on information levels lower than PPI-level.**

Each bar represents a separate FDR calculation performed at different information levels while considering all CSMs (naive) or heteromeric crosslinks only (heteromeric). The y-axis displays the observed errors of the respective information level at 5% FDR. Observed error in **a** identified CSMs calculated on redundant or (unique) CSM-level, **b** identified peptide pairs calculated on redundant CSM-level, CSM-level and peptide pair-level, **c** identified residue pairs calculated on redundant CSM-level, CSM-level, peptide pair-level and residue pair-level, respectively. Individual values from DSSO and BS3 datasets are depicted by dots; the dashed line indicates the respective 5% FDR threshold. Source data are provided as a Source Data file.



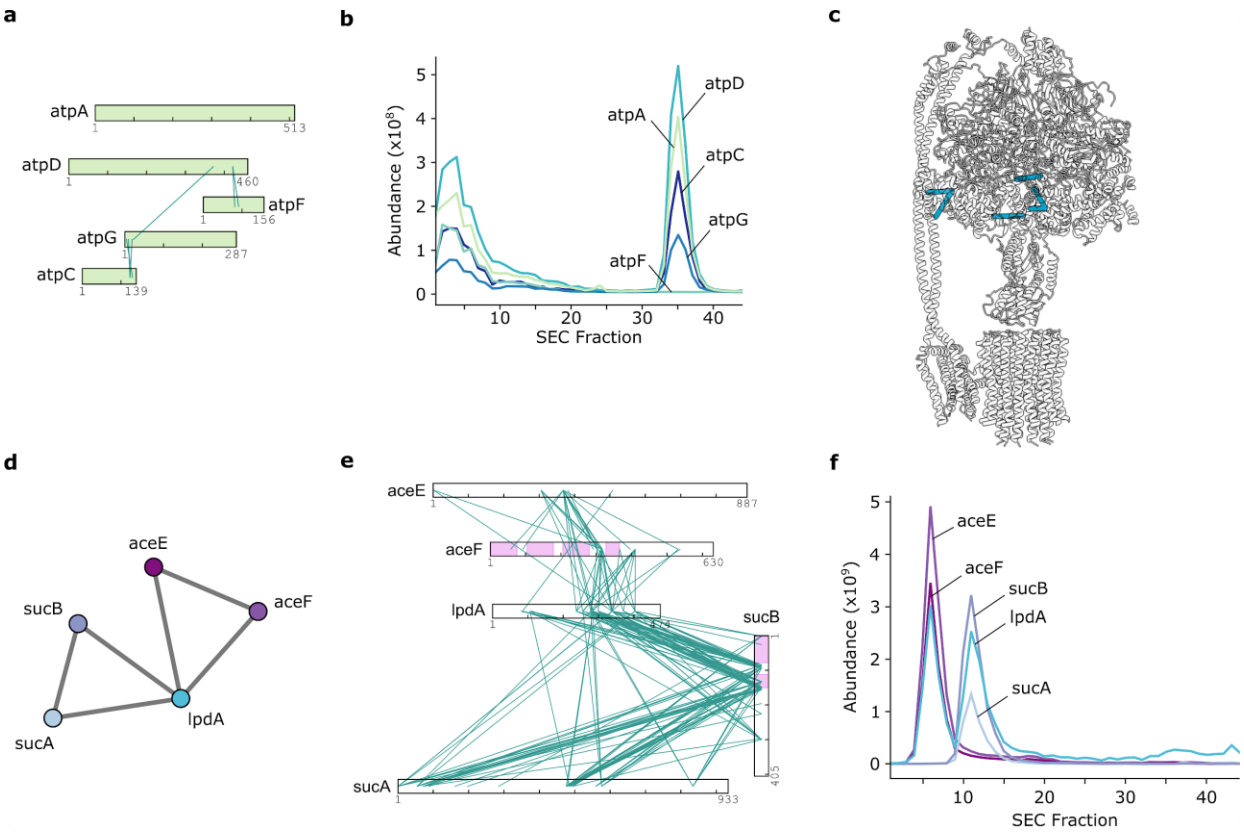
Supplementary Figure 4. Results of positive controls.

a, b Fraction of coeluting PPIs (correlation coefficient > 0.5) among the heteromeric PPIs passing a given FDR threshold, applying different published FDR approaches separated for **a** BS3 and **b** DSSO. **c**, **d** Fraction of PPIs present in the STRING database (STRING combined score >= 150) among the PPIs passing a given FDR threshold, applying different published FDR approaches (Supplementary Table 1) for **c** BS3 and **d** DSSO. Source data are provided as a Source Data file.



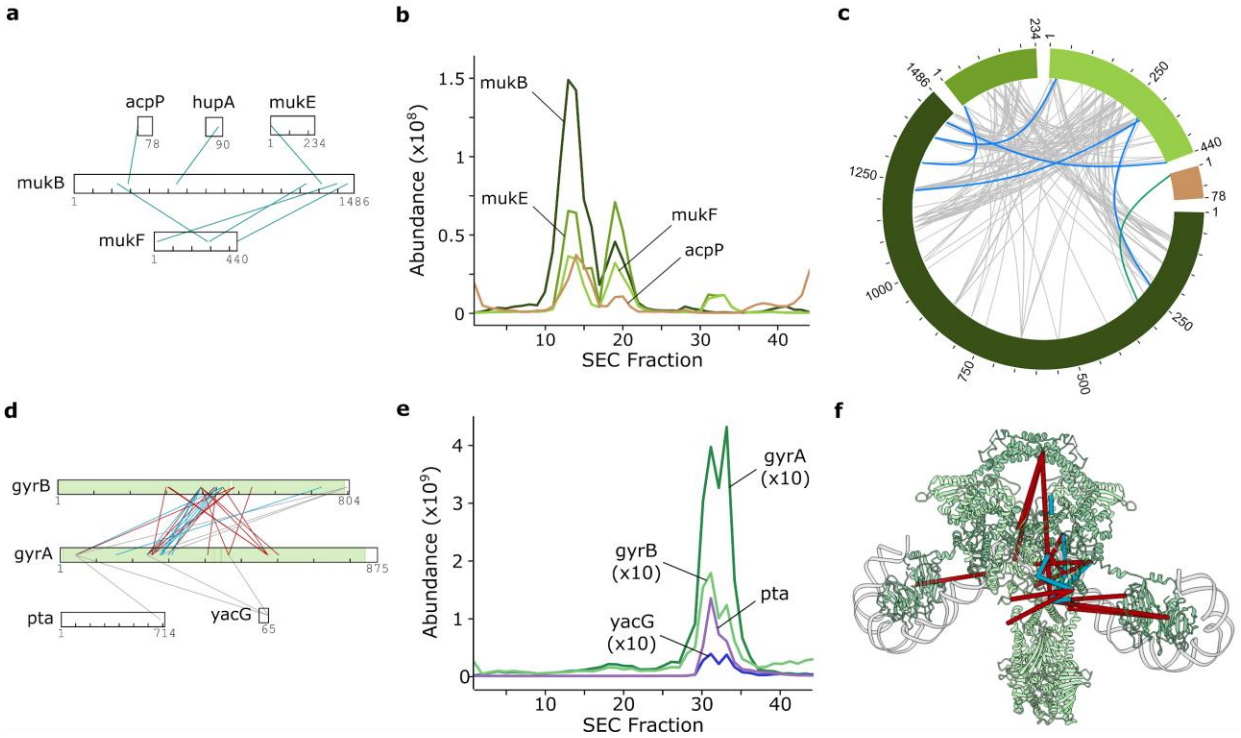
Supplementary Figure 5: Properties of final crosslink PPI network.

a The overall heteromeric PPI counts and the respective fraction of false PPIs are shown for varying FDR thresholds. Expected false PPIs were calculated based on decoy matches. For this dataset increasing FDR thresholds up to 4% add more true than false matches. **b** Distribution of residue pairs per PPI. **c** Abundance of identified crosslinked proteins in the respective categories. Heteromeric-only proteins are significantly more abundant than all identified proteins ($p = 0.044$ using a one-sided Kolmogorov–Smirnov test). **d** Heteromeric-only proteins tend to be smaller than the median of proteins in the database, and are therefore less likely to produce self-links. Boxplots in **c** and **d** depict the median (middle line), upper and lower quartiles (boxes), 1.5 times of the interquartile range (whiskers) as well as outliers (single points). Source data for panels c and d are provided as a Source Data file.



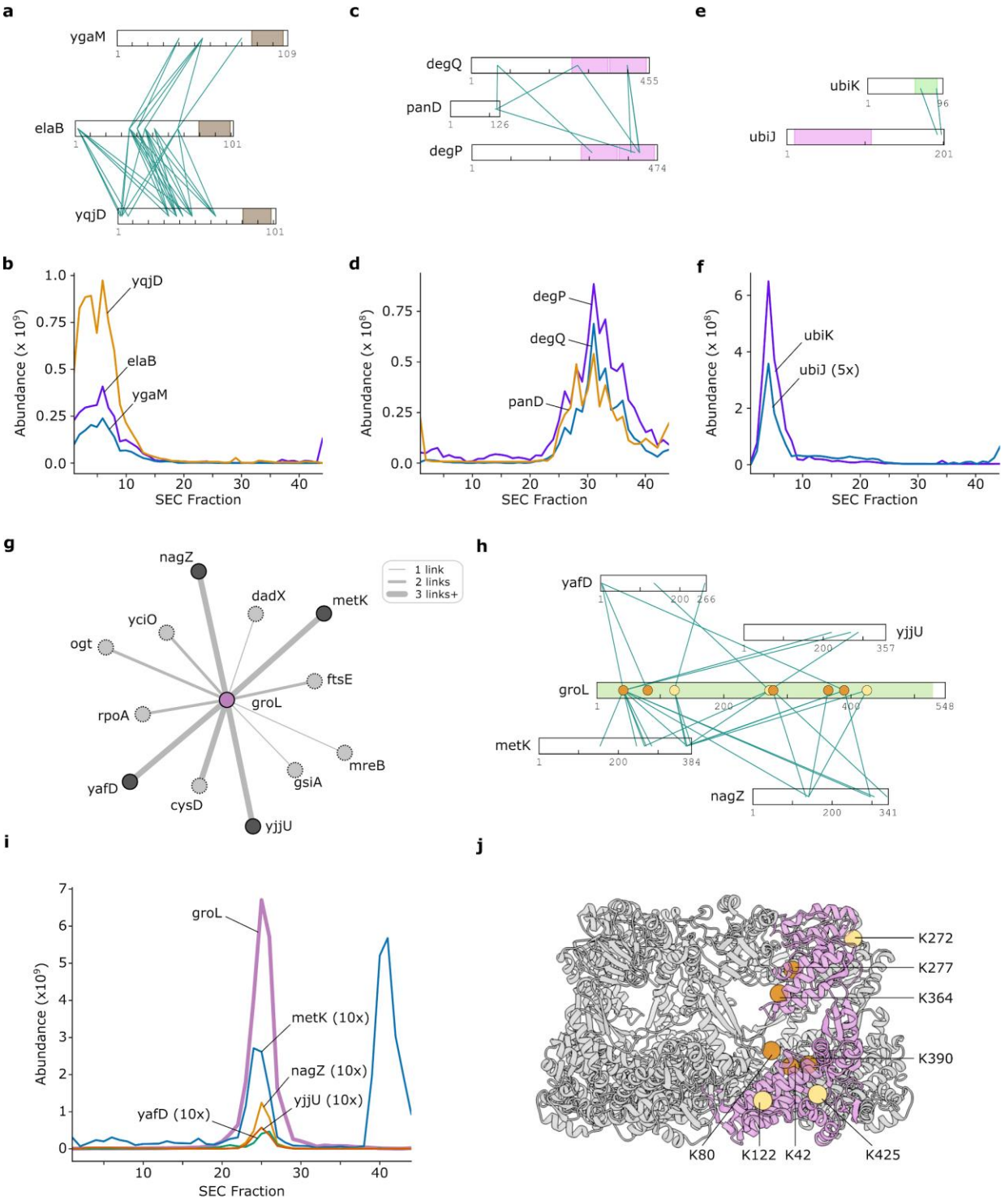
Supplementary Figure 6: ATP synthase and Pyruvate dehydrogenase Crosslinking MS subnetworks.

a PPI subnetwork of ATP synthase in xiNET[13]. Green shade on protein sequences illustrate sequence areas covered by the PDB model (PDB 5T4O)[14]. **b** SEC-Coelution traces of ATP synthase subunits. Note that all subunits also coelute in a very early fraction, probably containing lipid vesicles. **c** Structural model of ATP synthase with mapped heteromeric crosslinks (PDB 5T4O). All protein chains are colored in grey. Heteromeric links are below 35 Å and are colored in blue. **d** PPI subnetwork of pyruvate dehydrogenase and 2-oxoglutarate complexes with collapsed protein nodes and **e** with proteins shown as bars in xiNET[13]. **f** Coelution of pyruvate dehydrogenase / 2-oxoglutarate complex components showing lpdA eluting with both.



Supplementary Figure 7: MukBEF complex and DNA gyrase Crosslinking MS subnetworks.

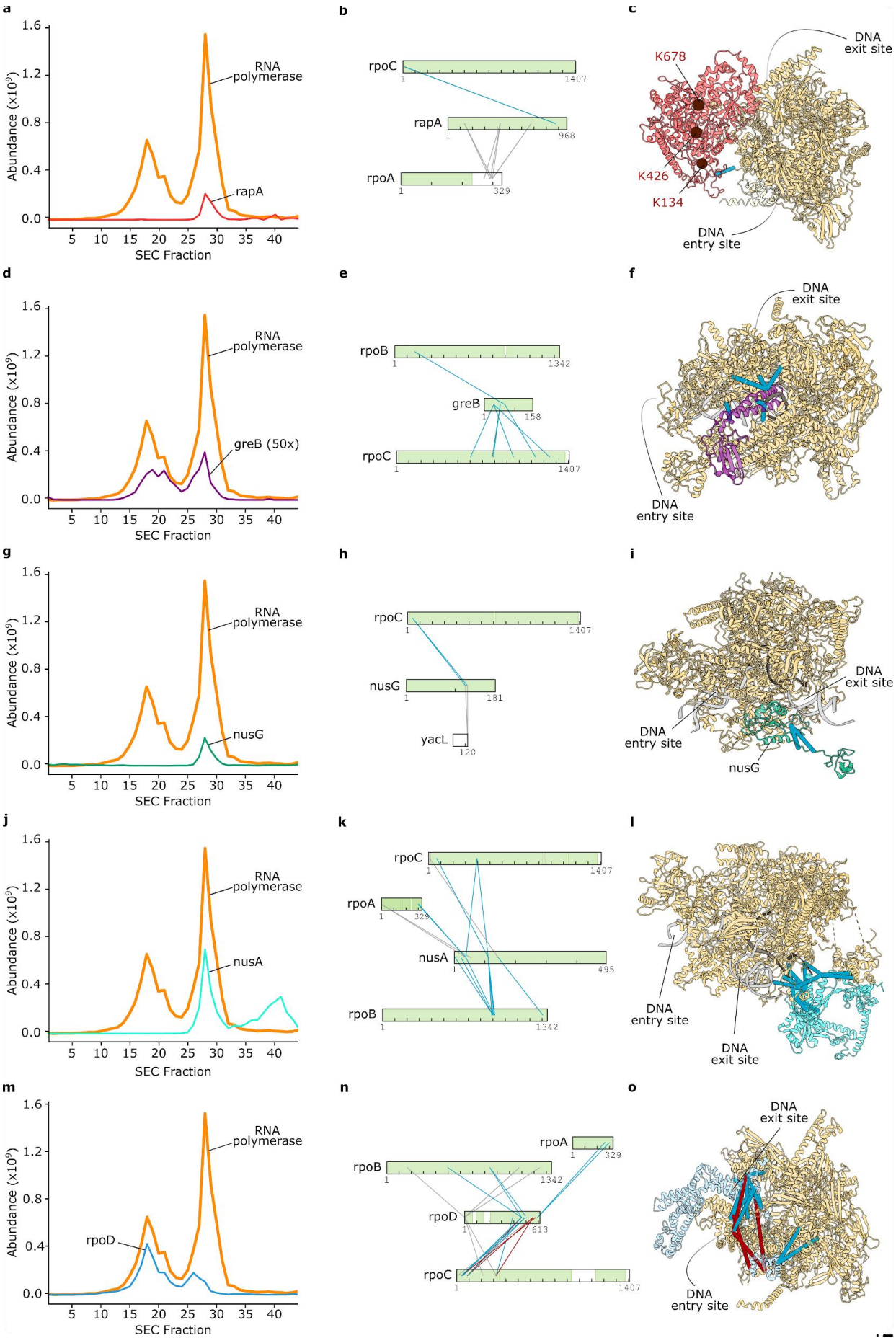
a PPI subnetwork of MukBEF complex with proteins shown as bars in xiNET[13]. **b** Coelution traces of MukBEF complex with its binder AcpP (acpP). The MukBEF complex (fraction 13) dissociated into another MukBEF assembly with lower occupancy for MukB (fraction 19) as judged by coelution. The interactor AcpP is a known binder of MukBEF that could be confirmed by Crosslinking MS and coelution. **c** Crosslink distribution on MukBEF protein sequences compared to a recent *in-vitro* study[15]. Crosslinks are colored in blue if shared between the studies, in grey for *in-vitro* data only and in green if unique to this study. **d** PPI subnetwork of DNA gyrase in xiNET[13]. Green shade on protein sequences illustrate sequence areas covered by the PDB model (PDB 6RKW) [16]. Heteromeric links are colored in blue if below 35 Å, red when exceeding and grey when absent from the model used. **e** Coelution of DNA gyrase complex components with its inhibitor YacG and phosphate acetyltransferase Pta. The abundances for GyrA, GyrB and YacG were magnified 10-fold. **f** Structural model of DNA gyrase with mapped heteromeric crosslinks (PDB 6RKW) [16]. GyrA is shown in dark green and GyrB in light green. DNA is shown in grey. For crosslink coloration see panel **d**. Of note, the DNA gyrase inhibitor YacG was suggested to associate with proteins involved in coenzyme A metabolism such as Pta[17], which is supported by this analysis.



Supplementary Figure 8: Novel identified PPIs.

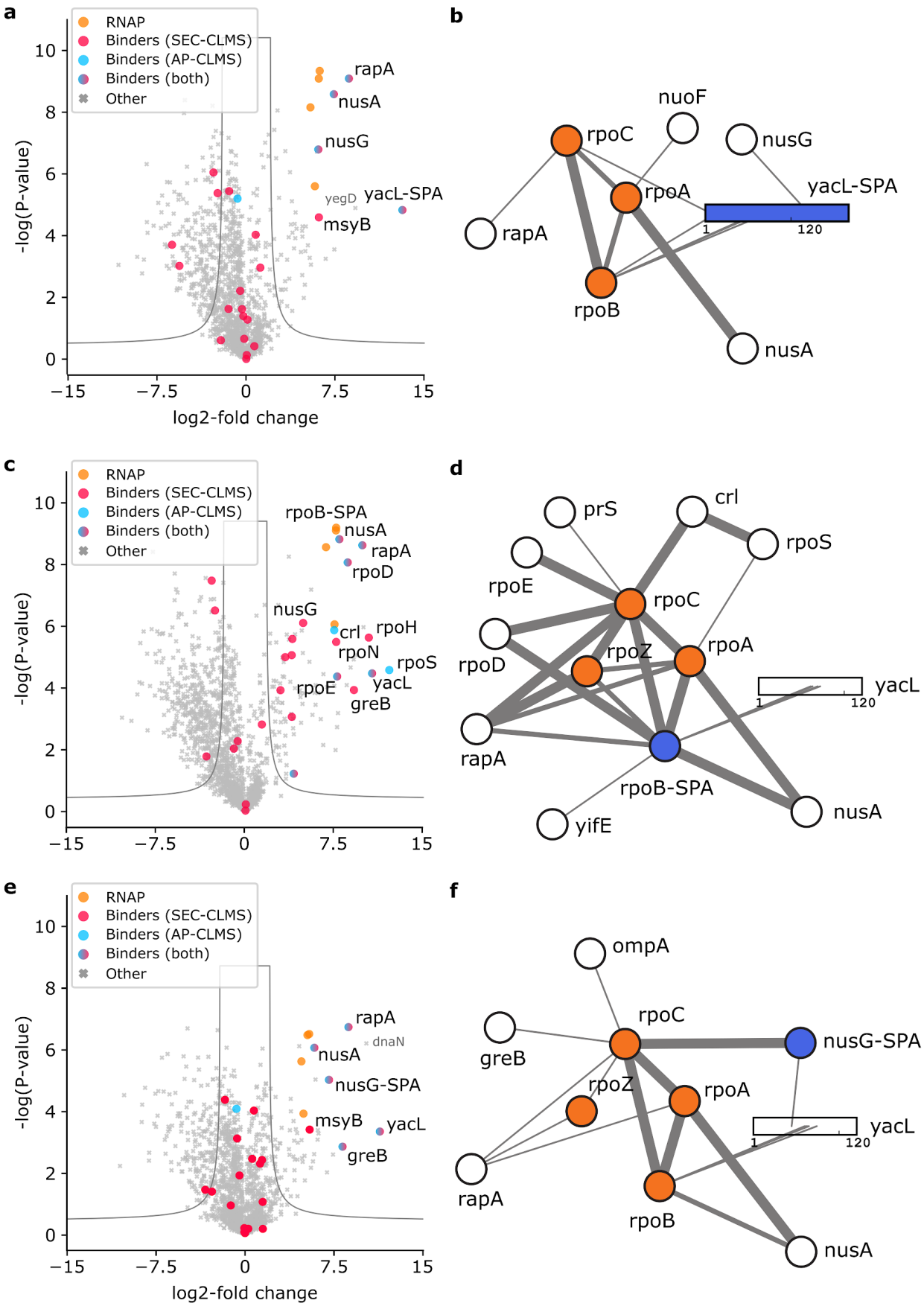
a, c, e, h Selected heteromeric PPI networks with proteins shown as bars or **g** in a network diagram[13] **b, d, f, i** with their corresponding SEC elution traces. Coloration in panels **a, c, e** represent: transmembrane domains for elaB network; PDZ domains for degP and degQ (pink); coiled-coil domain (green) for ubiK and SCP2 domain for ubiJ (pink). The abundance for ubiJ in

panel **f** was magnified 5-fold; abundances of GroEL interactors in panel **i** were magnified 10-fold. High-confidence binders of GroEL, displayed with darker nodes in panel **g**, were selected for the panels **h** & **i**. **h** Selected binders of GroEL shown as bars[13]. Green areas mark GroEL's aligned region with a structural model (PDB 4PKO)[18]. Lysines on the inside of GroEL are highlighted with an orange circle, the ones on the outside with a circle in light yellow. **i** SEC coelution trace of GroEL and selected binders. **j** GroEL structural model (PDB 4PKO) [18] with crosslinked residues highlighted. One ring of the GroEL barrel assembly is shown with one subunit of GroEL colored in pink. Crosslinked lysine residues are indicated as spheres and colored as in panel **h**.


a
Abundance (x10⁹)
SEC Fraction
RNA polymerase
rapA
b
rpoC
rapA
rpoA
c
DNA exit site
K678
K426
K134
DNA entry site
d
greB (50x)
e
rpoB
greB
rpoC
f
g
nusG
h
yacL
i
j
nusA
k
l
m
rpoD
n
o

Supplementary Figure 9: Selected RNA polymerase binders.

a, d, g, j, m Collected experimental support for heteromeric PPIs between RNAP and selected binders - rapA, greB, nusG, nusA and rpoD - is shown from SEC elution profiles, **b, e, h, k, n** Crosslinking MS-based PPI screening in xiNET[13] and **c, f, i, l, o** matching to existing structural PPI models. The abundance of greB was magnified 50-fold. Green shade on protein sequence bars illustrate areas covered by PDB models that were used to map heteromeric links onto structures. In protein crosslink diagrams and protein structural models, heteromeric crosslinks satisfying euclidean distance thresholds of 35 Å are shown in blue while longer links are colored in red. The structural models used were: rapA (PDB 4S20)[19], greB (PDB 6RIN)[20], nusG (PDB 5MS0)[21], nusA (PDB 6FLQ)[22] and rpoD (PDB 4ZH3)[23]. If present in the model, DNA is colored in grey and RNA in black. Labeled spheres in panel **c** mark lysine residues crosslinked to C-termini of RNAP's α-subunits (absent from this model).


a
RNAP
Binders (SEC-CLMS)
Binders (AP-CLMS)
Binders (both)
Other
-log(P-value)
log2-fold change
rapA
nusA
nusG
yegD
yacL-SPA
msyB
b
nuoF
rpoC
nusG
rpoA
yacL-SPA
rapA
rpoB
nusA
c
rpoB-SPA
nusA
rapA
rpoD
nusG
crl
rpoH
rpoN
rpoS
rpoE
yacL
greB
d
prS
crl
rpoS
rpoE
rpoC
rpoD
rpoZ
rpoA
yacL
rapA
rpoB-SPA
yifE
nusA
e
rapA
dnaN
nusA
nusG-SPA
msyB
yacL
greB
f
ompA
rpoC
greB
nusG-SPA
rpoZ
rpoA
yacL
rapA
rpoB
nusA

Supplementary Figure 10: AP-MS and sulfo-SDA crosslinking.

Enrichment analysis from affinity-enriched **a** YacL-SPA, **c** RpoB-SPA and **e** NusG-SPA with **b, d, f** corresponding crosslink subnetworks from sulfo-SDA crosslinked eluates from these enrichments. In the volcano plots, proteins of interest are labelled as follows: components of RNA polymerase in orange; proteins found crosslinked to RNA polymerase from lysate SEC fractionation in red; proteins found crosslinked to RNA polymerase from a given affinity-enrichment experiment in blue; overlapping proteins between the two sample workflows in red and blue; all other proteins are indicated as grey crosses. In the crosslink subnetworks generated with xiNET[13], RNA polymerase constituents are labelled in orange and SPA-tagged proteins are highlighted in blue. YacL is shown in expanded sequence view. The thickness of the lines represents the number of unique residue pairs between the proteins.

Supplementary References

1. Gonzalez-Lozano, M. A. *et al.* Stitching the synapse: Cross-linking mass spectrometry into resolving synaptic protein interactions. *Sci Adv* **6**, eaax5783 (2020).
2. Bartolec, T. K. et al. Cross-linking Mass Spectrometry Analysis of the Yeast Nucleus Reveals Extensive Protein-Protein Interactions Not Detected by Systematic Two-Hybrid or Affinity Purification-Mass Spectrometry. Anal. Chem. 92, 1874–1882 (2020).
3. Linden, A. *et al.* A cross-linking mass spectrometry approach defines protein interactions in yeast mitochondria. *Mol. Cell. Proteomics* (2020) doi:10.1074/mcp.RA120.002028.
4. 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).
5. 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).
6. Chavez, J. D., Keller, A., Zhou, B., Tian, R. & Bruce, J. E. Cellular Interactome Dynamics during Paclitaxel Treatment. *Cell Rep.* **29**, 2371–2383.e5 (2019).
7. Yugandhar, K. *et al.* MaXLinker: Proteome-wide Cross-link Identifications with High Specificity and Sensitivity. *Mol. Cell. Proteomics* **19**, 554–568 (2020).
8. Steigenberger, B., Pieters, R. J., Heck, A. J. R. & Scheltema, R. A. PhoX: An IMAC-Enrichable Cross-Linking Reagent. *ACS Cent Sci* **5**, 1514–1522 (2019).
9. Makepeace, K. A. T. *et al.* Improving Identification of In-organello Protein-Protein Interactions Using an Affinity-enrichable, Isotopically Coded, and Mass Spectrometry-cleavable Chemical Crosslinker. *Mol. Cell. Proteomics* **19**, 624–639 (2020).
10. Mendes, M. L. *et al.* An integrated workflow for crosslinking mass spectrometry. *Mol. Syst. Biol.* **15**, e8994 (2019).
11. Ryl, P. S. J. et al. In Situ Structural Restraints from Crosslinking Mass Spectrometry in

Human Mitochondria. J. Proteome Res. (2019) doi:10.1021/acs.jproteome.9b00541.

12. O'Reilly, F. J. *et al.* In-cell architecture of an actively transcribing-translating expressome. *Science* **369**, 554–557 (2020).
13. Combe, C. W., Fischer, L. & Rappsilber, J. xiNET: cross-link network maps with residue resolution. *Mol. Cell. Proteomics* 14, 1137–1147 (2015).
14. Sobti, M. *et al.* Cryo-EM structures of the autoinhibited E. coli ATP synthase in three rotational states. *Elife* **5**, (2016).
15. Bürmann, F. *et al*. A folded conformation of MukBEF and cohesin. *Nat. Struct. Mol. Biol.* **26**, 227–236 (2019).
16. Vanden Broeck, A., Lotz, C., Ortiz, J. & Lamour, V. Cryo-EM structure of the complete E. coli DNA gyrase nucleoprotein complex. *Nat. Commun.* **10**, 4935 (2019).
17. Vos, S. M. *et al*. Direct control of type IIA topoisomerase activity by a chromosomally encoded regulatory protein. *Genes Dev*. **28**, 1485-1497 (2014).
18. Fei, X., Ye, X., LaRonde, N. A. & Lorimer, G. H. Formation and structures of GroEL:GroES2 chaperonin footballs, the protein-folding functional form. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 12775–12780 (2014).
19. Liu, B., Zuo, Y. & Steitz, T. A. Structural basis for transcription reactivation by RapA. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 2006–2010 (2015).Asda
20. Abdelkareem, M. 'men *et al.* Structural Basis of Transcription: RNA Polymerase Backtracking and Its Reactivation. *Mol. Cell* **75**, 298–309.e4 (2019).
21. Said, N. *et al.* Structural basis for λN-dependent processive transcription antitermination. *Nat Microbiol* **2**, 17062 (2017).
22. Guo, X. *et al.* Structural Basis for NusA Stabilized Transcriptional Pausing. *Mol. Cell* **69**, 816–827.e4 (2018).
23. Feng, Y. *et al.* Structural Basis of Transcription Inhibition by CBR Hydroxamidines and CBR Pyrazoles. *Structure* **23**, 1470–1481 (2015).