TECHNISCHE UNIVERSITÄT BERLIN

DOCTORAL THESIS

Computational Methods and Machine Learning for Crosslinking Mass Spectrometry Data Analysis

Vorgelegt von:
Sven Hans-Joachim GIESE
0000-0002-9886-2447

Promotionsausschuss:
Gutachter:
Prof. Dr. Juri RAPPSILBER
Prof. Dr. Matthias SELBACH

Vorsitzende:
Prof. Dr. Vera MEYER

an der
Fakultät III - Prozesswissenschaften
der
Technischen Universität Berlin
zur Erlangung des akademischen Grades

doctor rerum naturalium
– Dr. rer. nat. –

genehmigte Dissertation

Tag der wissenschaftlichen Aussprache: 18. September 2020

Berlin 2021

Declaration of Authorship

I, Sven Hans-Joachim GIESE, declare that this thesis titled, "Computational Methods and Machine Learning for Crosslinking Mass Spectrometry Data Analysis" 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. With the exception of 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.

Signed:

Date:

TECHNISCHE UNIVERSITÄT BERLIN

Abstract

Fakultät III - Prozesswissenschaften
Institute of Biotechnology

doctor rerum naturalium
– Dr. rer. nat. –

Computational Methods and Machine Learning for Crosslinking Mass Spectrometry Data Analysis

by Sven Hans-Joachim GIESE

A central part in understanding complex biological systems is to uncover the function and structure of proteins. The elucidation of a protein's structure and understanding its function are tightly connected. The underlying paradigm that **structure defines function**, has led to the development of many methods to derive the three-dimensional structure of proteins and protein complexes. Crosslinking mass spectrometry (CLMS) is a comparatively new tool for the analysis of single proteins, multi-protein complexes, and protein-protein interactions. CLMS poses several challenges for mass spectrometry-based proteomics, which include understanding the fragmentation behavior of crosslinked peptides to design efficient database search strategies and improved acquisition settings.

CLMS builds upon the preservation of distance information by crosslinking reagents, which is relayed by mass spectrometric analysis. To identify a crosslink in a standard database search, theoretically all pairwise peptide combinations need to be considered. Without the use of isotope-labeled or cleavable crosslinkers, applying a standard crosslinking approach using homobifunctional NHS-ester crosslinker reagents, an exhaustive peptide identification strategy becomes quickly unfeasible because of the dynamic explosion of the search space. Therefore, robust heuristics are needed to make the identification of crosslinks in complex samples feasible. This endeavor is even further hindered by the unequal fragmentation of the two peptides in a crosslink under collision-induced dissociation conditions. The subsequent coverage gap between the two peptides in a crosslink may lead to misidentifications. This thesis presents computational approaches and machine learning methods to improve the identification of crosslinked peptides.

First, an efficient strategy is outlined, based on an explorative study about the fragmentation behavior of crosslinked peptides. Most importantly, the presented search strategy shows that the information from isotope-labeled and cleavable crosslinkers can be partially retrieved by computational processing of the spectra and adequate mass spectrometric acquisition settings. A key concept builds upon the ability to recognize crosslinked fragments from their mass and charge. This allows to identify the two linked peptides in a sequential manner without searching all peptide combinations exhaustively.

Second, to reduce the coverage gap, modern mass spectrometers offer versatile fragmentation methods. For most crosslinks, electron-transfer dissociation combined with higher-energy collision dissociation (HCD) yields the highest coverage.

HCD remains an important choice because of its fast acquisition speed and competitive sequence coverage.

Third, to avoid severe bias through the identification of noncovalently associated peptides as crosslinks, multiple solutions are feasible. For example, disruptive ionization settings can be used to avoid noncovalently associated peptides entering the mass spectrometer. Alternatively, post-acquisition heuristics using the retention time difference between linear and crosslinked peptides add valuable information to recognize noncovalent peptide associations.

Fourth, since complex crosslinking experiments with deep-proteome coverage require extensive fractionation, being able to predict the retention behavior may prove beneficial for peptide identification. In addition, mechanistic understanding of the separation process helps to further improve the chromatographic separation. For hydrophilic anion exchange chromatography (hSAX), the separation is heavily influenced by charged amino acids and aromatics. Most importantly, the retention behavior of linear peptides can be accurately predicted through deep neural networks.

Fifth, the ability to predict not only hSAX, but also strong cation exchange (SCX) and reversed-phase retention times indeed proves to be a valuable addition for the identification of crosslinked peptides. Siamese neural network architectures offer elegant solutions to encode crosslinked peptides. Multi-task learning of several chromatography domains at the same time allows robust and fast prediction of all chromatography domains. Accurate reversed-phase predictions together with hSAX and SCX fraction prediction allows rescoring already identified peptide spectrum matches with a support vector machine. This workflow leads to more identified protein-protein interactions at constant false discovery rate from a deep-fractionated *Escherichia coli* sample.

The integration of advancements in crosslinking chemistry, sample acquisition, database search, and machine learning together are essential stepping-stones for the identification of crosslinked peptides in complex samples.

TECHNISCHE UNIVERSITÄT BERLIN

Zusammenfassung

Fakultät III - Prozesswissenschaften
Institute of Biotechnology

doctor rerum naturalium
– Dr. rer. nat. –

Computational Methods and Machine Learning for Crosslinking Mass Spectrometry Data Analysis

von Sven Hans-Joachim GIESE

Ein zentraler Bestandteil zum Verständnis komplexer biologischer Systeme ist die Aufdeckung der Funktion und Struktur von Proteinen. Die Aufklärung der Struktur eines Proteins und das Verständnis seiner Funktion sind eng miteinander verbunden. Das zugrunde liegende Paradigma, **Struktur definiert Funktion**, hat zur Entwicklung vieler Methoden zur Bestimmung der dreidimensionalen Struktur von Proteinen und Proteinkomplexen geführt. Die quervernetzende Massenspektrometrie (CLMS) ist ein vergleichbar neues Werkzeug für die Analyse von einzelnen Proteinen, Multiproteinkomplexen und Protein-Protein-Interaktionen. CLMS stellt die Massenspektrometrie-basierte Proteomik vor mehrere Herausforderungen, darunter das Verständnis des Fragmentierungsverhaltens von quervernetzten Peptiden, um effiziente Datenbank-Suchstrategien und verbesserte instrumentelle Aufnahmestrategien zu entwerfen.

CLMS baut auf der Erhaltung von Abstandsinformationen durch die massenspektrometrische Analyse unter Verwendung von Quervernetzungsreagenzien auf. Beim universellen Ansatz, d.h. ohne die Verwendung isotopenmarkierter oder spaltbarer Quervernetzer, wird eine erschöpfende Peptididentifikationsstrategie aufgrund der dynamischen Explosion des Suchraums schnell undurchführbar. Daher sind robuste Heuristiken erforderlich, um die Identifizierung von Quervernetzungen in komplexen Proben durchführbar zu machen. Dieses Bestreben wird durch die ungleiche Fragmentierung der beiden Peptide in einer Quervernetzung unter kollisionsinduzierter Dissoziation behindert. Mit der daraus resultierenden Sequenzabdeckungslücke zwischen den beiden Peptiden in einer Quervernetzung kann es zu einer Fehlidentifizierung kommen oder sogar zu einer starken Verzerrung der Identifikationsergebnisse. In dieser Arbeit werden rechnergestützte Ansätze und Methoden des maschinellen Lernens vorgestellt, um die Identifizierung von quervernetzten Peptiden zu verbessern.

Zuerst wird eine effiziente Suchstrategie skizziert, basierend auf einer explorativen Studie über das Fragmentierungsverhalten von quervernetzten Peptiden. Die vorgestellte Suchstrategie zeigt, dass die Informationen von isotopenmarkierten und spaltbaren Quervernetzern teilweise durch rechnerische Verarbeitung der Spektren und geeignete massenspektrometrische Aufnahmeeinstellungen ersetzt werden können. Ein Schlüsselkonzept basiert auf der Fähigkeit, quervernetzte Fragmente anhand ihrer Masse und Ladung zu erkennen, um die beiden quervernetzten Peptide sequenziell zu identifizieren.

Zweitens bieten moderne Massenspektrometer vielseitige Fragmentierungsmethoden an, um die Sequenzabdeckung zu erhöhen. Für die meisten Quervernetzungen liefert die Elektronentransferdissoziation in Kombination mit der hochenergetischen Kollisionsdissoziation (HCD) die höchste Sequenzabdeckung. Die HCD bleibt wegen der schnellen Aufnahmegeschwindigkeit und der kompetitiven Sequenzabdeckung eine wichtige Option.

Drittens ist es möglich, schwere Verzerrungen durch die Identifizierung von nichtkovalent assoziierten Peptiden als quervernetzte Peptide zu vermeiden. Zum Beispiel können disruptive Ionisierungseinstellungen verwendet werden, um nichtkovalent assoziierte Peptide daran zu hindern in das Massenspektrometer zu gelangen. Alternativ liefern Heuristiken nach der Akquisition unter Verwendung der Retentionszeitdifferenz von linearen und quervernetzten Peptiden wertvolle Informationen hinzu, um nicht-kovalent assoziierte Peptide zu erkennen.

Viertens, da komplexe Quervernetzungsexperimente mit tiefer Proteomabdeckung eine umfangreiche Fraktionierung erfordern, kann sich die Vorhersage des Retentionsverhaltens als vorteilhaft für die Peptididentifizierung erweisen. Darüber hinaus hilft das mechanistische Verständnis des Trennprozesses, die chromatographische Trennung weiter zu verbessern. Bei der hydrophilen Anionenaustauschchromatographie (hSAX) wird die Trennung stark durch geladene Aminosäuren und Aromaten beeinflusst. Am wichtigsten ist, dass das Retentionsverhalten von linearen Peptiden durch tiefe neuronale Netzwerke genau vorhergesagt werden kann.

Fünftens erweist sich die Fähigkeit, nicht nur hSAX, sondern auch den starken Kationenaustausch (SCX) und die Retentionszeiten in Umkehr-Phase Chromatographie vorherzusagen, in der Tat als wertvolle Ergänzung für die Identifizierung quervernetzter Peptide. Siamesische neuronale Netzwerke bieten elegante Lösungen zur Kodierung quervernetzter Peptide. Das Multi-Task-Lernen mehrerer Chromatographie-Domänen zur gleichen Zeit ermöglicht eine robuste und schnelle Vorhersage. Genaue Umkehrphasenvorhersagen zusammen mit hSAX- und SCX- Fraktionsvorhersagen erlauben es, bereits identifizierte Peptidspektrum-Identifikationen mit einer *Support-Vektor-Maschine* neu zu bewerten. Dabei können die identifizierten Protein-Protein-Interaktionen von einer tief-fraktionierten *Escherichia coli* Probe um das Zweifache erhöht werden bei konstanter Falschfindungsrate.

Fortschritte in der Quervernetzungschemie, der Probenvorbereitung, der Datenbanksuche und dem maschinellen Lernen zusammen bilden wesentliche Sprungbretter für die Identifizierung von quervernetzten Peptiden in komplexen Proben.

Acknowledgements

First and foremost, I want to thank my beloved family for their ongoing support. There are no proper words to express my deep gratitude towards my mother who has always supported and encouraged me. My brother's dedication to his family is unique and I am glad that I could always count on him when I needed help. In the last years, my family has also grown a bit and I would not be the person I am, without my wife Katja and my daughter Elisa. Katja has probably suffered as much as I have during the last month of my thesis and I want to apologize and thank her for coping with everything so patiently. She was unlucky enough to listen to my whining and (maybe sometimes) unstructured thoughts when I needed her expertise.

"Science is magic that works" is the first google autocomplete entry when one types "Science is" into the search field (2020, April, 25th, 0:09 am). While I am not a magician (unfortunately), magic or magic tricks require years of training, dedicated preparation for the big blast, and a team that supports you in case the lock is stuck while you are in a water tank. Luckily, my supervisor Juri Rappsilber knows that these ingredients are very vital for a successful PhD. Together with a jump into the *unknown* scientists can do what they are best at: being creative and solving problems. For living this philosophy, I am very grateful that Juri and I have come together. I am very thankful for the continued support and freedom Juri gave me to follow-up on own project ideas. In addition, I would like to emphasize that I enjoyed the last 6 years a lot. Building an entire department at a university is a daunting task. Seeing your hard work flourish and being able to contribute so significantly towards our teaching and research activities makes me proud. When I started to jump into teaching, I was very lucky to have Christian at my side. Thanks for taking the countless hours for slide polishing, planning and optimizing our classes. Doing excellent research *and* excellent teaching is nearly impossible on your own. I am very thankful to all the contributions of our excellent tutors and extended teaching team over the years (alphabetical order): Anne, Benni, Edward, Eva, Fabi, Franzi, Gerrit, Henning, Iva, Jaqueline, Jessica, Julia, Kord, Leon, Lisa, Martina, My, Renate and Tarek.

Joining Juri's group also meant to make a *brief* detour to Scotland. While this move was certainly a challenge, it was a great opportunity. I was very lucky to experience such a warm (well... it is still Scotland, right?) and friendly welcome from the group members which were then in Edinburgh. Special thanks go to Lutz, Angel, and Colin who took me into their homes and provided me with shelter and food until I had my own place. In addition, Lutz deserves another sentence, just dedicated to him: thanks for helping me get started in the group and your answers to all kinds of stupid questions. With the move to Berlin and the group growing, also a lot of additional expertise came into the group: thanks to Ludwig, Petra, and Fränze for sharing it! Thanks to Adam and Ludwig for all the samples they prepared which turned out to be the basis for many chapters in this thesis. Lastly, thanks to Swantje for her (mostly :P) positive attitude and sharing the most entertaining reviewer experiences.

Even though this thesis bears my name, I am very thankful for the many (direct or indirect) contributions from the Rappsilber lab.

— Thanks.

Contents

List of Abbreviations

acc	Accuracy
APMS	Affinity-Purification Mass Spectrometry
BS3	Bis(sulfosuccinimi-dyl)suberate
CL	Crosslinked
CLMS	Crosslinking Mass Spectrometry
CSM	Crosslinked Peptide Spectrum Match
CV	Crossvalidation
D	Decoys
DDA	Data-Dependent Acquisition
DIA	Data-Independent Acquisition
EM	Electron Microscopy
ETD	Electron-Transfer Dissociation
FDR	False Discovery Rate
FN	False Negative
FP	Flase Positive
HCD	Higher-Energy Collision Dissociation
HSA	Human Serum Albumin
hSAX	Hydrophilic Anion Exchange Chromatography
ISCID	In-Source Collision-Induced Dissociation
LC	Liquid-Chromatography
LN	Linear
NAP	Noncovalent Peptide Associations
NCE	Normalized Collision Energy
NMR	Nuclear Magnetic Resonance Spectroscopy
PPI	Protein-Protein Interaction
racc	Relaxed Accuracy
RDP	Ranked Dot Product
RP	Reversed-Phase
RT	Retention Time
SCX	Strong Cation Exchange
SEC	Size Exclusion Chromatography
sulfo-SDA	Sulfosuccinimidyl 4,4′-azipentanoate
SVM	Support Vector Machine
T	Targets
TN	True Negative
TP	True Positive
Y2H	Yeast Two-Hybrid

Chapter 1

Introduction

The cell. A fundamental goal in molecular biology is to understand the complex processes that govern life in a functional cell. The central biomolecules that are involved in virtually any life-dependent process are proteins (Keskin, Tuncbag, and Gursoy, 2016). Only by the manifold interactions between proteins, cells can harvest energy, produce metabolites, fight diseases, and reproduce. Therefore, the study of the proteome and its rich interaction network will eventually enable modern life science to intervene and optimize the cell's response to the desired outcome.

Protein function. Proteins can function independently, in signal cascades, or even in large molecular machines. The functional landscape that makes these interactions possible is diverse. Based on the chemical properties of the involved proteins, the nature of the interaction can be transient or permanent (Nooren, 2003). Transient interactions have only short half-lives, while permanent interactions are irreversible. This makes transient protein complexes, prime regulators in signal cascades (Acuner Ozbabacan et al., 2011). Unfortunately, transient interactions are challenging to study because of their unstable nature. Therefore, the methods that are required to study both types of interactions are also diverse. But why do we strive to study the structure of proteins and protein complexes at all? A common paradigm that architects and biologists try to exploit is "structure determines function". Thus, knowing the structure of a protein will hopefully help us to understand its function. Determining the structure of a protein is often achieved by assuming that they have a single structural confirmation. To capture the structure of a protein, it is important to realize that there are many copies of the same protein in the cell – and even across different cells. Depending on the environment and the interactions the conformational states of a single protein vary (Miao and Cao, 2016). Examples include GTP-binding proteins, motor proteins, carrier proteins (Chen, Kurgan, and Ruan, 2007). In addition, protein structures can have rigid and flexible parts (e.g. unordered regions) that are vital for protein function (Craveur et al., 2015).

Structure determination. Capturing the structural information from peptides to protein-protein interactions (PPIs) is difficult with a single technique alone. A plethora of methods have been developed that each have their strengths and weaknesses. Traditionally, X-ray crystallography is the most used technique to deliver structural information for proteins and protein complexes (Sali et al., 2003). X-ray crystallography is a very well-developed method that delivers high-resolution structures if the targeted protein complex crystallizes. Even if the protein crystalizes, the question remains whether the adopted protein structure represents the physiological structure. This caveat makes X-ray crystallography less suited to study dynamic PPIs (in vivo). Nuclear magnetic resonance spectroscopy (NMR) and single-particle electron microscopy (cryo-EM) on the other hand are well suited to study dynamic protein interactions. Cryo-EM and the recent breakthroughs in sample preparation,

camera technology, and computational processing have led to high-resolution structures near to atomic resolution of 3 Å to 5 Å (Cheng, 2015). NMR, another technique, is expanding its application towards in-cell analysis under physiological conditions (Ikeya, Güntert, and Ito, 2019) and atomic resolution (Luchinat and Banci, 2017). For more details on the field of structural biology and the study of PPIs the reader is referred to the following publications: (Leitner, 2016; Liu and Hsu, 2005; Miura, 2018; Sali et al., 2003; Smits and Vermeulen, 2016).

Interactome screening. Despite the success and many applications of the above-described methods in elucidating protein structures, they do not provide PPI information on a system-wide scale. A selection of current high-throughput methods for elucidating PPI information include: Affinity-purification mass spectrometry (APMS), Yeast Two-Hybrid (Y2H) screens, and crosslinking mass spectrometry (CLMS). Depending on the analyzed species and technology, the magnitude in coverage ranges from several thousand to tens of thousands of discovered PPIs (Mehta and Trinkle-Mulcahy, 2016). A few years ago, CLMS would not have been considered in the context of large PPI screening experiments. In this thesis I show the improvements and contributions to the data analysis of crosslinking mass spectrometry that contributed to the maturation of the field from the analysis of single proteins to interactome studies.

1.1 Crosslinking Mass Spectrometry

CLMS. The design of CLMS experiments is very similar to standard proteomic workflows (Steen and Mann, 2004). A minimal CLMS experiment needs to perform cell lysis, crosslinking, protein digestion, LCMS, and data analysis (Rappsilber, 2011; Sinz, 2018). The introduced crosslinker covalently binds proximal amino acids in a single protein or between multiple proteins. This bond remains intact through the entire sample preparation and MS analysis. Thus, the identified crosslink spectrum matches (CSMs) represent peptide pairs that were spatially close during the crosslinking reaction. Depending on the crosslinking chemistry used the respective distance constraints vary. A common crosslinker, that is used in the universal approach (O'Reilly and Rappsilber, 2018) is bis(sulfosuccinimi-dyl)suberate (BS3). Universally applicable crosslinkers consist of two functional groups and a spacer. The theoretical distance constraints are typically derived by adding the spacer length, the length of the reactive amino acid's side chain, and a spatial tolerance parameter leading to approximately 27.4 Å for BS3 (Chen et al., 2010). This cutoff is supported by molecular dynamic simulations, recommending 26 Å to 30 Å as a sensible cutoff region (Merkley et al., 2014).

Relevance. CLMS is applicable to purified proteins (Belsom et al., 2016), protein complexes (Kao et al., 2012; Walzthoeni et al., 2013), organelles (Liu et al., 2018; Ryl et al., 2020) and entire proteomes (Götze et al., 2019; Liu et al., 2015). However, the strength of CLMS lies within the possibility to analyze protein structures in their native environment. This benefit comes with the drawback of medium-resolution information compared to X-ray crystallography, NMR, or EM. For the development of CLMS the availability of 3D-models from established methods is a blessing and a curse; on the one hand available structures make it easy to validate identified crosslinks, on the other hand discrepancies between 3D-models and crosslinks are difficult to interpret (Chu, Thornton, and Nguyen, 2018). They could either be the

result of different conformational states that were only captured by CLMS, or a 3D-model that does not reflect the native structure, or a misidentified crosslink. Nevertheless, available high-resolution 3D-models for proteins are often used to evaluate self and heteromeric crosslinks in studies of single proteins or complex samples (Mendes et al., 2019). A very active area of research is now elegantly combining the information from the different structure determination methods into integrated structural biology approaches (Cerofolini et al., 2019; Robinson et al., 2015; Schmidt and Urlaub, 2017). Of particular interest is the combination of cryo-EM and CLMS where the distance constraints together with density maps seem to complement each other in computational modeling approaches (Steigenberger et al., 2020; O'Reilly et al., 2020).

Peptide identification. Before distance constraints can be used in structural modeling, the acquired mass spectra need to be matched to peptide pairs. This process is usually done via a database search (Yılmaz et al., 2018) or spectral libraries for data-independent acquisition (Müller et al., 2019). The database search poses a major challenge for the universal crosslinking approach. Especially, the rapid increase in search space commonly referred to as the *n-square* problem. The target database of all possible crosslinks grows quadratically with the number of peptides. The rapid explosion of the search space from combining two peptides from 50 proteins, to a crosslink, reaches the search space size of a standard linear search (including roughly 20000 proteins) (Yılmaz et al., 2018). This effect only gets amplified with the use of promiscuous crosslinker chemistry (Belsom et al., 2016) or sequential digestion (Mendes et al., 2019). The exhaustive construction of the target database often becomes unfeasible for complex samples. Therefore, many search engines employ a heuristic two-pass strategy which tries to identify both peptides individually (Yılmaz et al., 2018). However, various implementations of the two-step identification process are found in the literature. Improvements for the data analysis in the universal crosslinking approach are of central importance for the CLMS field and the structural biology community, since approximately 78% of the studies between 2009 and 2019 use non-cleavable crosslinkers (Steigenberger et al., 2020).

Search heuristics. In Kojak (Hoopmann et al., 2015) the implemented two-pass strategy is similar to an open-modification search strategy (Joice et al., 2014). Linear peptides are searched with modifications on the crosslinkable residues to compute a list of high scoring candidates. In a second step, these candidates are combined if their summed mass and crosslinker modification fit the precursor mass. A very similar strategy is used in pLink (Yang et al., 2012) with the distinction that the list of candidate peptides is split into alpha and beta peptides based on the precursor mass. The two candidate lists are then combined and scored as peptide pair. This strategy was changed in pLink 2 (Chen et al., 2019), where first alpha peptide candidates are identified with query-peaks[1] from the spectrum. Then the alpha peptide candidates are used to retrieve beta peptide candidates that match the precursor mass minus the alpha peptide mass and the crosslinker mass. The peptide pairs are then fine-scored to compute the final list of CSMs. A similar search strategy was proposed earlier from an analysis of the fragmentation behavior of collision-induced dissociation (CID) fragmented crosslinked peptides (Giese, Fischer, and Rappsilber, 2015). The implementation of this search strategy, xiSEARCH (Mendes et al., 2019) exploits several characteristics from the fragmentation behavior analysis. Instead of placing the modifications of unknown mass on the crosslinkable residues, the crosslinked

[1] Unfortunately, there was no formal definition of this term in the pLink 2 paper.

fragments are computationally linearized. During this linearization, the most intense ions are selected, and their m/z's are transformed by applying a heuristic to yield the m/z of linear fragments. This heuristic builds upon two observations: first, crosslinked fragments have a higher mass and charge than linear fragments. Second, the linear fragments can be computed by subtracting the crosslinked fragment mass from the precursor mass. From the linearized fragment ions, alpha peptide candidates are retrieved and subsequently for each alpha candidate matching beta peptides are computed.

Coverage gap. Regardless of the applied search paradigm, the fragmentation behavior of crosslinked peptides has proven difficult for the reliable identification of both peptides. Instead of only considering targets (T) and decoys (D) as in a linear search, CL engines must consider TT, TD, and DD matches. Ideally, in TT matches both peptides achieve a high score, in TD matches the T achieves a high score and the D achieves a low score, and in DD matches both peptides achieve a low score. Unfortunately, this scenario does not necessarily reflect reality. Trnka et al. (Trnka et al., 2014) reported for their search engine that the first peptide contributes a much larger proportion of the final score than the second peptide. Unequal coverage of the two peptides under high-energy collisional dissociation (HCD) conditions was the main reason for this score gap. Under CID conditions it was even reported that the intensities of the fragment ions differ considerably (Giese, Belsom, and Rappsilber, 2016). In addition to the different fragmentation behavior of CL peptides, spectra of poor quality will also contribute to low scoring matches, as observed for linear peptides (Renard et al., 2010).

Identification bias. The coverage gap also introduces potential biases to the peptide identification. A common challenge are isobaric peptide combinations. With the large search space many peptide combinations can match the precursor mass. If only one peptide is well-characterized, the second peptide can be matched with only a few or no matched fragment ions by the precursor mass, leaving considerable uncertainty on the second peptide identification. A potential source for misidentifications are consecutive peptides from a single protein that get identified as crosslinked (Iacobucci and Sinz, 2017). If the crosslinker only hydrolyzes partially on the consecutive peptide, the mass is the same as if there was a crosslinked peptide. This source of error can easily be avoided by excluding all consecutive peptides identifications after the database search. Another potential bias rises from ambiguous crosslink site assignments, especially with promiscuous crosslinking reagents such as sulfosuccinimidyl 4,4'-azipentanoate (sulfo-SDA) (Belsom et al., 2016). No or limited backbone fragmentation from the crosslinked peptides can lead to severe artifacts that prevent the meaningful interpretation of crosslinked data. For sulfo-SDA it was observed that under specific liquid chromatography conditions noncovalent peptide associations (NAPs) get identified as crosslinked peptides (Giese et al., 2019). The shown characteristics of NAPs revealed the vulnerability of search engines if no proper fragmentation of the crosslinked fragments is required for identification. In the above-mentioned publication the false discovery estimation (FDR), as judged by the percentage of long-distance links, was three times higher than the estimated FDR. While NAPs can largely be avoided by disruptive ionization settings, the underlying missing evidence in the MS2 domain and resulting coverage gap remains a challenge. To overcome the coverage gap, experimental and computational solutions are feasible.

Experimental Solutions

Fragmentation methods. Modern mass spectrometers such as the Orbitrap Fusion Lumos (Thermo Fisher Scientific) offer versatile acquisition strategies. Apart from the choice of the mass analyzer for the MS1 and MS2 spectra acquisition, a central part is the choice of the fragmentation method. Traditionally, peptide fragmentation was mostly centered around CID or HCD fragmentation for linear and crosslinked peptides. Alternative fragmentation methods include electron-transfer dissociation (ETD) or combinations of ETD and CID/HCD (EThciD/EThcD). For linear proteomics, decision trees have been proposed to enhance fragmentation efficiency based on precursor charge and m/z (Swaney, McAlister, and Coon, 2008). This concept has also been investigated for CLMS and the universal crosslinking approach. The combined fragmentation method EThcD yielded on average the highest crosslinked peptide sequence coverage (Giese, Fischer, and Rappsilber, 2016). For the individual peptides, the coverage gap was largest in CID but comparable in ETD, ETciD, EThcD, and HCD. The increased acquisition time in EThcD comes with an overhead in cycle time and thus fewer acquired spectra compared to HCD. To maximize the number of scans and sequence coverage a data-dependent decision tree was proposed also for CLMS (Giese, Fischer, and Rappsilber, 2016). The combination of different fragmentation methods is not the only choice to optimize fragmentation efficiency. Since most fragmentation methods need to be parameterized (e.g. collision energy for HCD, or reaction time for ETD), they can also be optimized individually (Diedrich, Pinto, and Yates, 2013). This process is not straight forward since b- and y-ions seem to have different optimal normalized collision energies (NCE) (Révész et al., 2018).

Cleavable crosslinker. In addition to the universal approach a specialized MS-cleavable approach can also be used (O'Reilly and Rappsilber, 2018). Here, crosslinker reagents containing MS-cleavable chemical groups are used that under CID/HCD conditions results in characteristic peak doublets that reveal the individual peptide masses (Kao et al., 2011; Sinz, 2017). This approach simplifies the identification of crosslinks by reducing the computational complexity to that of linear peptide identification. Similarly, optimized acquisition schemas have been investigated that make use of sequential CID-ETD fragmentation (Liu et al., 2017) and MS3 scans of the individual peptides. Since the alpha and beta peptides are then fragmented independently, preferred backbone cleavage in one of the peptides does not influence the other peptide. However, the involved acquisition cycle for a single precursor typically results in reduced numbers of acquired spectra and thus a loss in sensitivity. This method is most powerful if all signature peaks are observed, because only then the MS3 acquisition is triggered. Currently this acquisition schema is recommended when using the software XlinkX (Gonzalez-Lozano et al., 2020; Liu et al., 2017). Unfortunately, observing both doublets is rarely the case (Lu et al., 2018). Another acquisition strategy builds upon a stepped NCE in which low NCEs are used to produce (preferentially) reporter ion peaks, and higher NCEs are used for fragment ions (Stieger, Doppler, and Mechtler, 2019). This approach is attractive since acquisition time is saved by avoiding MS3 spectra but introduces again the coherent fragmentation of the two peptides. So far, experimental approaches have not yet solved the coverage gap satisfactory. For a more comprehensive review of current acquisition trends, crosslinker chemistry, and computational solutions the reader is referred to the following articles: (O'Reilly and Rappsilber, 2018; Steigenberger et al., 2020; Yu and Huang, 2018).

Computational Solutions

RT in CLMS. CLMS relies on the coordinated development of computational solutions for novel acquisition schemas and crosslinkers. Many lessons can be learned from linear proteomics, e.g. from considering multiple monoisotopic peaks (Lenz et al., 2018) or using Skyline (MacLean et al., 2010) and Spectronaut for quantitation of crosslinked peptides (Müller et al., 2018; Müller et al., 2019). Another idea from linear proteomics is to filter identified peptide spectrum matches (PSMs) based on the difference between observed and predicted retention times (Klammer et al., 2007). Theoretically, this approach introduces orthogonal information from the physicochemical properties of the peptides in the identification routine. For reversed-phase (RP) chromatography, which is typically directly coupled to the MS, the mass of the peptide correlates with the observed RT. Therefore, the readout of the precursor mass is only partially useful to distinguish false from correct identifications. For CSMs, predicted RTs may offer a loophole to distinguish correct from incorrect identifications. A typical scenario where misidentification occurs is when one peptide is very thoroughly fragmented and the other is not. This could either stem from the physicochemical properties of the crosslinked peptide or from the involvement of an incorrect match. One would expect that the RT characteristics from these two scenarios would help to distinguish correct from incorrect matches. Since CLMS experiments often include an additional fractionation step, the RT prediction would not necessarily be limited to RP.

CLMS experiments. Fractionation requires more material but increases sensitivity and analysis depth compared to multiple injections of the same sample. Ideally, the applied fractionation method can be used to enrich for a specific type of molecule in early or later phases of the applied gradients and is orthogonal to the chromatography method coupled to the MS. Typical fractionation methods for CLMS include size exclusion chromatography (SEC), strong anion exchange chromatography (SCX), or hydrophilic strong anion exchange chromatography (hSAX) which all can be used to enrich for crosslinked peptides. In addition, SCX and hSAX showed, at least in linear proteomics, orthogonal separation capabilities (Ritorto et al., 2013; Ruprecht et al., 2017). To decrease the complexity even further, multi-dimensional fractionation methods are also being used in CLMS (Ryl et al., 2020). The extensive use of fractionation techniques leaves room for a multi-dimensional RT prediction to supplement the identification of crosslinks. Based on earlier observations that the elution behavior of hSAX is predictable (Giese, Ishihama, and Rappsilber, 2018) this approach was investigated for CLMS (Chapter 6).

Using retention time predictions. Assuming that an accurate RT model can be trained, the challenge remains how to incorporate the additional information into a search engine score. PeptideProphet (Ma, Vitek, and Nesvizhskii, 2012) and Percolator (Käll et al., 2007) were among the first tools to use machine learning for this task. Both tools aim to discriminate between correct and incorrect PSMs from multiple score features. For the CLMS analysis Kojak (Hoopmann et al., 2015) allows the use of either tool in the post-processing of the results. So far, no rescoring method has been tested together with the RT prediction of crosslinked peptides. First results with unoptimized rescoring algorithms seem promising, suggesting to increase the number of identifications at constant FDR (Chapter 6).

Machine learning in proteomics. The development of machine learning applications in proteomics is mainly driven by the availability of large data sets. For linear peptides millions of acquired spectra were used to develop accurate MS2 intensity prediction tools (C. Silva et al., 2019; Gessulat et al., 2019; Zhou et al., 2017).

Together with the availability of large data sets and advancements in the field of deep learning (LeCun, Bengio, and Hinton, 2015) MS-based proteomics can be supplemented with machine learning models for protein digestion, retention time, and peptide identification (Bouwmeester et al., 2020). Especially, the application of deep learning methods is attractive since they are suited for complex tasks. Another advantage is that deep learning methods allow the application of transfer learning. In these applications a deep neural network gets trained in a domain where a lot of data is available, e.g. general-purpose image classification. In the transfer step, the already trained network is fine-tuned for a more specific problem, e.g. lung cancer detection from imaging systems (Fang, 2018). Conceptually, the advantage is that the trained network already developed a robust method to recognize important elements in an image. Therefore, much less data is required on the specialized task (e.g. lung cancer detection). Similarly, the availability of large data sets in linear proteomics might be useful for the development of specialized applications in the CLMS analysis.

1.2 Contributions and Main Findings

In this chapter I list my contributions to advance the field of crosslinking mass spectrometry. Manuscripts 1-5 have been peer-reviewed and accepted for publication. All manuscripts were written through contributions from all co-authors.

Manuscript 1

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

The first publication entitled *A Study into the Collision-induced Dissociation (CID) Behavior of crosslinked Peptides* (Chapter 2), is a fundamental contribution to improve the understanding on the behavior of crosslinked peptides in the mass spectrometer. This knowledge played a crucial role in the development of efficient search engines that exploit the main findings. The use of isotope-labeled crosslinkers is not required for an efficient search strategy in the universal approach. Instead crosslinked peptide and fragment ion properties allow a heuristic reduction of the search complexity due their higher mass and charge. In addition, it became apparent that the coverage gap in crosslinks is also amplified by the unequal distribution of fragment intensities. For this manuscript, I performed the exploratory data analysis, analyzed all the data and wrote the manuscript.

Manuscript 2

Giese, S. H., Belsom, A., & Rappsilber, J. (2016). Optimized fragmentation regime for diazirine photo-crosslinked peptides. Analytical Chemistry, 88(16), 8239–8247. https://doi.org/10.1021/acs.analchem.6b02082

The second publication entitled *Optimized fragmentation regime for diazirine photo-crosslinked peptides* (Chapter 3), is a direct follow-up to address the shortcomings that were revealed in the first manuscript. We observed that HCD fragmentation is the fastest available method and thus generated most links in an SDA crosslinked single protein sample. However, the coverage of the second peptide could be greatly improved by using EThcD fragmentation as well as the site-calling precision. To optimize the total coverage in an experiment we recommend using a data-dependent decision tree that chooses the best fragmentation method based on the m/z and charge of a peptide. I thank Adam Belsom for acquiring the data for the project. I performed the exploratory data analysis, analyzed all the data, and wrote the manuscript.

Manuscript 3

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). https://doi.org/10.1021/acs.analchem.8b04037

The third publication entitled *Noncovalently Associated Peptides Observed during Liquid Chromatography-Mass Spectrometry and Their Effect on crosslink Analyses* (Chapter 4), demonstrates that poor fragmentation of crosslinked peptides can lead to unwanted artefacts during the identification process under certain liquid chromatography conditions. To reduce the identification artefacts from noncovalent peptide associations (NAPs) we recommend using more disruptive ionization settings on the mass spectrometer. For already acquired data a heuristic retention time filter can help to recognize NAPs. Again, I thank Adam Belsom and Ludwig Sinn for acquiring the data. I performed the exploratory data analysis, analyzed all the data, and wrote the manuscript.

Manuscript 4

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

The fourth manuscript entitled *Peptide Retention in Hydrophilic Strong Anion Exchange Chromatography Is Driven by Charged and Aromatic Residues* (Chapter 5), was not based on crosslinking mass spectrometry data. Instead, we analyzed a large set of linear peptide identifications to model and understand the retention time behavior of peptides separated by hydrophilic strong anion exchange chromatography (hSAX). This fundamental work showed that anion exchange chromatography is also largely identified by aromatic amino acids. The development of a first predictor for the retention times in hSAX, allowed us to transfer the knowledge to crosslinking mass spectrometry (next section). I performed the exploratory data analysis, analyzed all the data, and wrote the manuscript.

Manuscript 5

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). https://doi.org/10.1038/s41467-021-23441-0

The fifth manuscript entitled *Retention time prediction using neural networks increases identifications in crosslinking mass spectrometry* introduces the first model to perform multi-dimensional peptide retention time prediction for crosslinked peptides (Chapter 6). The data consists of deep-fractionated crosslinked samples by SCX, hSAX and RP coupled to the mass spectrometer. The developed tool xiRT can be used to generate accurate RT predictions for RP as well as fraction-based predictions for SCX and hSAX. The application of predicted RTs is demonstrated in a crosslinked spectrum match rescoring approach using a support vector machine. This workflow increased the number of PPIs at constant 1% PPI FDR. I thank Ludwig Sinn and Fritz Wegner for acquiring the data. I performed the exploratory data analysis, developed the RT prediction models, analyzed all the data, and wrote the manuscript together with Ludwig Sinn (shared first author).

1.3 Additional Publications

In addition to the manuscripts in this thesis, I authored or co-authored the following publications:

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. Mol. Syst. Biol. 2019, 15 (9), e8994.
An integrated workflow for crosslinking mass spectrometry. https://doi.org/10.15252/msb.20198994

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

Karayel, Ö., Şanal, E., **Giese, S. H.**, Üretmen Kagıalı, Z. C., Polat, A. N., Hu, C.-K., Özlü, N. (2018). Comparative phosphoproteomic analysis reveals signaling networks regulating monopolar and bipolar cytokinesis. Scientific Reports, 8(1), 2269. https://doi.org/10.1038/s41598-018-20231-5

Belsom, A., Mudd, G., **Giese, S. H.**, Auer, M., & Rappsilber, J. (2017). Complementary Benzophenone Cross-Linking/Mass Spectrometry Photochemistry. Analytical Chemistry, 89(10), 5319–5324. https://doi.org/10.1021/acs.analchem.6b04938

Giese, S. H., Zickmann, F., & Renard, B. Y. (2016). Detection of Unknown Amino Acid Substitutions Using Error-Tolerant Database Search. Methods in Molecular Biology (Clifton, N.J.), 1362(9), 247-264. https://doi.org/10.1007/978-1-4939-3106-4_16

Polat, A. N., Karayel, Ö., **Giese, S. H.**, Harmanda, B., Sanal, E., Hu, C. K., Özlü, N. (2015). Phosphoproteomic analysis of aurora kinase inhibition in monopolar cytokinesis. Journal of Proteome Research, 14(9), 4087–4098. https://doi.org/10.1021/acs.jproteome.5b00645

Giese, S. H., Zickmann, F., & Renard, B. Y. (2014). Specificity control for read alignments using an artificial reference genome-guided false discovery rate. Bioinformatics, 30(1), 9–16. https://doi.org/10.1093/bioinformatics/btt255

Chapter 2

Manuscript 1. A Study into the Collision-induced Dissociation (CID) Behavior of Cross-Linked Peptides

Research

A Study into the Collision-induced Dissociation (CID) Behavior of Cross-Linked Peptides*[S]

Sven H. Giese‡§, Lutz Fischer§, and Juri Rappsilber‡§¶

Cross-linking/mass spectrometry resolves protein–protein interactions or protein folds by help of distance constraints. Cross-linkers with specific properties such as isotope-labeled or collision-induced dissociation (CID)-cleavable cross-linkers are in frequent use to simplify the identification of cross-linked peptides. Here, we analyzed the mass spectrometric behavior of 910 unique cross-linked peptides in high-resolution MS1 and MS2 from published data and validate the observation by a ninefold larger set from currently unpublished data to explore if detailed understanding of their fragmentation behavior would allow computational delivery of information that otherwise would be obtained via isotope labels or CID cleavage of cross-linkers. Isotope-labeled cross-linkers reveal cross-linked and linear fragments in fragmentation spectra. We show that fragment mass and charge alone provide this information, alleviating the need for isotope-labeling for this purpose. Isotope-labeled cross-linkers also indicate cross-linker-containing, albeit not specifically cross-linked, peptides in MS1. We observed that acquisition can be guided to better than twofold enrich cross-linked peptides with minimal losses based on peptide mass and charge alone. By help of CID-cleavable cross-linkers, individual spectra with only linear fragments can be recorded for each peptide in a cross-link. We show that cross-linked fragments of ordinary cross-linked peptides can be linearized computationally and that a simplified subspectrum can be extracted that is enriched in information on one of the two linked peptides. This allows identifying candidates for this peptide in a simplified database search as we propose in a search strategy here. We conclude that the specific behavior of cross-linked peptides in mass spectrometers can be exploited to relax the requirements on cross-linkers. ***Molecular & Cellular Proteomics 15: 10.1074/mcp.M115.049296, 1094–1104, 2016.***

Cross-linking/mass spectrometry extends the use of mass-spectrometry-based proteomics from identification (1, 2), quantification (3), and characterization of protein complexes (4) into resolving protein structures and protein–protein interactions (5–8). Chemical reagents (cross-linkers) covalently connect amino acid pairs that are within a cross-linker-specific distance range in the native three-dimensional structure of a protein or protein complex. A cross-linking/mass spectrometry experiment is typically conducted in four steps: (1) cross-linking of the target protein or complex, (2) protein digestion (usually with trypsin), (3) LC-MS analysis, and (4) database search. The digested peptide mixture consists of linear and cross-linked peptides, and the latter can be enriched by strong cation exchange (9) or size exclusion chromatography (10). Cross-linked peptides are of high value as they provide direct information on the structure and interactions of proteins.

Cross-linked peptides fragment under collision-induced dissociation (CID) conditions primarily into b- and y-ions, as do their linear counterparts. An important difference regarding database searches between linear and cross-linked peptides stems from not knowing which peptides might be cross-linked. Therefore, one has to consider each single peptide and all pairwise combinations of peptides in the database. Having n peptides leads to $(n2 + n)/2$ possible pairwise combinations. This leads to two major challenges: With increasing size of the database, search time and the risk of identifying false positives increases. One way of circumventing these problems is to use MS2-cleavable cross-linkers (11, 12), at the cost of limited experimental design and choice of cross-linker.

In a first database search approach (13), all pairwise combinations of peptides in a database were considered in a concatenated and linearized form. Thereby, all possible single bond fragments are considered in one of the two database entries per peptide pair, and the cross-link can be identified by a normal protein identification algorithm. Already, the second search approach split the peptides for the purpose of their identification (14). Linear fragments were used to retrieve candidate peptides from the database that are then matched based on the known mass of the cross-linked pair and scored as a pair against the spectrum. Isotope-labeled cross-linkers were used to sort the linear and cross-linked fragments apart. Many other search tools and approaches have been developed since (10, 15–19); see (20) for a more detailed list, at

From the ‡Department of Bioanalytics, Institute of Biotechnology, Technische Universität Berlin, 13355 Berlin, Germany; §Wellcome Trust Centre for Cell Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3BF, United Kingdom

Received February 24, 2015, and in revised form, December 3, 2015

Published, MCP Papers in Press, December 30, 2015, DOI 10.1074/mcp.M115.049296

Author contributions: S.H.G., L.F., and J.R. designed the research; S.H.G., L.F., and J.R. performed the research; S.H.G. and J.R. analyzed data; and S.H.G. and J.R. wrote the paper.

least some of which follow the general idea of an open modification search (21–24).

As a general concept for open modification search of cross-linked peptides, cross-linked peptides represent two peptides, each with an unknown modification given by the mass of the other peptide and the cross-linker. One identifies both peptides individually and then matches them based on knowing the mass of cross-linked pair (14, 22, 24). Alternatively, one peptide is identified first and, using that peptide and the cross-linker as a modification mass, the second peptide is identified from the database (21, 23). An important element of the open modification search approach is that it essentially converts the quadratic search space of the cross-linked peptides into a linear search space of modified peptides. Still, many peptides and many modification positions have to be considered, especially when working with large databases or when using highly reactive cross-linkers with limited amino acid selectivity (25).

We hypothesize that detailed knowledge of the fragmentation behavior of cross-linked peptides might reveal ways to improve the identification of cross-linked peptides. Detailed analyses of the fragmentation behavior of linear peptides exist (26–28), and the analysis of the fragmentation behavior of cross-linked peptides has guided the design of scores (24, 29). Further, cross-link-specific ions have been observed from higher energy collision dissociation (HCD) data (30). Isotope-labeled cross-linkers are used to distinguish cross-linked from linear fragments, generally in low-resolution MS2 of cross-linked peptides (14).

We compared the mass spectrometric behavior of cross-linked peptides to that of linear peptides, using 910 high-resolution fragment spectra matched to unique cross-linked peptides from multiple different public datasets at 5% peptide-spectrum match (PSM)[1] false discovery rate (FDR). In addition, we repeated all experiments with a larger sample set that contains 8,301 spectra—also including data from ongoing studies from our lab (Supplemental material S9-S12). This paper presents the mass spectrometric signature of cross-linked peptides that we identified in our analysis and the resulting heuristics that are incorporated into an integrated strategy for the analysis and identification of cross-linked peptides. We present computational strategies that indicate the possibility of alleviating the need for mass-spectrometrically restricted cross-linker choice.

EXPERIMENTAL PROCEDURES

Spectra Collection and Filtering—We collected database search results from experiments that were acquired and described in previous publications (31–33) (Pride: PXD002142, PXD001835, PXD001454) and accumulated cross-linked and linear peptide spectrum matches (PSMs). All data were acquired in CID mode on hybrid linear iontrap-Orbitrap mass spectrometers (LTQ Orbitrap Velos, Thermo Scientific, Bremen, Germany). The cross-linker in all searches was bis(sulfosuccinimidyl)suberate or its isotopic variant bis(sulfosuccinimidyl)suberate-d4. A typical search was performed using Xi (ERI Edinburgh, UK) and the following parameters: MS accuracy, 6 ppm; MS/MS accuracy, 20 ppm; enzyme, trypsin; maximum missed cleavages, 4; maximum number of modifications, 3; fixed modification, carbamidomethylation on cysteine; variable modifications, oxidation on methionine; and modification by the hydrolyzed or the ammonia reacted cross-linker on lysine, serine, threonine, tyrosine, and the protein N terminus. Cross-linking was allowed to involve lysine, serine, threonine, tyrosine, and the protein N terminus. To ensure the analysis of high-quality data, we extracted 910 PSMs to unique cross-linked peptides at a 5% FDR cutoff using XiFDR (v. 1.0.4.13, (31)). Along with the 910 cross-linked PSMs, we extracted 4,161 linear PSMs from the cross-linking acquisitions as a reference data set for linear peptides. Detailed information about each PSM is available in the Supplemental Table S1 along with the annotation of the cross-linked peptides (Supplementary File S2). In addition, we repeated all experiments with a larger sample set that contains 8,301 spectra—also including data from ongoing studies from our lab (Supplemental material S9-S12). To provide a comparison on the specific mass-spectrometric properties, we included search results from a linear peptide identification experiment using MaxQuant from a cyclin-dependent kinase (CDK)-regulated chicken chromatin dataset (34) on our machine with 1% FDR.

Data Extraction—Software written in Python (2.7, www.python.org) was used to extract relevant fragmentation information from the local PostgreSQL database containing details about search settings and spectra annotations. For each PSM involving a cross-linked peptide, the match score, peptide sequence (alpha and beta), precursor charge, experimental mass, and cross-link position (alpha and beta peptide) were extracted. In addition, the identified fragments were stored with each PSM. For each fragment, the *m/z*, charge, fragment type, intensity, and associated isotope cluster information were stored. When isotope clusters were identified, the summed intensity over all isotope peaks was used instead of the intensity of the monoisotopic peak. Similarly, linear PSMs were extracted. After extracting all fragments, the intensity for each fragment was normalized by division by the most intense fragment from the respective PSM. In addition, the respective intensity rank for each matched fragment was stored. A high rank refers to a high intensity and a rank of one to the lowest intensity in that PSM. For example, a spectrum containing three matched peaks with fictive intensities (10, 3, 1) was first normalized by the base peak to arrive at (1, 0.3, 0.1). Then, the ranks were derived such that the intensities are converted to (3, 2, 1). To compare the ranked intensity among peptides of different length (as done in Figs. 3*A* and 3*B*), the rank was further normalized by the number of matched peaks per spectrum. Thereby, the highest intense peak received a normalized rank value of 1. For the fictive example, this led to peak intensities of (1, 0.66, 0.33). We then compared b- or y-ion intensities for fragments in relation to the linker position or the peptide length, disregarding the specific ion index information (*e.g.* y7).

Similarity Computation of Linear and Cross-Linked Spectra—The similarity comparison of two spectra was realized via an adapted ranked dot product scoring scheme. The ranked dot product is usually used in spectral library searches where acquired spectra are compared *versus* annotated spectra from previous database identifications (35). Here, we define the ranked dot product as follows:

$$RDP = \frac{S_r \times T_r}{\sqrt{S_r^2 \times T_r^2}} \quad \text{(Eq. 1)}$$

where $S_r \times T_r$ is the scalar product of the two vectors S_r and T_r that represent the identified ions of the source and target peptide, respectively. Usually, the vectors S_r and T_r contain binned intensity values

[1] The abbreviations used are: PSM, peptide spectrum match.

from the observed spectrum and the target spectrum from the spectral library. Here, we adapted the scoring scheme such that only nonlossy, b- and y-ion intensities were compared. For example, for a peptide of length five, the vectors S_r and T_r have length eight. Moreover, instead of using actual intensity values, we replaced intensity values by intensity ranks (35). If a specific ion type was present in the source but not in the target peptide, the intensity for that particular ion in the target peptide was set to zero and *vice versa*. Otherwise, the intensity for each ion was derived via its rank. To evaluate the scoring behavior, a reference similarity distribution of random pairings of cross-linked peptides was computed. The reference distribution was derived by computing the similarity of 1,000 random peptide combinations. We made sure that no comparison of peptides with the same sequence is included. The resulting random score distribution was used to evaluate all other score distributions.

Evaluation of the Predictive Power to Distinguish Linear and Cross-Linked Fragments—Based on the ground truth of 910 PSMs, we evaluated the predictive power of the relative fragment mass and the charge state as indicators whether or not a fragment is cross-linker containing. The applied constraints were the fragment mass divided by the precursor mass, the charge state of the fragment, and the combination of both. Only fragments with isotope clusters were used for this analysis. The performance of the classification was evaluated via the sensitivity, defined as $sn = \frac{TP}{TP + FN}$, and the specificity, defined as $sp = \frac{TP}{TP + FP}$, where a true positive (*TP*) is a fragment that was annotated as cross-linker containing and is also predicted as such, a false positive (*FP*) is a fragment that was annotated as linear but was predicted as cross-linked, and a true negative (*TN*) is a fragment that was annotated as linear and was also predicted linear. Lastly, a false negative (*FN*) is a fragment that was annotated as cross-linked but was not recognized as such.

RESULTS AND DISCUSSION

Mass and Charge of Cross-Linked Peptides Can Be Used to Direct Data-Dependent Acquisition—Digestion of cross-linked proteins yields both linear and cross-linked peptides. We wondered if the signals of cross-linked peptides either in MS1 or MS2 differed systematically from those of linear peptides. Note that we are focusing here on the most frequent form of cross-linked peptides: tryptic peptides that are cross-linked via lysine residues or serine, threonine, and tyrosine.

The precursor masses of (tryptic, Lys/Ser/Thr/Tyr-linked) cross-linked peptides and (tryptic) linear peptides have a large overlap in their mass distribution (Fig. 1*A*). However, in the margin area, *i.e.* considering all masses up to 1,300 Da, linear peptides are more frequently observed than cross-linked peptides. Given a mass cutoff of *e.g.* 1,300 Da, it is possible to reduce the complexity of the sample dramatically, *i.e.* 33.3% of the linear spectra can be disregarded. This benefit comes with a loss of 2% in unique cross-linked peptides. Often, these hits are disputable because both or one of the peptides in the cross-linked product is rather short. In these cases, reliable identification is usually not possible. Thus, restricting acquisition to precursors above 1,300 Da appears a viable strategy to enrich for cross-linked peptides. Cross-linked peptides having a larger size than linear peptides can be rationalized by them being a pair of peptides. In addition, a protease-cleavage site is frequently blocked when using lysine-reactive cross-linkers and trypsin. Cross-linked peptides would then be expected to be a pair of peptides each having

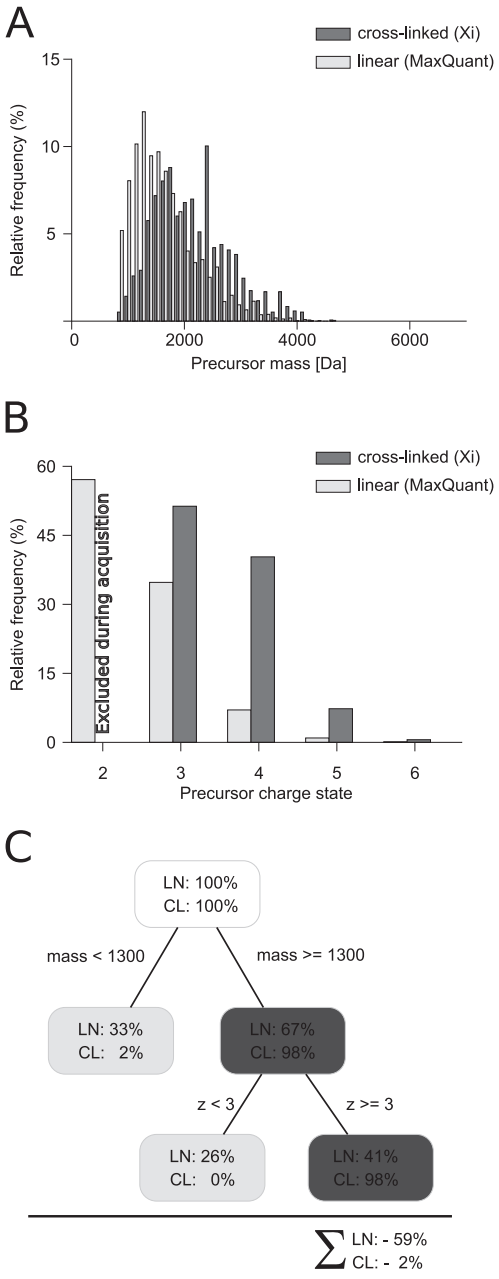


FIG. 1. **Precursor properties of linear and cross-linked peptides.** (*A*) Comparison of precursor masses from linear and cross-linked identifications. (*B*) Comparison of the charge state from cross-linking acquisitions (charge state 1 and 2 were excluded during acquisition) and noncross-linked acquisitions (charge state 1 was excluded). (*C*) Decision tree to enrich for cross-linked peptides. The cross-linking results are derived from 1,255 PSMs identified with a 5% false discovery rate and a minimum peptide length of 4. The linear identifications contain 14,361 PSMs with a 1% false discovery rate.

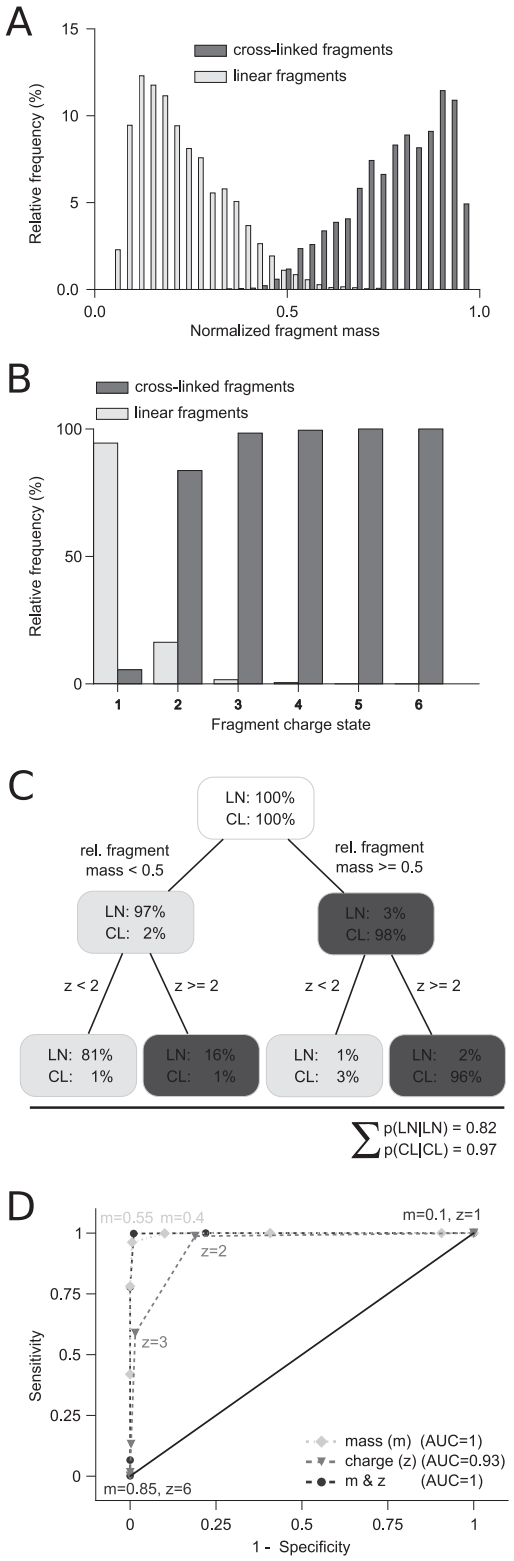


FIG. 2. **Fragment properties of cross-linked peptides.** (*A*) Comparison of cross-linked and linear fragment masses of cross-linked peptides normalized by their precursor mass. (*B*) Distribution of assigned charge states from isotope clusters distinguished in cross-linked and linear fragments of cross-linked peptides. (*C*) Decision tree visualizing the process to decide if a fragment is cross-linked or linear based on charge and mass. (*D*) Receiver operating characteristic curve showing the sensitivity ($TP/(TP + FN)$) and specificity ($TN/(FP + TN)$) for assigning a cross-linked fragment as cross-linked and linear fragments as noncross-linked. Thresholds are annotated and based on charge and/or mass. The data were derived from 910 high-confidence identifications with a 5% false discovery rate (FDR) and a minimum peptide length of 6. Abbreviations: TP, true positives; FN, false negatives; TN, true negatives; FP, false positives.

a missed cleavage site and thus in total four-times the mass of a linear peptide on average. This was already utilized in sample preparation by enriching for cross-linked peptides in size exclusion chromatography (10).

Cross-linked peptides are often higher charged than linear peptides (Fig. 1*B*), as was noted based on smaller sample sizes previously (9, 14). We investigated this here in detail based on our set of 910 PSMs. All our data in cross-link analyses were acquired excluding charge states 1 and 2, based on our initial observations (9). Therefore, nothing more can be said here on the occurrence of cross-linked peptides in these charge states. Looking at linear peptides from non-cross-linked samples, more than half (57%) are doubly charged. This supports the current strategy of at least excluding doubly charged precursors during data acquisition (9, 14). Adding triply charged precursors to the exclusion (14) further improves on this by removing an additional 35% of linear peptides. However, excluding triply charged precursors from fragmentation analysis also reduces the number of identified cross-linked peptides by almost half (48%). Given this considerable loss of cross-linked peptides, it appears advisable to exclude only doubly and not also triply charged precursors from the analysis, at least when working with ionization conditions similar to ours (9).

In summary, an enrichment of 2.3-fold could be achieved for cross-linked over linear peptides. This is based solely on MS1 peak characteristics and comes at no additional experimental costs. It should be noted that this is comparable and possibly complementary to the fold enrichment achieved by the currently widely used chromatographic enrichment strategies, strong cation exchange (9) or size exclusion chromatography (10) for cross-linking experiments. In chromatographic methods, about 50% of the linear peptides never reach the mass spectrometer. In the acquisition-based approach, they do but are not selected for MS2.

Mass and Charge Reveal the Cross-Link Status of Fragments without Using Isotopes—Extending the mass and charge analysis to fragments (Fig. 2) leads to the observation that linear fragments can be distinguished from cross-linked fragments with high confidence. We define the normalized fragment mass as the fragment mass divided by the precursor mass. Looking at the normalized fragment mass reveals that the distributions for cross-linked and linear fragments are very

well separated. Linear fragments tend to have a smaller mass than 50% of the precursor mass. In contrast, cross-linked fragments tend to have a larger mass than 50% of the precursor mass. Very few linear fragments (2.5%) and cross-linked fragments (1.6%) are not following this rule. Consequently, the mass-based prediction is highly successful. Setting the decision boundary to 50% precursor mass yields a sensitivity of 0.99 and a specificity of 0.97. The corresponding receiver operating characteristic curve yields an area under the curve of 0.996 (Fig. 2*D*) by relative fragment mass alone.

In addition to the fragment mass, the charge state distribution differs for linear and cross-linked fragments (Fig. 2*D*). Essentially all fragments with charge state one are linear. Similarly, the vast majority of fragments that are triply or higher charged are cross-linked. If the fragment is doubly charged, the probability that the fragment is cross-linked is four-times higher than being linear. Hence, cross-linked and linear fragments can be very well separated by evaluating the charge state of the fragment, reaching a sensitivity of 0.98 and a specificity of 0.81, respectively. The charged-based prediction yields an overall area under the curve of 0.93. A combined approach of normalized fragment mass and charge state to detect cross-linked fragment species provides additional resolving power and increases the area under the curve to ~1 (Fig. 2*D*).

One of the first search algorithms for the identification of cross-links by database searching builds on the idea of knowing which fragments are linear and which are cross-linked (14). The cross-link status of the fragments was assessed through isotope labeling. Using isotope-labeled cross-linker, cross-linked fragments experience a mass shift in the fragmentation spectra of light and heavy cross-linked peptides. In contrast, linear fragments are observed with identical mass in both fragmentation spectra. While being intriguing, this approach for determining the cross-link status of fragments has a number of inherent setbacks: (1) Both peaks of a labeled cross-linked peptide have to be selected for fragmentation; selecting only one does not yield the required information. (2) The MS1 signal of the cross-linked peptide is split into two, whereas other peptides are seen with their original intensity. (3) The choice of cross-linker is limited. (4) Any use of isotope labeling increases the complexity of the sample. (5) Use of isotopes for this purpose complicates their use for quantitation. We here present an alternative to isotope labeling for high-resolution fragmentation spectra. If the fragment charge can be determined and thus also the fragment mass, isotopes are not needed to determine the cross-link status of fragments. By using high-resolution data, the search algorithm can benefit for free from the confident distinction of linear and cross-linked fragments. This leaves isotopes for quantification of cross-links (31).

Cross-Linked Peptides Fragment Similar to the Corresponding Linear Peptides—Cross-linked peptides are expected to fragment like linear peptides to generate b- and y-ions under CID conditions. However, the extent to which this fragmentation is affected by the cross-link or the presence of two peptides in close proximity in the gas phase is not immediately clear. As a first step, we compared the fragmentation spectrum of the cross-linked peptide pair AEFAEVSKLVTDLTK–AFKAWAVAR with those obtained for both peptides individually (Fig. 3*A*, see Supplemental Fig. S6 for annotation of the individual spectra). For ease of comparison, the b- and y-ion signals of the noncross-linked peptides were moved to the same *m/z* value of the corresponding b- or y-ion in the cross-linked peptide. The two fragmentation spectra of the linear peptides together show a marked resemblance to the fragmentation spectrum of the cross-linked peptide pair, albeit some fragment yields are affected by the linkage. Furthermore, there was no dominant presence of double fragmentation observed. This means that despite a cross-linked peptide being more complex and having more parameters, its fragmentation follows essentially the same rules as apply to linear peptides. In essence, the cross-linked peptide fragmented like two linear peptides, each bearing the respective other peptide as a modification. This opens the prospect of at least initially dealing with both peptides individually during the identification process. Even if the final evaluation of matches should be done as a cross-linked pair, first candidates could be extracted from a linear instead of a quadratic search space.

Cross-linked peptide CID spectra contain fragments from two peptides but at unequal contribution. Usually, one of the two partners of a cross-linked peptide shows superior fragmentation, measured in the number of fragments and their intensities. Asymmetric sequence coverage of the two peptides in a cross-link has been observed previously, under HCD fragmentation conditions (30). We call the more dominant fragmented peptide the alpha peptide and the submissive peptide the beta peptide. Formally, the alpha peptide was defined as the peptide with more identified ions among the ten most intense peaks (Fig. 3*B*). On average, 78% of the fragments within the ten highest intense matched fragments are attributed to the alpha peptide (Fig. 3*C*). Alpha peptides show consistently higher intensities for b- and y-ions, whereas y-ions for both peptides are more intense than b-ions (Fig. 3*D*). As the two peptides differ in the intensity of their fragments, one could envision to use intensity as a means to separate the otherwise superimposed fragmentation spectra of both peptides of the cross-link. This suggests the possibility of separating the fragmentation spectra of alpha and beta peptides computationally, similarly to the use of MS2-cleavable cross-linkers experimentally (11, 12). MS2-cleavable cross-linkers, in addition, provide a route to the mass of the alpha and beta peptides but restrict the choice of the cross-linker. Also, normal cross-linkers cleave to some extent under HCD fragmentation at the bond between the cross-linker and the peptide (30). At least under our experimental conditions,

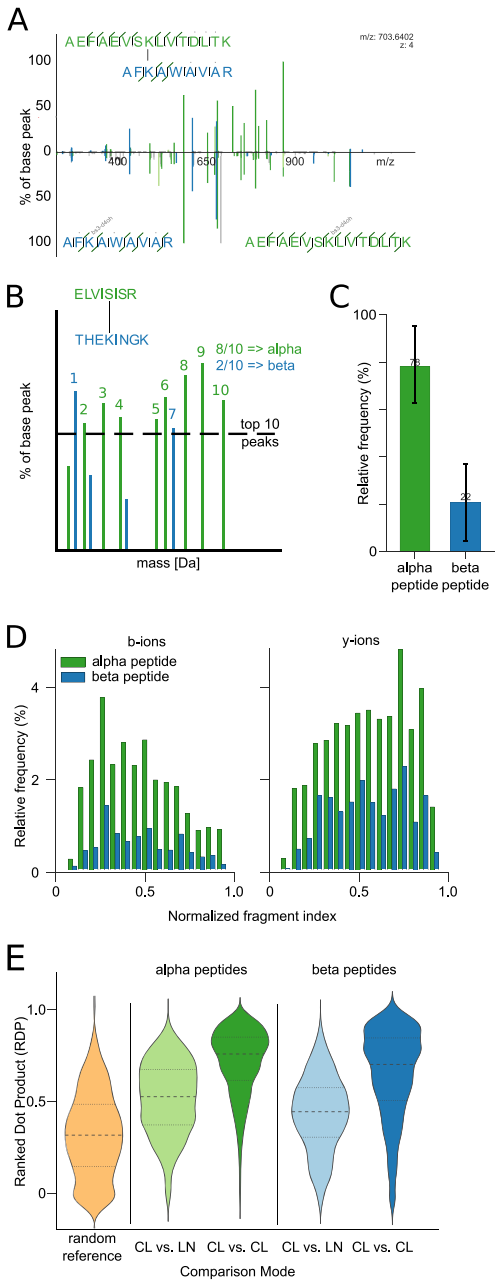


FIG. 3. **Fragmentation patterns of cross-linked peptides.** (*A*) Spectral comparison of a cross-linked peptide (*upper part*) and an overlay of the individual linear peptide spectra (*lower part*). Equivalent fragments from the cross-linked peptide and the respective linear peptides have been aligned to facilitate direct comparison (Supplemental Fig. S6 shows the individual spectra with annotations). (*B*) Visualization of an idealized (hypothetical) cross-linked peptide spectrum that is divided into alpha and beta peptides. The alpha peptide is defined as the peptide that has more ions among the ten most intense ions. (*C*) Distribution of annotated fragment peaks among the ten most intense ions of identified ions. The height refers to the mean with the standard deviation as error bars. (*D*) Quantitative analysis of b- and y-ion fragment peak intensities of alpha and beta peptides, respectively. (*E*) Quantitative comparison of the spectral similarity between linear (LN) and cross-linked (CL) peptides. A reference distribution is derived by randomly comparing spectra of cross-linked peptides. The data for (*B–E*) was derived from 910 high-confidence identifications with a 5% false discovery rate (FDR). Abbreviations: CL, cross-linked; LN, linear.

this is a rare event (10% of cross-linked peptides) (Supplemental Table S8).

Investigating more systematically the fragmentation similarity of peptides in cross-links and their linear counterparts reinforced the above conclusions. A quantitative view at the spectral similarity was achieved through exhaustive comparisons of cross-linked and linear peptides through the ranked dot product. For the systematic assessment of the spectral similarity, we used the linear peptide identifications from cross-link database searches and compared the spectra to all cross-linked peptides with the same sequence (Fig. 3*E*). Peptides in cross-links display large spectral similarity to their linear counterparts, regardless if alpha or beta peptides are considered. However, subspectra for a peptide in a cross-link look more alike, independent of the partner peptide or link position, than to the spectra of the corresponding linear peptide. Beta peptides generally perform less well in these comparisons. They tend to have less intense ions and also fewer ions. This reduces the overlap of beta peptide fragment ions between spectra. With a higher overlap in fragment ions, the spectral similarity increases and *vice versa*. Factors that potentially influence the fragmentation are the charge state, cross-linked residue, or the partner peptide. Of these, the highest influence on the fragmentation behavior comes from the charge state with minor, but present, effects from the other factors (see Supplemental Fig. S4).

Uncross-Linking Peptides by Data Analysis Resolves the n^2 Problem of Their Identification—In order to identify a pair of peptides that are cross-linked, one needs to consider the pairwise combination of all peptides in a database. As databases become bigger, this space grows quadratically. An exhaustive database construction could be avoided if a few candidates for at least one of the two peptides could be identified in a simplified first search. Ideally, one was to isolate the fragment peaks of one peptide. An adapted linear search can then retrieve candidates for this peptide without having to actually select a single one as the final match. Once having candidates for this peptide, one would know the mass of the corresponding second peptide, extract all mass matches from the original database of linear peptides, and construct a concentrated "bonsai" database of peptide pairs that would largely enrich for the cross-linked peptide. As observed above, intensity enriches fragment ions of the alpha peptide over those of the beta peptide. So, a stepwise extraction of candidates appears possible.

Extracting candidates for the alpha peptide as a linear peptide without knowing its peptide mass is complicated by the intense presence of cross-linked fragments and by the

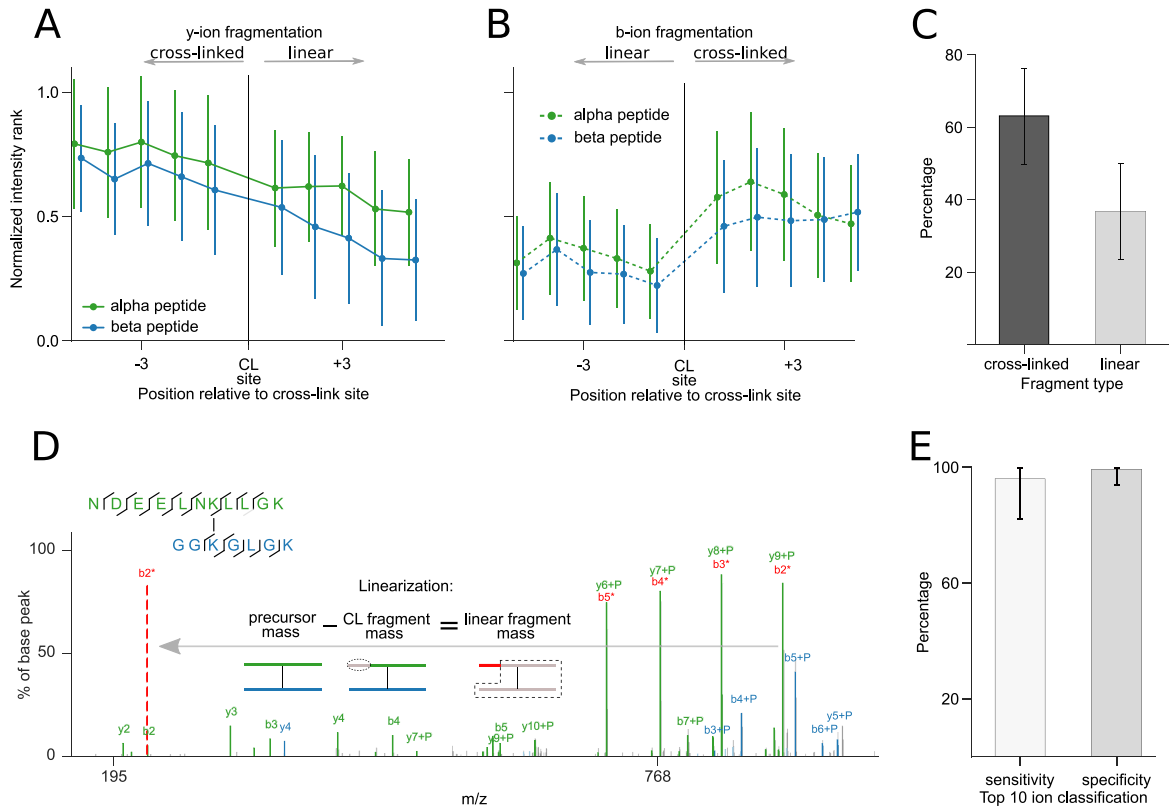


FIG. 4. **Cross-linked peptide fragmentation patterns.** Influence of the cross-link site (CL site) on y-ion (*A*) and b-ion yield (*B*), respectively. Longer y-ions are located to the left of the cross-link site; longer b-ions are located to the right of the cross-link site. Fragment intensities were transformed to ranks, with high intensities having a higher rank, and then normalized by the number of fragments in a spectrum. Error bars correspond to the standard deviation of all measured intensities at a relative position. (*C*) Distribution of cross-linker containing and linear fragments in cross-linked peptide spectra, respectively. (*D*) Example spectrum reflecting preferred cleavage of cross-linked fragments and an exemplary linearization of the cross-linked y7-ion of the alpha peptide. As shown in the pictogram of the linearization process, the y9-ion is transformed to the b2 ion (which was also observed as low intense peak) by subtracting the fragment mass from the precursor mass. Similarly, the y8 ion can be transformed to the b3 ion, which is indicated by the annotation with a '*' in the spectrum. (*E*) Sensitivity and specificity of correctly assigned cross-linked and linear fragments by their charge and mass from the top ten identified ions. The underlying data were extracted from 910 PSMs at a 5% FDR. For bar plots, the height and error bars refer to the mean and the standard deviation of all evaluated PSMs. The linear alpha peptides are also shown in Supplemental Fig. S7.

presence of fragments of the beta peptide. The dominance of the alpha peptide suggests the possibility of extracting a subspectrum that enriches for fragmentation information of this peptide. This could be achieved by simply taking the ten most intense fragments. However, this means that one looks primarily at cross-linked fragments (Fig. 4). For y-ions, longer fragments were seen with higher intensities (Fig. 4*A*). This favors cross-linked fragments that tend to be larger. For b-ions, there is no continuous effect. Instead, the cross-link site appears to exert a direct effect, leading primarily to cross-linked b-ions (Fig. 4*B*). The apparent influence of the link site on b-ions can be mechanistically explained through the presence of the second peptide. Charge in fragments is primarily carried by basic residues. y-ions of tryptic peptides have one by default at their C terminus. b-ions lack this terminal basic residue. However, cross-linked b-ions are modified by the second peptide. In this way, like y-ions they carry a C-terminal basic residue. The general dominance of cross-linked fragments (Fig. 4*C*) complicates the identification of the alpha peptide as they can only be used if the modification mass is known. However, this mass is inaccessible. The only solution would be to uncross-link the fragments.

Importantly, cross-linked fragments can be converted during data processing into their linear counterparts. Above, we established a reliable method to distinguish signals of cross-linked and not cross-linked fragments. The challenge is in converting cross-linked fragments into not cross-linked ones. Interestingly, any fragment also carries with its mass the information of the matching counterpart that is missing to make the whole peptide. Looking at this relation as a formula and resolving this formula to the missing fragment defines the mass of the fragment as the mass of the peptide less the mass of the observed fragment. If the peptide is cross-linked and the fragment is as well, then the missing fragment is

linear. In this way, we can convert a cross-linked fragment into a linear fragment. Since cross-linked fragments are generally observed more frequently (Fig. 4*C*), the linearization step provides a valuable information gain. As all fragment ions are linearized, the matching of fragments can be done entirely free of having to consider cross-links. In consequence, the processed MS2 spectrum contains the fragments of two linear peptides and thus provides much of the value of CID-cleavable cross-linkers.

The linearization is straightforward and highly reliable. For instance, the alpha peptide fragment ions y6+P, y7+P, y8+P, and y9+P (+P refers to the cross-linked partner peptide) (Fig. 4*D*) fulfill the 50% precursor rule, *i.e.* their fragment mass is larger than 50% of the precursor mass (note that they are the base peaks). To remove the dependence of P—and perform a simple linear matching—the cross-linked ions need to be linearized: The y6+P-ion is converted into its complementary b5-ion. y7+P becomes b4, y8+P becomes b3, and y9+P becomes b2. Note that the b-ions were also observed on their own in our example spectrum. However, this is not always the case, and they would not have made it under the ten most intense ions on their own. After the linearization, we only have linearized fragments in the spectrum that can be matched by standard database search approaches. We established above that 60–100% of the top ten matched peaks in the fragmentation spectrum of cross-linked peptides derived from alpha peptide fragmentation. Among these, we detected the cross-linked fragments with high success (~98% average per spectrum) by their charge or mass alone (Fig. 4*E*). In consequence, we can extract a subspectrum that is largely enriched in linear fragments of the alpha peptide, thus substituting further aspects of CID-cleavable cross-linkers.

Open modification search also resolves the n^2 problem, but it is still necessary to look for large modification masses. Knowing which fragments are cross-linked (and knowing how to linearize them) allows us to simplify the open modification search paradigm: Instead of considering wide gap mass ranges or all possible modification sites, a standard linear search is sufficient to identify candidates for at least one of the two peptides in the cross-linked peptide. Considering open modifications is computationally expensive. Possibly as a consequence, the prevalence of secondary cross-link reactions, *i.e.* serine, threonine, or tyrosine cross-links with bis-(sulfosuccinimidyl)suberate, are generally neglected (30). Our results show that these reactions make up ~14% of all cross-links and thus contribute largely to the outcome of an analysis (Supplemental Fig. S3). Identifying peptides with multiple cross-link sites becomes even more challenging if photoactivatable cross-linkers, such as sulfosuccinimidyl 4,4′-azipentanoate (sulfo-SDA) are used (36). Sulfo-SDA links some nucleophilic amino acids (lysine, serine, threonine, tyrosine, and the protein N terminus) with any other amino acid by having a standard N-hydroxysuccinimide (NHS)-activated ester on one side and a highly reactive diazirine group on the other. Therefore, open modification search paradigms would need to consider almost as many linkable residues as there are amino acids in the peptide to generate the right theoretical spectrum for each cross-linkable site. Existing search engines could utilize the highly reliable linearization process to avoid the probing of all possible cross-link sites.

An Integrated Search Strategy for Cross-Linked Peptides—With the above observations and concepts in hand, an integrated search strategy becomes possible. The quadratic search problem of cross-linked peptides can be simplified if the database size is decreased. Instead of combining exhaustively all peptides of the database, we first identify a set of candidates for one peptide. In a second step, all peptides can be extracted from the database that complete these candidates to obtain the mass of the observed cross-linked peptide. Combining these two sets of candidate linear peptides gives a focused database of candidate cross-linked peptides. The final identification is done against this largely reduced database. The stepwise candidate extraction is facilitated by the asymmetric fragmentation yield of cross-linked peptides. One peptide tends to give more intense fragment signals. An intensity cutoff can enrich, therefore, for information of one peptide in a simplified subspectrum comprising the n most intense peaks, *e.g.* $n = 10$. Unfortunately, cross-linked fragments contribute the majority to this subset of signals. However, using charge and relative mass as indicators, these can confidently be revealed and then converted into linear fragments. This removes any dependence of fragments on knowing the other peptide. Candidates for the first peptide can now be identified based on linear fragment data alone. Having a small set of candidates of the first peptide allows calculating the mass of the respective partner peptide candidates by simple algebra from the mass of the cross-link. Consequently, candidates for the second peptide can be extracted from the database by mass look-up. In this way, an initial set of peptide pair candidates is generated guided by data rather than following exhaustive combination of all peptides in the database. Exhaustive database search in this hugely reduced peptide pair database then allows identifying the final match.

We have implemented this search strategy in Xi and used it successfully in several studies (33, 37–44). In concrete terms (Fig. 5), we start with the full spectrum of all peaks from a MS2 scan. After charge state assignment and removal of isotopic peaks (1) the linearization of alpha peptide candidate ions is performed (2). The decision whether or not a fragment is going to be linearized depends on the relative precursor mass and the charge. If either the relative precursor mass is $>= 0.5$ or the charge state $>= 2$, the given fragment will be linearized. After the linearization, the ten highest ion signals are selected for a dedicated linear database search for alpha peptide candidates (3). The first search is a means to extract a moderate number of candidates for the alpha peptide without knowing the mass or location of the cross-link. A small number of peaks is usually sufficient to extract the true alpha peptide as

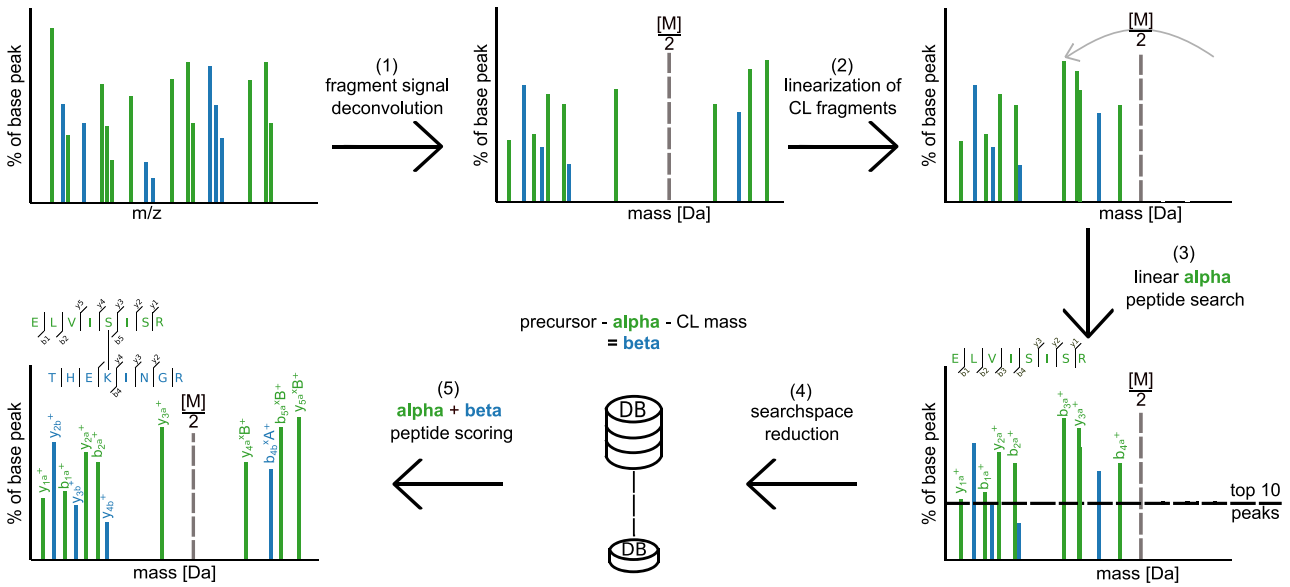


FIG. 5. **A search strategy for the identification of cross-linked peptides based on their CID behavior.** (*1*) A mass spectrum is processed by peak picking, deisotoping, resolving losses, and decharging. (*2*) Putative cross-linked fragment peaks are converted to linear fragment peaks. (*3*) The top ten peaks are extracted and matched against a linear database version. (*4*) *n* candidates for the alpha peptide are extracted. For each alpha peptide candidate, *m* beta peptide candidates are extracted such that each alpha/beta pair adds up to the precursor mass. (*5*) The combined identifications of alpha and beta peptides are then scored together.

one of the candidates from the database. However, the actual identification is done on the full spectrum together with the beta peptide. The list of alpha peptides is used to generate corresponding beta peptides by a precursor mass filter (4). Corresponding beta peptides are extracted by subtracting the alpha peptide mass and the cross-linker mass from the measured precursor, as also explained above. Finally, the matching peptide pairs (alpha + beta peptide candidates) are reevaluated on the initial, untreated spectrum to localize the cross-link site and perform a scoring with all fragment ions present. Only in this step, the final alpha and beta peptide pairing/scoring is done. The final match for any given spectrum is the one with the highest scoring pair. After all spectra have been processed, separate FDR estimation needs to be performed. Elements of our stepwise identification have been described previously (21).

Influence of the Sample Size on Our Analysis—We analyzed the precursor and fragment information of 910 PSMs that were identified in a collection of published experiments conducted in our lab. In addition, we challenged the presented analysis with all ongoing studies of our lab—yielding a total of 8,301 PSMs—to question if our set of 910 PSMs was large enough to arrive at general conclusions (see Supplemental material S9-S12). For example, the enrichment possibilities during acquisition for cross-linked peptides have been investigated in Fig. 1—coming to the conclusion that a charge and mass based selection filter greatly enriches cross-linked peptides (Supplemental Fig. S9). Based on 910 PSMs our analysis arrives at excluding 59% of linear peptides at the expense of losing 7% cross-linked peptides. Based on all our data, we conclude 59% of linear peptides can be excluded at the expense of losing 4% cross-linked peptides. As a second example, to distinguish cross-linked from linear fragments, we introduced a mass cutoff of 50% precursor mass: In 910 PSMs, 2.5% of linear fragments and 1.6% of cross-linked fragments were not following this 50% rule. For the larger collection (8,301 PSMs), 2.25% of linear fragments and 1.8% of cross-linked fragments did not follow this rule. Finally, 77% of the top ten fragment peaks derive from the alpha peptide in 910 PSMs. This contrasts to 80% in 8,301 PSMs. Fragment peak intensity is hence a reliable filter to assign a subset of fragments to one of the two linked peptides. In summary, the 8,301 PSMs confirm the observations made on the basis of 910 PSMs, suggesting that our analysis was not limited by sample size.

CONCLUSION

In this paper, we developed computational solutions to three experimental problems, building upon in-depth data mining of MS1 and MS2 properties of cross-linked peptides (1). The enrichment of cross-linked peptides is crucial to the success of cross-linking experiments. We show that focused acquisition can reach similar enrichment success for cross-linked peptides as chromatographic methods (2). Fragmentation spectra of cross-linked peptides contain fragments of two peptides. We find that fragments of the alpha peptide can be enriched through selection of the most intense peaks. Computationally, this parallels at least in part the use of

MS2-cleavable cross-linkers. A benefit of doing this computationally is not relying on cross-linker properties and thus potentially being universally applicable (3). Finally, fragmentation spectra of cross-linked peptides contain linear and cross-linked fragments. We show that cross-linked fragments have a distinguishable signal in CID (mass and charge). Thus, there is no need for labeling strategies to recognize cross-linked fragments. Our resulting search strategy sees the linearization of cross-linked fragments to collect enough evidence to extract candidates for one of the cross-linked peptides before the other, an approach that avoids the large search space of cross-linked peptides. In conclusion, computational approaches prove highly valuable in complementing experimental strategies in the endeavor of simplifying the identification of cross-linked peptides.

Acknowledgments—We thank Morten Rasmussen for initial work on this project and Marta Mendez for comments on the manuscript.

* The Wellcome Trust generously funded this work through a Senior Research Fellowship to J.R. (103139), a Centre core Grant (092076) and an instrument Grant (091020).

[S] This article contains supplemental material.

¶ To whom correspondence should be addressed: E-mail: juri.rappsilber@tu-berlin.de.

REFERENCES

1. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422,** 198–207
2. Yates, J. R., Ruse, C. I., and Nakorchevsky, A. (2009) Proteomics by mass spectrometry: Approaches, advances, and applications. *Annu. Rev. Biomed. Eng*. **11,** 49–79
3. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., and Kuster, B. (2007) Quantitative mass spectrometry in proteomics: A critical review. *Anal. Bioanal. Chem*. **389,** 1017–1031
4. Yin, S., and Loo, J. A. (2009) Mass spectrometry detection and characterization of noncovalent protein complexes. *Mass Spectrometry of Proteins and Peptides*, Clifton, NJ: Springer. pp 273–282
5. Sinz, A. (2006) Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein–protein interactions. *Mass Spectrom. Rev*. **25,** 663–682
6. Gingras, A.-C., Gstaiger, M., Raught, B., and Aebersold, R. (2007) Analysis of protein complexes using mass spectrometry. *Nat. Rev. Mol. Cell Biol*. **8,** 645–654
7. Petrotchenko, E. V., and Borchers, C. H. (2010) Crosslinking combined with mass spectrometry for structural proteomics. *Mass Spectrom. Rev.* **29,** 862–876
8. Rappsilber, J. (2011) The beginning of a beautiful friendship: cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. *J. Struct. Biol*. **173,** 530–540 doi: 10.1016/j.jsb.2010.10.014
9. Chen, Z. A., Jawhari, A., Fischer, L. Buchen, C., Tahir, S., Kamenski, T., Rassmussen, M., Lariviere, L, Bukowski-Will, J.-C., Nilges, M., Cramer, P., and Rappsilber, J. (2010) Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry. *EMBO J.* **29,** 717–726
10. Leitner, A., Reischl, R., Walzthoeni, T., Herzog, F., Bohn, S., Förster, F., and Aebersold, R. (2012) Expanding the chemical cross-linking toolbox by the use of multiple proteases and enrichment by size exclusion chromatography. *Mol. Cell. Proteomics* **11,** M111.014126
11. Müller, M. Q., Dreiocker, F., Ihling, C. H., Schäfer, M., and Sinz A. (2010) Cleavable cross-linker for protein structure analysis: Reliable identification of cross-linking products by tandem MS. *Anal. Chem*. **82,** 6958–68
12. Kao, A., Chiu, C., Vellucci, D., Yang, Y., Patel, V. R., Guan, S., Randall, A., Baldi, P., Rychnovsky, S. D., and Huang, L. (2011) Development of a novel cross-linking strategy for fast and accurate identification of cross-linked peptides of protein complexes. *Mol. Cell. Proteomics* **10,** M110.002212
13. Maiolica, A., Cittaro, D., Borsotti, D., Sennels, L., Ciferri, C., Tarricone, C., Musacchio, A., and Rappsilber, J. (2007) Structural analysis of multiprotein complexes by cross-linking, mass spectrometry, and database searching. *Mol. Cell. Proteomics* **6,** 2200–2211
14. Rinner, O., Seebacher, J., Walzthoeni, T., Mueller, L. N., Beck, M., Schmidt, A., Mueller, M., and Aebersold, R. (2008) Identification of cross-linked peptides from large sequence databases. *Nat. Methods* **5,** 315–318
15. Choi, S., Jeong, J., Na, S., Lee, H. S., Kim, H. Y., Lee, K. J., and Paek, E. (2010) New algorithm for the identification of intact disulfide linkages based on fragmentation characteristics in tandem mass spectra. *J. Proteome Res*. **9,** 626–635
16. Du, X., Chowdhury, S. M., Manes, N. P., Wu, S., Mayer, M. U., Adkins, J. N., Anderson, G. A., and Smith, R. D. (2011) Xlink-Identifier: An automated data analysis platform for confident identifications of chemically cross-linked peptides using tandem mass spectrometry. *J. Proteome Res*. **10,** 923–931
17. Yang, B., Wu, Y.-J., Zhu, M., Fan, S.-B., Lin, J., Zhang, K., Li, S., Chi, H., Li, Y.-X., Chen, H.-F., Luo, S.-K., Ding ,Y.-H., Wang, L.-H., Hao, Z., Xiu, L.-Y., Chen, S., Ye, K., He, S.-M., and Dong, M.-Q. (2012) Identification of cross-linked peptides from complex samples. *Nat. Methods* **9,** 904–906
18. Götze, M., Pettelkau, J., Schaks, S., Bosse, K., Ihling, C. H., Krauth, F., Fritzsche, R., Kühn, U., and Sinz, A. (2012) StavroX-a software for analyzing crosslinked products in protein interaction studies. *J. Am. Soc. Mass Spectrom*. **23,** 76–87
19. Holding, A. N., Lamers, M. H., Stephens, E., and Skehel, J. M. (2013). Hekate: software suite for the mass spectrometric analysis and three-dimensional visualization of cross-linked protein samples. *Journal of Proteome Research*, **12,** 5923–5933
20. Mayne, S. L. N., and Patterton, H.-G. (2011) Bioinformatics tools for the structural elucidation of multi-subunit protein complexes by mass spectrometric analysis of protein-protein cross-links. *Brief Bioinform*. **12,** 660–71
21. Rasmussen, M., Tahir, S., A. Chen, Z., and Rappsilber, J. (2008) Identification of cross-linked peptides in proteomic-scale experiments. Proc. 56 ASMS Conf. Mass Spectrom. Color, June 1–5
22. Singh, P., Shaffer, S. A., Scherl, A., Holman, C., Pfuetzner, R. A., Larson Freeman, T. J., Miller, S. I., Hernandez, P., Appel, R. D., and Goodlett, D. R. (2008) Characterization of protein cross-links via mass spectrometry and an open-modification search strategy. *Anal. Chem*. **80,** 8799–806
23. Chu, F., Baker, P. R., Burlingame, A. L., and Chalkley, R. J. (2010) Finding chimeras: A bioinformatics strategy for identification of cross-linked peptides. *Mol. Cell. Proteomics* **9,** 25–31
24. Wang, J., Anania, V. G., Knott, J., Rush, J., Lill, J. R., Bourne, P. E., and Bandeira, N. (2014). Combinatorial approach for large-scale identification of linked peptides from tandem mass spectrometry spectra. *Molecular & Cellular Proteomics* **13,** 1128–1136
25. Blencowe, A., and Hayes, W. (2005) Development and application of diazirines in biological and synthetic macromolecular systems. *Soft Matter* **1,** 178
26. Tabb, D. L., Smith, L. L., Breci, L. A., Wysocki, V. H., Lin, D., and Yates, J. R. 3rd (2003) Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Anal. Chem.* **75,** 1155–1163
27. Waldera-Lupa, D. M., Stefanski, A., Meyer, H. E., and Stühler, K. (2013) The fate of b-ions in the two worlds of collision-induced dissociation. *Biochim. Biophys. Acta* **1834,** 2843–2848
28. Degroeve, S., and Martens, L. (2013) MS2PIP: A tool for MS/MS peak intensity prediction. *Bioinformatics* **29,** 3199–3203
29. Li, W., O'Neill, H. A., and Wysocki, V. H. (2012) SQID-XLink: Implementation of an intensity-incorporated algorithm for cross-linked peptide identification. *Bioinformatics* **28,** 2548–2550
30. Trnka, M. J., Baker, P. R., Robinson, P. J. J., Burlingame, a L., and Chalkley, R. J. (2014). Matching cross-linked peptide spectra: only as good as the worse identification. *Molecular & Cellular Proteomics* **13,** 420–434
31. Fischer, L., Chen, Z. A., and Rappsilber, J. (2013) Quantitative cross-linking/mass spectrometry using isotope-labelled cross-linkers. *J. Proteomics* **88,** 120–128

32. Barysz, H., Kim, J. H., Chen, Z. A., Hudson, D. F., Rappsilber, J., Gerloff, D. L., and Earnshaw, W. C. (2015) Three-dimensional topology of the SMC2/SMC4 subcomplex from chicken condensin I revealed by cross-linking and molecular modelling. *Open Biol.* **5,** 150005
33. Trowitzsch, S., Viola, C., Scheer, E., Conic, S., Chavant, V., Fournier, M., . . . Berger, I. (2015). Cytoplasmic TAF2-TAF8-TAF10 complex provides evidence for nuclear holo-TFIID assembly from preformed submodules. *Nature Communications* **6,** 6011
34. Kustatscher, G., Hégarat, N., Wills, K. L. H., Furlan, C., Bukowski-Wills, J. C., Hochegger, H., and Rappsilber, J. (2014) Proteomics of a fuzzy organelle: Interphase chromatin. *EMBO J.* **33,** 648–664
35. Yen, C.-Y., Houel, S., Ahn, N. G., and Old, W. M. (2011) Spectrum-to-spectrum searching using a proteome-wide spectral library. *Mol. Cell. Proteomics* **10,** M111.007666
36. Belsom, A., Schneider, M., Fischer, L., Brock, O., and Rappsilber, J. (2015) Serum albumin domain structures in human blood serum by mass spectrometry and computational biology. *Mol. Cell. Proteomics* **3,** 1105–1116
37. 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., and Rappsilber, J. (2010) Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry. *EMBO J.* **29,** 717–726
38. Braun, N., Zacharias, M., Peschek, J., Kastenmuller, A., Zou, J., Hanzlik, M., Haslbeck, M., Rappsilber, J., Buchner, J., and Weinkauf, S. (2011) Multiple molecular architectures of the eye lens chaperone B-crystallin elucidated by a triple hybrid approach. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 20491–20496
39. Soares, D. C., Bradshaw, N. J., Zou, J., Kennaway, C. K., Hamilton, R. S., Chen, Z. A., Wear, M. A., Blackburn, E. A., Bramham, J., Böttcher, B., Millar, J. K., Barlow, P. N., Walkinshaw, M. D., Rappsilber, J., and Porteous, D. J. (2012) The mitosis and neurodevelopment proteins NDE1 and NDEl1 form dimers, tetramers, and polymers with a folded back structure in solution. *J. Biol. Chem*. **287,** 32381–32393
40. Lauber, M. A., Rappsilber, J., and Reilly, J. P. (2012). Dynamics of ribosomal protein S1 on a bacterial ribosome with cross-linking and mass spectrometry. *Molecular & Cellular Proteomics* **11,** 1965–1976
41. Fischer, L., Chen, Z. A., and Rappsilber, J. (2013). Quantitative cross-linking/mass spectrometry using isotope-labelled cross-linkers. *Journal of Proteomics*, **88,** 120–128
42. Miell, M. D. D., Fuller, C. J., Guse, A., Barysz, H. M., Downes, A., Owen-Hughes, T., Rappsilber, J., Straight, A. F., and Allshire, R. C. (2013) CENP-A confers a reduction in height on octameric nucleosomes. *Nat. Struct. Mol. Biol.* **20,** 763–765
43. Abad, M. A., Medina, B., Santamaria, A., Zou, J., Plasberg-Hill, C., Madhumalar, A., Jayachandran, U., Redli, P. M., Rappsilber, J., Nigg, E. A., and Jeyaprakash, A. A. (2014) Structural basis for microtubule recognition by the human kinetochore Ska complex. *Nat. Commun.* **5,** 2964
44. Kostan, J., Salzer, U., Orlova, A., Toro, I., Hodnik, V., Senju, Y., Zou, J., Schreiner, C., Steiner, J., Merilainen, J., Nikki, M., Virtanen, I., Carugo, O., Rappsilber, J., Lappalainen, P., Lehto, V.-P., Anderluh, G., Egelman, E. H., and Djinovic-Carugo, K. (2014) Direct interaction of actin filaments with F-BAR protein pacsin2. *EMBO Rep.* **15,** 1154–1162

Chapter 3

Manuscript 2. Optimized fragmentation regime for diazirine photo-cross-linked peptides

analytical chemistry

Article

pubs.acs.org/ac

Optimized Fragmentation Regime for Diazirine Photo-Cross-Linked Peptides

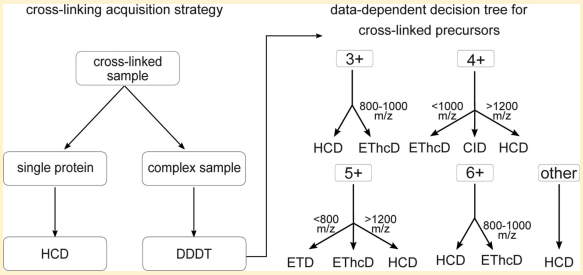
Sven H. Giese,[†,‡] Adam Belsom,[‡] and Juri Rappsilber*[,†,‡]

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

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

Supporting Information

ABSTRACT: Cross-linking/mass spectrometry has evolved into a robust technology that reveals structural insights into proteins and protein complexes. We leverage a new tribrid instrument with improved fragmentation capacities in a systematic comparison to identify which fragmentation method would be best for the identification of cross-linked peptides. Specifically, we explored three fragmentation methods and two combinations: collision-induced dissociation (CID), beam-type CID (HCD), electron-transfer dissociation (ETD), ETciD, and EThcD. Trypsin-digested, SDA-cross-linked human serum albumin (HSA) served as a test sample, yielding over all methods and in triplicate analysis in total 2602 matched PSMs and 1390 linked residue pairs at 5% false discovery rate, as confirmed by the crystal structure. HCD wins in number of matched peptide-spectrum-matches (958 PSMs) and identified links (446). CID is most complementary, increasing the number of identified links by 13% (58 links). HCD wins together with EThcD in cross-link site calling precision, with approximately 62% of sites having adjacent backbone cleavages that unambiguously locate the link in both peptides, without assuming any cross-linker preference for amino acids. Overall quality of spectra, as judged by sequence coverage of both peptides, is best for EThcD for the majority of peptides. Sequence coverage might be of particular importance for complex samples, for which we propose a data dependent decision tree, else HCD is the method of choice. The mass spectrometric raw data has been deposited in PRIDE (PXD003737).



Current methods of structural biology have left a systematic and large gap in our knowledge of protein structures.[1] Cross-linking/mass spectrometry (CLMS) is an emerging tool that helps to gain structural information for challenging proteins and protein complexes. In CLMS experiments, protein complexes are chemically cross-linked, digested into peptides, and then analyzed via mass spectrometry and bioinformatics.[2−5] Identifying a cross-linked peptide pair or the linked residues within, defines their maximal distance in the folded protein. The derived distance constraints can then be used to determine the low-resolution arrangement of protein complexes[4,6,7] or even the high-resolution structure of a protein by the help of computational modeling.[8]

To identify cross-linked peptides, fragmentation spectra have to be matched with peptide sequences by database search. For this purpose, a number of tools have been developed,[9,10] for example, pLINK,[11] Protein Prospector,[12,13] StavroX,[14] xQuest,[15] Kojak,[16] Xi,[6,17] or even search engines[18] based on linear peptide identification search paradigms such as Mascot.[19] One of the challenges in identifying cross-linked peptides is the unequal fragmentation of the two linked peptides,[13,17] that is, often one of the two peptides is better fragmented and thus also better characterized by fragment ions. Under collision-induced dissociation (CID) conditions this has been investigated in more detail, revealing that the intensity of observed fragment ions is also affected.[17] This is important for the scoring of cross-linked peptides since in general the number of identified fragment ions and their intensity is used for spectra matching. Despite the obvious disadvantage of the unequal fragmentation, scoring mechanisms managed to successfully exploit this fact: To judge the complete cross-linked peptide-spectrum match (PSM), the two individual peptide scores are weighted differently.[13,16] However, this should only be an ad hoc solution; ideally the experimental setup can be changed in such a way that the sequence coverage for both peptides is increased. It is plausible that one of the available fragmentation methods performs better than the others, and a comparative analysis into the behavior of cross-linked peptides might reveal options for a refined acquisition strategy.

Throughout the manuscript we use CID for resonant excitation CID in the linear ion trap and HCD as the abbreviation for beam-type CID (HCD is also often referred to as higher-energy collisional dissociation). CID is one of the standard methods of fragmenting peptides in proteomics and has been used in many CLMS studies.[6,20−24] The details of CID of cross-linked peptides have recently been systematically

Received: May 27, 2016
Accepted: July 25, 2016
Published: July 25, 2016

assessed,[17] but a systematic comparison to other fragmentation methods such as HCD is lacking. HCD has also been used in many CLMS studies.[11,13,25,26] Neither a systematic analysis of cross-linked peptides under HCD exists nor under electron-transfer dissociation (ETD). ETD-based fragmentation, that is, ETD with and without supplemental activation of CID (ETciD) or HCD (EThcD)[27] has neither routinely been applied to cross-linked peptides nor investigated in much detail. A sequential fragmentation scheme of CID and ETD is reported to increase the identification and confidence levels of cross-linked peptides.[28] Another study acquired sequential CID and ETD fragmentation spectra as an optimized method for CID cleavable cross-linkers with signature peaks. Both spectra are then matched with their appropriate ion types and scored together, yielding an improved sequence coverage compared to CID alone.[29] Search strategies for noncleavable cross-linkers, however, do not rely on the detection of signature peaks, and thus, the time for the reisolation of the precursor can be saved by simply using ETD with supplemental activation. It was also shown that ETD alone can generate good ion coverage for both peptides using a novel cross-linker,[30] albeit the effect on peptides cross-linked with another cross-linker remains to be investigated. In contrast, ETD has been used frequently for complete proteins[31] or to characterize post-translational modifications (since it leaves the often labile peptide modifications intact[32]). Earlier studies stated that ETD fragment peptides with charge states higher than two, more extensively than CID.[33] However, the underlying effect seems to correlate with the mass-to-charge ratio (m/z) of the precursors.[34] For cross-linked peptides, we expect highly charged precursors[6,15,18] and, thus, potentially well-suited targets for ETD.

High-sequence coverage is important to ensure selectivity during database search when trying to identify the two cross-linked peptides from the large choice of alternatives offered by the database. Good backbone fragmentation should also be beneficial to pinpoint the exact location of the cross-link. Despite being the intuitive expectation, sequence coverage and site calling precision do not necessarily have to be linked directly. Properties of the linkage site might direct fragmentation toward neighboring backbone bonds or away from them. Also, for amine-reactive cross-linkers, pinpointing the exact position of the cross-link is assisted by the restricted chemical reactivity toward lysine, serine, threonine, tyrosine, or the protein N-terminus. Hence, depending on the peptide sequence there might only be a single amino acid amenable to the cross-linker reaction. For highly reactive cross-linkers such as succinimidyl 4,4-azipentanoate (SDA) each residue in a peptide needs to be considered when locating the linkage site. Therefore, pinpointing the cross-link sites potentially requires more complete backbone fragmentation than for more specific cross-linkers.

In this study we compared three different fragmentation techniques and two combined fragmentation schemes available on a novel tribrid mass spectrometer (Orbitrap Fusion Lumos, Thermo Fisher Scientific), CID, HCD, ETD, ETciD, and EThcD, on cross-linked peptides obtained by tryptic cleavage of SDA-cross-linked human serum albumin (HSA). The three-dimensional structure of HSA has been resolved by X-ray crystallography[35] and is used as ground-truth to evaluate the identification results. The right choice of fragmentation method allows the number of identified linkage sites to be increased; increasing the sequence coverage of both linked peptides boosts the confidence of the matches and also the correct localization of the cross-link site.

METHODS

Sample Preparation. Purified HSA (Sigma-Aldrich, St. Louis, MO) was cross-linked using different cross-linker-to-protein, weight-to-weight (w/w) ratios: 0.152:1, 0.203:1, 0.303:1, 0.406:1, 0.606:1, 0.811:1, 1.21:1, and 1.62:1. Aliquots of purified HSA (15 μg, 0.75 mg/mL) in cross-linking buffer (20 mM HEPES−OH, 20 mM NaCl, 5 mM $MgCl_2$, pH 7.8) were mixed with sulfo-SDA (Thermo Scientific Pierce, Rockford, IL) to initiate incomplete reaction of the protein with the sulfo-NHS ester component of the cross-linker. Human blood serum from a healthy donor (20 μg, 1.0 mg/mL) was cross-linked in a similar manner, using cross-linker-to-protein ratios (w/w) of 0.5:1, 1:1, 2:1, and 4:1. Total reaction volume in each case was 30 μL. For the second step of the cross-linking procedure, photoactivation of the diazirine group was carried out using UV irradiation from a UVP CL-1000 UV Cross-linker (UVP Inc.). Samples were irradiated for either 25 or 50 min for purified HSA samples, and either 10, 20, 40, or 60 min in the case of blood serum samples and separated using gel electrophoresis. Bands corresponding to monomeric HSA were excised from gels and the proteins reduced with DTT, alkylated using IAA, and digested using trypsin following standard protocols.[18] Peptides were loaded onto self-made C18 StageTips[36] and eluted using 80% acetonitrile and 20%, 0.1% TFA in water. The eluates from blood serum HSA and purified HSA digests were mixed 0.33:1 as a master mix to be used throughout this study. The two samples originally used in our structural analysis of HSA[8] were mixed here to gain enough material to perform the experiments of this study in triplicates.

Data Acquisition. Peptides were loaded directly (2% B, 500 nL/min) onto a spray emitter analytical column (75 μm inner diameter, 8 μm opening, 250 mm length; New Objectives) packed with C18 material (ReproSil-Pur C18-AQ 3 μm; Dr Maisch GmbH, Ammerbuch-Entringen, Germany) using an air pressure pump (Proxeon Biosystems).[37] The 0.1% formic acid served as mobile phase A and 0.1% formic acid/80% acetonitrile as mobile phase B. Peptides were eluted (200 nL/min, linear gradient of 2−40% B over 139 min) directly into an Orbitrap Fusion Lumos Tribrid mass spectrometer (Thermo Fisher Scientific, San Jose, CA). Survey spectra were recorded in the Orbitrap at 120000 resolution. Spectra for all fragmentation methods were acquired using a scan range of 300−1700 m/z. Precursor ion isolation was performed with the quadrupole and an m/z window of 1.6 Th. The precursor automatic gain control (AGC) target value was 4×10^5, maximum injection time 50 ms. For CID only, CID collision energy was set to 30%. For HCD only, HCD collision energy was set to 35%. For ETD only, the option to inject ions for all available parallelizable time was selected (anion AGC 5×10^4, 60 ms maximum injection time). Supplemental activation (SA) collision energy was set to 10% for ETciD, and 25% for EThcD.

Data Analysis. Raw files were preprocessed with MaxQuant (v. 1.5.2.8) with "Top MS/MS peaks per 100 Da" set to 100.[38] Resulting peak files (APL format) were subjected to Xi (ERI Edinburgh, v. 1.5.584) and searched with the following settings: MS accuracy, 6 ppm; MS/MS accuracy, 20 ppm; enzyme, trypsin; max. missed cleavages, 4; max. number of modifications, 3; fixed modification, none; variable modifications, carbamidomethylation on cysteine; oxidation on methionine; cross-linker, SDA (mass modification: 109.0396 Da). In

DOI: 10.1021/acs.analchem.6b02082
Anal. Chem. 2016, 88, 8239−8247

addition, variable modifications by the hydrolyzed cross-linker ("SDA-hyd", mass modification: 82.0413 Da) and loop-links ("SDA-loop", mass modification: 83.0491 Da) were allowed. SDA cross-link reactions were assumed to connect lysine, serine, threonine, tyrosine, or the protein N-terminus on the one end of the spacer with any other amino acid on the other end. FDR was estimated using XiFDR (v. 1.0.6.14)[39] on a 5% peptide spectrum match (PSM) level and 5% link-level only including unique PSMs. The reference database consisted of a single entry with the protein sequence of HSA (Uniprot: P02768). For further analysis, PSM information (precursor m/z, annotated fragments, score, peptide sequences, etc.) were extracted from a local PostgreSQL database. The annotated spectra are available in the Supporting Information (Figures S4–S8).

To derive a decision tree for an optimized fragmentation scheme for cross-linked peptides we divided the acquisition range into a grid of m/z bins of size 200 for each charge state from 3 to 7. After sorting all PSMs into this theoretical grid we assigned each cell the best performing and second best performing fragmentation method. The performance was assessed through the median achieved sequence coverage of the complete cross-linked peptide. Note, sequence coverage does not depend on the possible fragment ions but rather on the actual evidence (fragment ions) for specific n-terminal or c-terminal sequences. To decide whether or not a fragmentation method is favorable over another we conducted a simple, one-sided permutation test[40] with label swaps and 10.000 iterations. P values lower than 0.05 were regarded as significant. Permutation tests were only performed if more than 15 observations were in the best performing class. If the best and second best were too similar to give significant results the best performing method was also compared to all other methods.

All raw files are available via the PRIDE repository[41] (PDX: PXD003737) along with PSM results and the reference FASTA.

RESULTS AND DISCUSSION

We investigated the impact of five fragmentation techniques (CID, HCD, ETD, EThcD, ETciD) on the analysis of cross-linked peptides using a latest generation Orbitrap mass spectrometer (Orbitrap Fusion Lumos, Thermo Fisher Scientific). HSA was used as a model protein with a known crystal structure. Cross-linking experiments suffer under CID conditions from the underrepresentation of fragment ions from one of the two peptides.[13,17] Here we define the peptide with more intense ions among the ten most intense fragment ions as the α-peptide and the remaining peptide as the β-peptide.[17] Note that the nomenclature for the two peptides in a cross-link is not standardized; other definitions using the achieved search score[13] or the peptide's chain length or mass[4] are used. We hypothesized that the usage of other fragmentation techniques has an impact on the fragmentation pattern of cross-linked peptides and subsequently on the success rate of identification. In our analysis we applied two different FDR-levels according to the descriptive features that we evaluated.[39] For the evaluation of identification results on the crystal structure, a link-level FDR is used. For the evaluation of PSM properties (e.g., sequence coverage), a regular PSM FDR is used. An overview is available in Table S1.

HCD Fragmentation Gives the Highest Number of Identified Cross-Links. We compared the number of identified cross-links that passed a 5% link-level FDR and a 5% PSM-level FDR to assess which fragmentation approach leads to the highest identification success. The results, accumulating the three technical replicates for all fragmentation techniques, show that HCD (958 PSMs) gives the highest number of identifications followed by CID (604 PSMs, Figure 1A). ETciD fragmentation achieves the lowest number of identified cross-links with 296 PSMs. This order is closely related to the number of acquired spectra in all replicates. While HCD is the fastest acquisition technique producing ~109000 MS2 spectra ETciD and ETD only produce ~80000 spectra (Table S1). While the number of PSMs is only a proxy for the success of CLMS experiments, the true value of CLMS data comes from the corresponding distance constraints. Therefore, for the comparison of cross-linking data it makes sense to compare the results on the link-level. For the comparison on the link-level only unique links are regarded for further analysis. A unique link is defined by the combination of residues involved in a cross-link, that is, a unique residue pair.

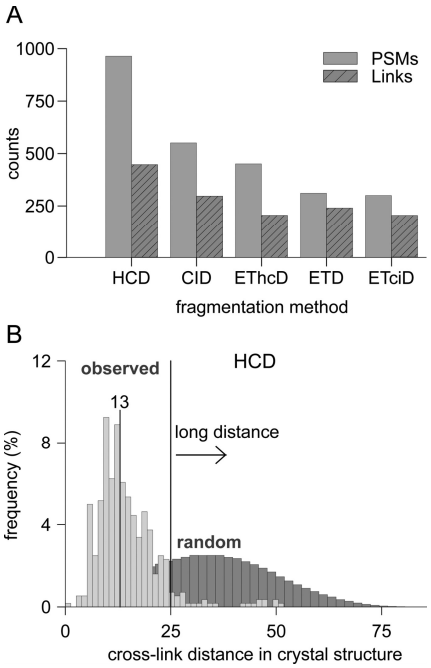


Figure 1. Number of SDA-induced cross-links identified in HSA using different fragmentation techniques. (A) Identified PSMs and links were computed for 5% FDR-level on the respective category. (B) Evaluation of the identified cross-links against the crystal structure of HSA. The light gray distribution reflects the distance measurement between identified residues in a cross-link mapped to the crystal structure (the median is shown above the vertical line). The dark gray distribution reflects all pairwise combinations of cross-linkable residues in the crystal structure. The black vertical line at 25 Å is used to classify cross-links as long distance or not.

As is the case for PSMs, HCD fragmentation also returns the highest number of identified links (Figure 1A). In total 1390 links (972 unique) were identified with the various methods: Of the unique links HCD observed 446 links (46%), CID 297 links (31%), EThCD 240 links (25%), ETciD 205 links (21%), and ETD 202 (21%). Note, the comparison of the links is not straightforward if the cross-link site is ambiguous. We applied a simple heuristic that assigns the linkage site to the c-terminal

DOI: 10.1021/acs.analchem.6b02082
Anal. Chem. 2016, 88, 8239–8247

residue in ambiguous linkage windows. As HSA's three-dimensional structure has been resolved, it is possible to utilize it as ground truth and further evaluate the quality of the identified links. We used 25 Å here for SDA as the maximal α-carbon distance of two linkable amino acids in the three-dimensional structure. This provides a clear distinction between true positive and false positive identifications. Each identified link that lies within 25 Å in the crystal structure is plausibly a true positive. Accordingly, every link that is further than 25 Å apart is plausibly a false positive. This is a simplified approach, as links shorter than 25 Å will also contain false positives as a result of random matching, and conversely, longer links may be true and result from protein structural flexibility. Comparing the link information from all five fragmentation techniques shows that the overall quality of the results is comparable across all fragmentation modes and distinctly different from random results. The derived distance distributions have a median of 12−13 Å and are very distinct from the random distance distribution (Figure 1B). In addition, the results are comparable in meeting the approximated 5% FDR. FDR analysis for the HCD data and the ETciD data slightly underestimates the number of false positives by 1% and 2.5%, respectively (Figure S1). These can be partially explained by the definition of the FDR itself, which only gives an approximation of the true false discovery rate. Furthermore, the hard cutoff that was used has a large impact on the computed FDR. For example, the ETciD distances showed a larger peak just to the right of the desired distance cutoff, indicating that a small increase in the maximal allowed distance would give an FDR closer to the desired 5%. The HCD distance distribution looks similar to a small enrichment of false positives just outside the maximal allowed distance. Thus, accounting for more flexibility would change the FDR and suggests that the different methods lead to data of comparable quality but different quantity.

Having a preranking of the individual fragmentation techniques in terms of number of PSMs and unique cross-links is desirable to maximize the information content in a single run. Depending on the peptide properties, some fragmentation methods might be more suited for a certain group of peptides, and thus, using two (or more) orthogonal fragmentation techniques may increase the overall yield in peptide identifications and thus distance constraints. Disregarding the link information to focus first on the identified peptide pairs shows that HCD fragmentation also yields the largest number of unique peptide pairs (Figure 2A). A total of 43% (201 peptide pairs) are shared between at least two fragmentation techniques. The remaining 57% (269 peptide pairs) are unique to one of the five fragmentation techniques. To maximize the information content, HCD should be combined with CID fragmentation to increase the number of unique links by 58 (Figure 2B). Interestingly, ETD fragmentation can maximally increase the number of unique links by 41 by using EThcD. We suspect that the difference in the number of acquired spectra and actually identified PSMs is the main driver for this effect. We define the identification rate, IR, as IR $= \frac{N_{id}}{N_{acq}}$, where N_{id} is the number of identified unique PSMs and N_{acq} is the total number of acquired MS2 spectra (Table S1). The IR reveals that HCD not only acquires most spectra, but also has the highest success rate of 0.88% compared to CID (0.61%), EThcD (0.52%), and ETD/ETciD (0.38%). If speed and reliability of ETD-based fragmentation should change in the future, this order of complementarity may change. In comparison with linear peptide identifications, where the IR reaches up to 54%[42] (depending on the instrumentation), the success rate of cross-link identification is much lower. A contributing factor will be the generally low abundance of cross-linked peptides when compared to linear peptides, which will reduce their frequency of selection for MS2, especially in competition with the linear peptides also present. Other factors will include poorer database matching due to often lower intensity, but also more complex spectra and a larger search space.

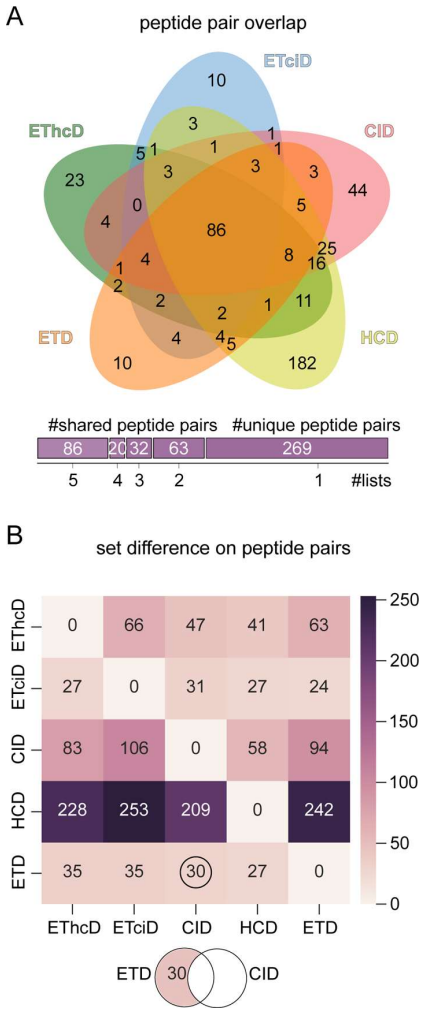


Figure 2. Pairwise result overlaps of fragmentation techniques. (A) Overlap of identified peptide pairs (disregarding link-site positions) between fragmentation techniques (Venn diagram generated with Jvenn[49]). (B) Set difference matrix shows the number of uniquely identified peptide pairs (disregarding link-site positions) by one fragmentation technique (*y*-axis) when compared to another one (*x*-axis).

ETD-Aided Fragmentation Improves the Coverage of the Second Peptide. The identification of cross-linked peptides poses two challenges: First, finding the correct peptide pair, and second, assigning the correct cross-link site. High peptide sequence coverage for both individual peptides should be beneficial to assigning the correct site. Site calling will be

DOI: 10.1021/acs.analchem.6b02082
Anal. Chem. 2016, 88, 8239−8247

especially challenging when considering cross-linkers such as SDA, where the number of cross-link target sites is large.

Under HCD conditions the coverage distribution for the α-peptide is the lowest, with a mean coverage around 50% (Figure 3A). The other four fragmentation techniques perform very similar to only small improvements in the coverage of the α-peptide with CID or EThcD fragmentation. Interestingly, ETD involving fragmentation schemes do not increase the fragmentation efficiency (measured by the peptide coverage) much for the α-peptide. In fact, the highest coverage values for the α-peptide were observed with CID fragmentation. In contrast, the sequence coverage for the beta peptide largely depends on the fragmentation method (Figure 3B). ETD, ETciD, EThcD, and HCD show a much better fragmentation compared to CID. Previously, ETD was reported to improve the sequence coverage compared to CID.[32,43] We observe here that for cross-linked peptides this effect is very pronounced for the β-peptide, but not for the α-peptide.

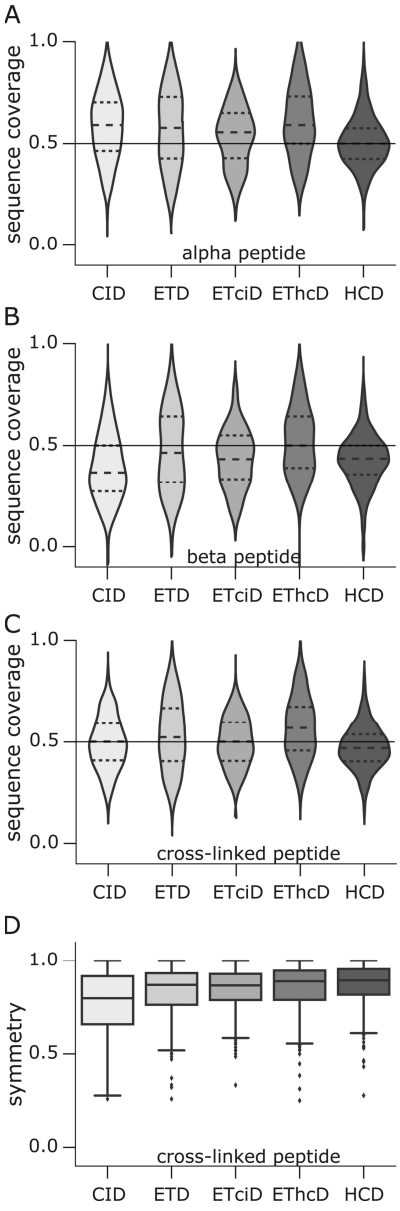


Figure 3. Achieved sequence coverage comparison. Coverage distribution of the α-peptide (A; more matches among the 10 most intense fragment ions) and the β-peptide (B). The vertical line in (A)−(C) reflects a reference value of 50% sequence coverage, meaning fragments (b, c, y, or z) match to half of the backbone links between residues along the sequence of the peptide. (C) Coverage distribution for the complete cross-linked peptide. (D) Symmetry (absolute coverage difference between alpha and beta peptide) distributions for the different fragmentation techniques. The data in (A)−(D) were analyzed using a 5% PSM FDR.

In general, in cross-linked peptides, one peptide matches more and with higher intense fragment ions than the other. All fragmentation methods yield at least an average coverage of around 50% for the α-peptide. For the β-peptide, the average coverage lies between 39% and 50%. CID would be the method of choice for high α-peptide coverage. However, CID is systematically disadvantaging the β-peptide. For the β-peptide, the other fragmentation methods perform much better: EThcD and HCD almost reach the same fragmentation efficiency as for the α-peptide. In numbers, the largest discrepancy between α- and β-peptide coverage was observed with CID, with a mean coverage difference (MCD) of 19%. EThCD and HCD show the lowest MCD of 8%. The overall best coverage is observed with EThcD fragmentation (Figure 3C). ETciD seems to be less effective, presumably as ETD in the first stage leads to charge reduction, and CID then fragments a single precursor, while HCD fragments all. Nevertheless, ETciD greatly improves the coverage of the second peptide when compared to CID.

To compare the fragmentation efficiency on both peptides in a cross-link more systematically, we define the symmetry factor (SF) as

$$\mathrm{SF} = |\mathrm{cov}_\alpha - \mathrm{cov}_\beta| \quad (1)$$

where cov_α and cov_β refer to the sequence coverage of the α- and β-beta peptide, respectively. For convenience, we use the negation SF′ of SF defined as

$$\mathrm{SF}' = 1 - \mathrm{SF} \quad (2)$$

A large SF′ means that α- and β-peptide coverage are very similar and vice versa. CID shows the smallest among the five fragmentation methods of ∼0.8. The other four methods perform better than CID, with a median of ∼0.9 (Figure 3D). In addition, ETD, ETciD, EThcD, and HCD have a smaller spread than CID. In summary, CID exasperates the second peptide problem. Nevertheless, CID still slightly outperforms HCD in overall cross-linked peptide sequence coverage. In order to maximize overall cross-linked peptide coverage ETD, ETciD, and EThcD are recommended, based on median coverage of the complete cross-linked peptide.

Precursor *m/z* Has a Large Effect on the Efficiency of the Fragmentation. To follow-up on the different fragmentation behavior of cross-linked peptides we investigated how the precursor properties influence the fragmentation efficiency. We first divided the m/z acquisition range into bins of m/z 150 (starting from m/z 550). For each bin we then collected the peptide identifications of all different fragmentation methods and investigated the sequence coverage based on the m/z of the precursor.

ETD and EThcD lead to the highest sequence coverage between m/z 500−800 (Figure 4A,B). However, ETD

DOI: 10.1021/acs.analchem.6b02082
Anal. Chem. 2016, 88, 8239−8247

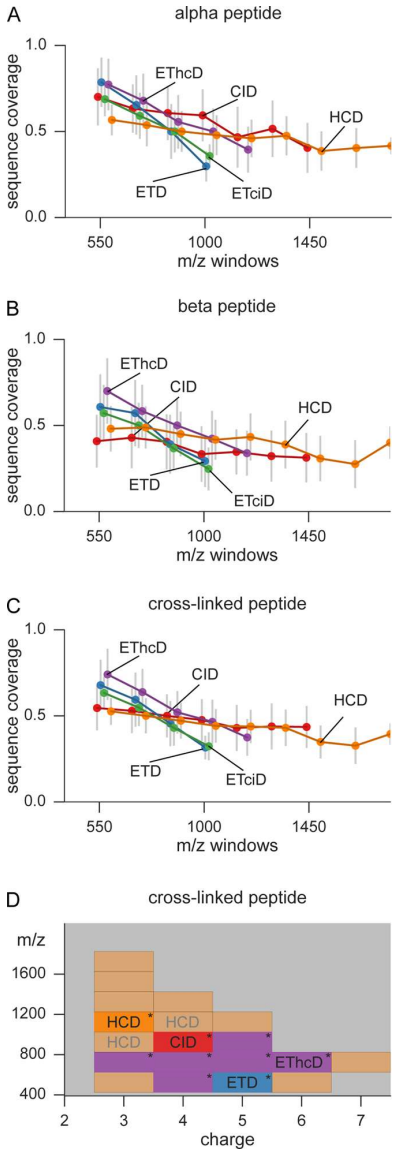


Figure 4. Sequence coverage depending on precursor m/z and charge. The average coverage values from (A) α-peptides, (B) β-peptides, and (C) the complete cross-linked peptide are plotted vs the precursor m/z. Each dot represents the median of all identified peptides in a window of m/z 150. Error bars show the standard deviation. (D) Decision surface to optimize the sequence coverage of cross-linked peptide. The acquisition range was divided into bins of 200 m/z per charge state. In each bin the best performing fragmentation method (judged by median achieved sequence coverage) is used to color that particular bin. The "*" denotes a significant improvement in sequence coverage by using the best performing fragmentation method over the second best. Areas with less than 15 observations are colored in light red, falling back to HCD as standard fragmentation technique. Gray annotations show areas where no significant improvement could be obtained by choosing one method over the others.

efficiency decreases steeply with higher m/z, making HCD and CID the better choice for precursors larger than m/z 1000. The same trend is observed for all ETD-based methods. These differences are more pronounced on the individual α- and β-peptides. When the complete peptide coverage is compared (Figure 4C), all methods stick more closely together but EThcD and ETD still outperform all other methods for precursors smaller than m/z 850. In higher m/z areas, only CID and HCD are able to still produce enough peptide identifications (data in the figure was limited to only include m/z bins with at least five observations).

As demonstrated in the sections above, there are differences in the efficiency of the fragmentation of cross-linked peptides. In a more detailed comparison, we divided the acquisition range into a grid made of charge bins of size one and m/z bins of size 200. In each of these cells we then tested how well the five different fragmentation methods performed. The performance was evaluated on the cross-linked peptide sequence coverage. For the majority of peptides, EThcD achieved significantly higher sequence coverage (Figure 4D) than the second best method between 600 and 800 m/z (precursor charge 3−6). In addition, the m/z cells 400−600 (z = 4) and 800−1000 (z = 5) are also favored by EThcD fragmentation. Since the majority of cross-linked peptides (71%) lie within 600−1000 m/z, the most important area is dominated by EThcD fragmentation. However, evaluated by pure numbers of identifications, EThcD is not the best performing method. On average, ∼35 PSMs are missed if EThcD is chosen over the method that achieves the highest number of identifications. If the evaluation metric is changed to the highest number of identifications, HCD is outperforming the other fragmentation methods for all m/z bins (Figure S3). Therefore, HCD was selected as default method for regions where no significant improvement could be observed by any of the other methods (Figure 4D, HCD written in gray).

HCD, EThcD, and ETD fragmentation define the cross-link site most unambiguously. The overall sequence coverage is a valuable feature to assess the quality of peptide identifications. However, for cross-linked peptides those fragments flanking the cross-linked residues are important to define the linkage site. This resembles the localization of post-translational modifications such as phosphorylation, which greatly benefited from the usage of combined fragmentation methods.[44] Limited information about the cross-link site is available when none of the fragments next to a cross-linked residue are observed; the cross-link site can then only be assigned by prior assumptions or to larger sequence windows, which becomes problematic if the site call is off by ±5 residues (at least in HSA and using current ab initio structure computation).[8] Given the information from correct fragment identifications, a combination of one c-terminal and one n-terminal ion is enough to locate the cross-link site unambiguously. Utilizing the high-resolution/accurate mass measurement in our experimental design, we thus assumed that each assigned fragment is correct for peptides passing the specified FDR.

The cross-link site in α-peptides could be assigned to a single residue in ∼65% of all PSMs identified with EThcD or HCD (Figure 5A). The second best performing method was ETD, with approximately 60% of PSMs where the cross-link could be assigned to a single residue. CID and ETciD PSMs show the lowest number of accurate site localizations to a single residue (below 50% of all PSMs). All methods placed the cross-link site on average within the critical 5 residue window for 97.2% ± 1.17 (α-peptides) and 95.6% ± 1.3 (β-peptides) of all PSMs. For the β-peptide, this looks very similar; EThcD and HCD show the best fragmentation behavior to localize the cross-link

DOI: 10.1021/acs.analchem.6b02082
Anal. Chem. 2016, 88, 8239−8247

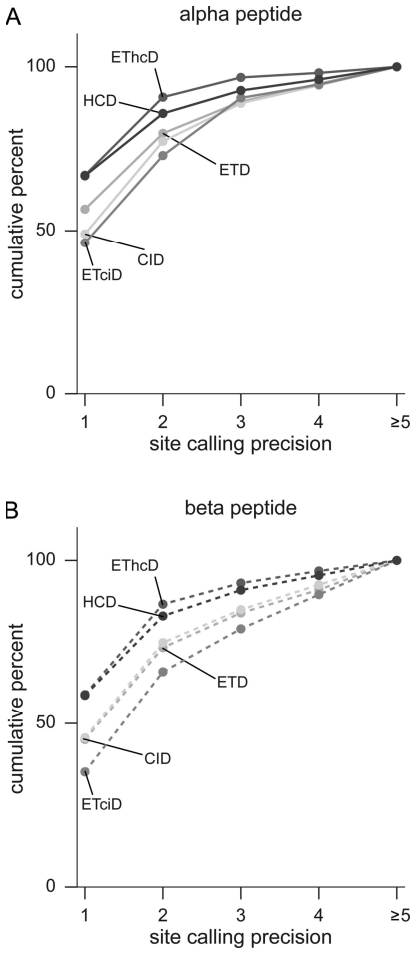


Figure 5. Cross-link site localization precision. (A) Cumulative precision curve for the α-peptide. (B) Cumulative precision curve for the β-peptide. With a precision value of one the cross-link site is unambiguously located by adjacent backbone fragments (b, c, y, or z) in the peptide. A value of two limits the cross-link site to two eligible residues.

site (Figure 5B). With approximately 50% of precisely localized cross-links in the β-peptide, the link-localization is less well for the β-peptide than for the α-peptide. However, this is not as pronounced as would be expected from the sequence coverage asymmetry. This is counterintuitive since the coverage distributions for HCD is among the lowest of all five fragmentation techniques for the β-peptide. For EThcD, the results for the determination of the cross-link site are more in line with the observed coverage distributions. Still, the large difference in the coverage distribution of the α- and β-peptides seems not to be as pronounced for the distribution of correct localizations of the cross-link site. One of the possible reasons is that the cleavage of the peptides before and after the cross-link site is preferred. For CID a statistical trend was reported that cross-linked fragments outnumber linear fragments and tend to have a higher intensity.[17] We encounter the opposite for HCD, linear fragments visibly outnumber cross-linked fragments (Figure S8).

Data-Dependent Decision Tree for Optimized Acquisition of Cross-Linked Peptides. CLMS studies vary in the degree of complexity: single proteins, multiple protein complexes or complete proteomes can be analyzed to generate protein−protein interaction information or the three-dimensional structure. Depending on the specific case we propose two different acquisition strategies (Figure 6A): First, for single

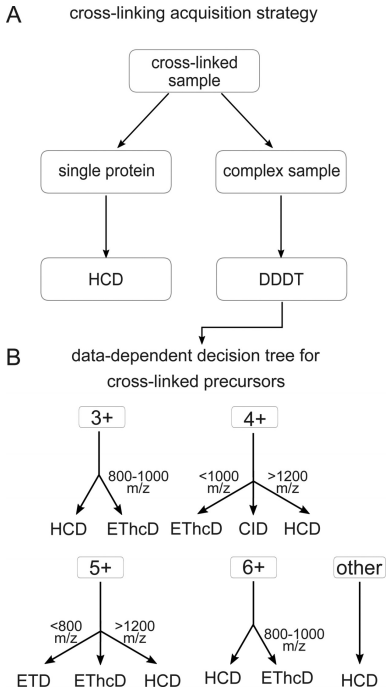


Figure 6. Acquisition strategy for cross-linked peptides. (A) Recommended acquisition scheme for cross-linking samples. (B) Data-dependent decision tree (DDDT) for cross-linked peptides. Depending on the precursor charge state (3+, 4+, 5+, 6+, and other) and the m/z, the appropriate fragmentation technique is selected.

proteins or small protein complexes, we recommend HCD as the method of choice. Since the complexity of the sample is not very high, cross-linked peptides can often be matched by precursor mass alone. In addition, HCD fragmentation generates enough fragments to precisely localize the cross-link site in the majority of cases. For the second case, that is, complex samples with many proteins not only the search space becomes an issue but also the associated random matches. A fragmentation scheme that generates highly discriminative scores for target and decoy peptides will identify more peptides under the same FDR threshold. The optimal fragmentation scheme for such an experiment is shown in Figure 6B. Earlier studies on the development of data dependent decision trees (DDDT) for the acquisition of linear peptides mainly support our conclusions: HCD gives the highest number of identifications, but ETD gives higher search engine scores[45] or, as in our case, higher sequence coverage. Compared to a DDDT for linear peptides our results are slightly different but still comparable. For example, linear DDDTs precursors with charge state 3+ have been analyzed with ETD up to 750 m/z[46] or 650 m/z,[45] we only use ETD from 600−800 m/z. In addition, instead of using ETD alone for 4+, 5+ precursors below 1000 m/z and 800 m/z, respectively, EThcD is used. In this study we investigated SDA-cross-linked, tryptic peptides. Other cross-linkers or enzymes may lead to peptide populations

DOI: 10.1021/acs.analchem.6b02082
Anal. Chem. 2016, 88, 8239−8247

with distinct fragmentation behavior due to differences in size or amino acid composition. Note, however, that the proposed fragmentation scheme is similar to the decision tree for linear peptides[45,46] and may therefore be of more general value.

CONCLUSION

For the majority of the peptides EThcD is the method of choice to achieve the highest sequence coverage. HCD is an important alternative because of its superior speed, with only somewhat reduced peptide sequence coverage. CID, ETD, and ETciD only play minor roles. We advise to adjust the acquisition scheme to follow the experimental setup: simple protein samples should be analyzed using only HCD to maximize number of observed links, which starts having value in protein structure determination.[8,47] For complex samples, we propose a decision tree that is mainly based on EThcD and HCD to maximize search specificity.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.6b02082.

Quality control results, details regarding the decision tree areas, the relative performance of the individual fragmentation techniques, and annotated spectra of all PSMs (PDF).

AUTHOR INFORMATION

Corresponding Author

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

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The Wellcome Trust generously funded this work through a Senior Research Fellowship to J.R. (103139), a Centre Core Grant (092076), and an Instrument Grant (108504).

REFERENCES

(1) Perdigão, N.; Heinrich, J.; Stolte, C.; Sabir, K. S.; Buckley, M. J.; Tabor, B.; Signal, B.; Gloss, B. S.; Hammang, C. J.; Rost, B.; et al. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 15898−15903.
(2) Rappsilber, J. *J. Struct. Biol.* **2011**, *173*, 530−540.
(3) Holding, A. N. *Methods* **2015**, *89*, 54−63.
(4) Sinz, A. *Mass Spectrom. Rev.* **2006**, *25*, 663−682.
(5) Schmidt, C.; Robinson, C. V. *Nat. Protoc.* **2014**, *9*, 2224−2236.
(6) 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. *EMBO J.* **2010**, *29*, 717−726.
(7) Walzthoeni, T.; Leitner, A.; Stengel, F.; Aebersold, R. *Curr. Opin. Struct. Biol.* **2013**, *23*, 252−260.
(8) Belsom, A.; Schneider, M.; Fischer, L.; Brock, O.; Rappsilber, J. *Mol. Cell. Proteomics* **2016**, *15*, 1105−1116.
(9) Mayne, S. L. N.; Patterton, H.-G. *Briefings Bioinf.* **2011**, *12*, 660−671.
(10) Sinz, A.; Arlt, C.; Chorev, D.; Sharon, M. *Protein Sci.* **2015**, *24*, 1193−1209.
(11) 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. *Nat. Methods* **2012**, *9*, 904−906.
(12) Chalkley, R. J.; Baker, P. R.; Medzihradszky, K. F.; Lynn, A. J.; Burlingame, A. L. *Mol. Cell. Proteomics* **2008**, *7*, 2386−2398.
(13) Trnka, M. J.; Baker, P. R.; Robinson, P. J. J.; Burlingame, a L.; Chalkley, R. J. *Mol. Cell. Proteomics* **2014**, *13*, 420−434.
(14) Götze, M.; Pettelkau, J.; Schaks, S.; Bosse, K.; Ihling, C. H.; Krauth, F.; Fritzsche, R.; Kühn, U.; Sinz, A. *J. Am. Soc. Mass Spectrom.* **2012**, *23*, 76−87.
(15) Rinner, O.; Seebacher, J.; Walzthoeni, T.; Mueller, L. N.; Beck, M.; Schmidt, A.; Mueller, M.; Aebersold, R. *Nat. Methods* **2008**, *5*, 315−318.
(16) Hoopmann, M. R.; Zelter, A.; Johnson, R. S.; Riffle, M.; MacCoss, M. J.; Davis, T. N.; Moritz, R. L. *J. Proteome Res.* **2015**, *14*, 2190−2198.
(17) Giese, S. H.; Fischer, L.; Rappsilber, J. *Mol. Cell. Proteomics* **2016**, *15*, 1094−1104.
(18) Maiolica, A.; Cittaro, D.; Borsotti, D.; Sennels, L.; Ciferri, C.; Tarricone, C.; Musacchio, A.; Rappsilber, J. *Mol. Cell. Proteomics* **2007**, *6*, 2200−2211.
(19) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551−3567.
(20) Fischer, L.; Chen, Z. A.; Rappsilber, J. *J. Proteomics* **2013**, *88*, 120−128.
(21) Pearson, K. M.; Pannell, L. K.; Fales, H. M. *Rapid Commun. Mass Spectrom.* **2002**, *16*, 149−159.
(22) Chu, F.; Maynard, J. C.; Chiosis, G.; Nicchitta, C. V.; Burlingame, A. L. *Protein Sci.* **2006**, *15*, 1260−1269.
(23) Müller, M. Q.; Dreiocker, F.; Ihling, C. H.; Schäfer, M.; Sinz, A. *Anal. Chem.* **2010**, *82*, 6958−6968.
(24) Leitner, A.; Reischl, R.; Walzthoeni, T.; Herzog, F.; Bohn, S.; Förster, F.; Aebersold, R. *Mol. Cell. Proteomics* **2012**, *11*, M111.014126.
(25) Shi, Y.; Fernandez-Martinez, J.; Tjioe, E.; Pellarin, R.; Kim, S. J.; Williams, R.; Schneidman-Duhovny, D.; Sali, A.; Rout, M. P.; Chait, B. T. *Mol. Cell. Proteomics* **2014**, *13*, 2927−2943.
(26) Nguyen-Huynh, N.-T.; Sharov, G.; Potel, C.; Fichter, P.; Trowitzsch, S.; Berger, I.; Lamour, V.; Schultz, P.; Potier, N.; Leize-Wagner, E. *Protein Sci.* **2015**, *24*, 1232−1246.
(27) Frese, C. K.; Altelaar, A. F. M.; van den Toorn, H.; Nolting, D.; Griep-Raming, J.; Heck, A. J. R.; Mohammed, S. *Anal. Chem.* **2012**, *84*, 9668−9673.
(28) Chowdhury, S. M.; Du, X.; Tolić, N.; Wu, S.; Moore, R. J.; Mayer, M. U.; Smith, R. D.; Adkins, J. N. *Anal. Chem.* **2009**, *81*, 5524−5532.
(29) Liu, F.; Rijkers, D. T. S.; Post, H.; Heck, A. J. R. *Nat. Methods* **2015**, *12*, 1179−1184.
(30) Trnka, M. J.; Burlingame, A. L. *Mol. Cell. Proteomics* **2010**, *9*, 2306−2317.
(31) Syka, J. E. P.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 9528−9533.
(32) Mikesh, L. M.; Ueberheide, B.; Chi, A.; Coon, J. J.; Syka, J. E. P.; Shabanowitz, J.; Hunt, D. F. *Biochim. Biophys. Acta, Proteins Proteomics* **2006**, *1764*, 1811−1822.
(33) Kim, M.-S.; Pandey, A. *Proteomics* **2012**, *12*, 530−542.
(34) Good, D. M.; Wirtala, M.; McAlister, G. C.; Coon, J. J. *Mol. Cell. Proteomics* **2007**, *6*, 1942−1951.
(35) Sugio, S.; Kashima, A.; Mochizuki, S.; Noda, M.; Kobayashi, K. *Protein Eng., Des. Sel.* **1999**, *12*, 439−446.
(36) Rappsilber, J.; Ishihama, Y.; Mann, M. *Anal. Chem.* **2003**, *75*, 663−670.
(37) Ishihama, Y.; Rappsilber, J.; Andersen, J. S.; Mann, M. *J. Chromatogr. A* **2002**, *979*, 233−239.
(38) Renard, B. Y.; Kirchner, M.; Monigatti, F.; Ivanov, A. R.; Rappsilber, J.; Winter, D.; Steen, J. A. J.; Hamprecht, F. A.; Steen, H. *Proteomics* **2009**, *9*, 4978−4984.
(39) Fischer, L.; Rappsilber, J. 2016, in preparation.
(40) EDGINGTON, E. S. *J. Psychol.* **1964**, *57*, 445−449.
(41) Vizcaíno, J. A.; Côté, R. G.; Csordas, A.; Dianes, J. A.; Fabregat, A.; Foster, J. M.; Griss, J.; Alpi, E.; Birim, M.; Contell, J.; et al. *Nucleic Acids Res.* **2013**, *41*, D1063−D1069.
(42) Hebert, A. S.; Richards, A. L.; Bailey, D. J.; Ulbrich, A.; Coughlin, E. E.; Westphall, M. S.; Coon, J. J. *Mol. Cell. Proteomics* **2014**, *13*, 339−347.
(43) Molina, H.; Matthiesen, R.; Kandasamy, K.; Pandey, A. *Anal. Chem.* **2008**, *80*, 4825−4835.

DOI: 10.1021/acs.analchem.6b02082
Anal. Chem. 2016, 88, 8239−8247

(44) Frese, C. K.; Zhou, H.; Taus, T.; Altelaar, A. F. M.; Mechtler, K.; Heck, A. J. R.; Mohammed, S. *J. Proteome Res.* **2013**, *12*, 1520−1525.
(45) Frese, C. K.; Altelaar, A. F. M.; Hennrich, M. L.; Nolting, D.; Zeller, M.; Griep-Raming, J.; Heck, A. J. R.; Mohammed, S. *J. Proteome Res.* **2011**, *10*, 2377−2388.
(46) Swaney, D. L.; McAlister, G. C.; Coon, J. J. *Nat. Methods* **2008**, *5*, 959−964.
(47) Belsom, A.; Schneider, M.; Brock, O.; Rappsilber, J. *Trends Biochem. Sci.* **2016**, *41*, 564−567.

DOI: 10.1021/acs.analchem.6b02082
Anal. Chem. 2016, 88, 8239−8247

Chapter 4

Manuscript 3. Noncovalently Associated Peptides Observed during Liquid Chromatography-Mass Spectrometry and Their Effect on Cross-Link Analyses

analytical chemistry

Article

Cite This: Anal. Chem. 2019, 91, 2678−2685

pubs.acs.org/ac

Noncovalently Associated Peptides Observed during Liquid Chromatography-Mass Spectrometry and Their Effect on Cross-Link Analyses

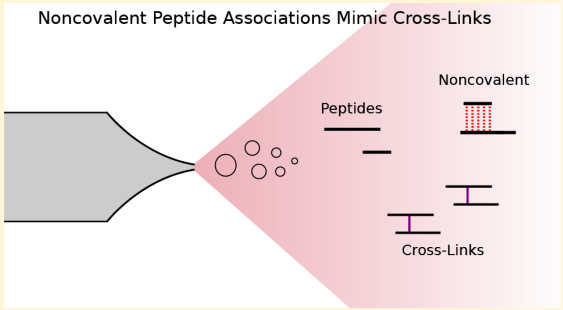
Sven H. Giese,[†] Adam Belsom,[†,‡] Ludwig Sinn,[†] Lutz Fischer,[†,‡] and Juri Rappsilber*[,†,‡]

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

[‡]Wellcome Centre for Cell Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH93BF, United Kingdom

Supporting Information

ABSTRACT: Cross-linking mass spectrometry draws structural information from covalently linked peptide pairs. When these links do not match to previous structural models, they may indicate changes in protein conformation. Unfortunately, such links can also be the result of experimental error or artifacts. Here, we describe the observation of noncovalently associated peptides during liquid chromatography-mass spectrometry analysis, which can easily be misidentified as cross-linked. Strikingly, they often mismatch to the protein structure. Noncovalently associated peptides presumably form during ionization and can be distinguished from cross-linked peptides by observing coelution of the corresponding linear peptides in MS1 spectra, as well as the presence of the individual (intact) peptide fragments in MS2 spectra. To suppress noncovalent peptide formations, increasingly disruptive ionization settings can be used, such as in-source fragmentation.



The preservation of noncovalent associations in electrospray ionization (ESI) has been widely used in the field of native mass spectrometry to study protein interactions. Major achievements of native mass spectrometry include analyzing the topology and stoichiometry of multiprotein complexes and the binding of small molecules to proteins.[1−3] The key premise of the field is that the observed noncovalent interactions in the gas phase are based on biologically relevant interactions in the aqueous phase.[4]

Another mass spectrometric field that investigates (non)-covalent interactions of proteins is cross-linking mass spectrometry (CLMS).[5−7] Here, spatially close amino acid residues in native proteins are covalently linked. This preserves spatial information throughout the subsequent non-native analytical process, comprising trypsin digestion of the proteins into peptides and their chromatographic separation for mass spectrometric detection. A key premise of this field is that the observed peptide interactions in the gas phase are exclusively based on covalent links. Note that, for synthetic peptides, gas-phase peptide−peptide complexes have been observed recently,[8] suggesting that not only proteins but also peptides can remain associated during mass spectrometric analysis.

In theory, one can construct peptide pairs where mass information alone cannot differentiate between covalent linkage and noncovalent association. A peptide pair can reach the same mass either by cross-linking or by noncovalent association if one of the two peptides carries a loop-link, that is, the frequent case of a cross-linker reacting with two amino acid residues so near in sequence that they fall into a tryptic peptide (Figure S1, Supporting Information). The concept of mass equivalence between cross-linked and non-cross-linked peptides has been exploited during data analysis, when using standard proteomics software for the analysis of cross-linked peptides, including Mascot[9] to identify cross-linked peptides[10] and quantitation software.[11,12] If such noncovalent associations physically arise, current cross-link analysis could be fooled into misidentifying analytical artifacts as spatial information.

We observed surprising differences when comparing the identified cross-links using data acquired on two different mass spectrometers: a hybrid linear ion trap-Orbitrap mass spectrometer (LTQ Orbitrap Velos, Thermo Fisher Scientific) and a hybrid quadrupole-Orbitrap mass spectrometer (Q Exactive, Thermo Fisher Scientific). This led us to investigate the formation of noncovalent peptide associations with and without cross-linking. We analyzed cross-linked human serum albumin (HSA). Using only the monomeric protein band obtained from sodium dodecyl sulfate polyacrylamide gel electrophoresis allowed identified links to be validated against an available three-dimensional structural model as "ground truth" to reveal suspicious peptide pairs for detailed interrogation. We then extended this data analysis to a four-protein

Received: September 3, 2018
Accepted: January 16, 2019
Published: January 16, 2019

mix without employing cross-linking to test if the noncovalent association is cross-linker-specific.

MATERIALS AND METHODS

Data Acquisition. *HSA Acquisition and Sample Preparation.* Human blood serum (20 μg aliquots, 1 μg/μL) was cross-linked using cross-linker-to-protein, weight-to-weight (w/w) ratios of 1:1 and 2:1. Aliquots of human serum diluted with cross-linking buffer (20 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES)−OH, 20 mM NaCl, 5 mM $MgCl_2$, pH 7.8) were incubated with sulfosuccinimidyl 4,4′-azipentanoate (sulfo-SDA) (Thermo Scientific Pierce, Rockford, IL), in a reaction volume of 30 μL for 1 h at room temperature. The diazirine group was then photoactivated by UV irradiation, for either 10, 20, 40, or 60 min using a UVP CL-1000 UV Cross-linker (UVP Inc.). Cross-linked samples were separated using gel electrophoresis, with bands corresponding to monomeric HSA excised and then reduced with dithiothreitol, alkylated with iodoacetamide, and digested using trypsin following standard protocols.[10] Peptides were then desalted using C18 StageTips[13] and eluted with 80% acetonitrile, 20% water, and 0.1% trifluoroacetic acid (TFA).

Peptides were analyzed on either a hybrid linear ion trap/Orbitrap mass spectrometer (LTQ Orbitrap Velos, Thermo Fisher Scientific) or a hybrid quadrupole/Orbitrap mass spectrometer (Q Exactive, Thermo Fisher Scientific). In both cases, peptides were loaded directly onto a spray analytical column (75 μm inner diameter, 8 μm opening, 250 mm length; New Objectives, Woburn, MA) packed with C18 material (ReproSil-Pur C18-AQ 3 μm; Dr. Maisch GmbH, Ammerbuch-Entringen, Germany) using an air pressure pump (Proxeon Biosystems).[14]

Orbitrap Velos Analysis. Mobile phase A consisted of water and 0.1% formic acid. Peptides were loaded using a flow rate of 0.7 μL/min and eluted at 0.3 μL/min, using a gradient with a 1 min linear increase of mobile phase B (acetonitrile and 0.1% v/v formic acid) from 1% to 9%, increasing linearly to 35% B in 169 min, with a subsequent linear increase to 85% B over 5 min. Eluted peptides were sprayed directly into the hybrid linear ion trap-Orbitrap mass spectrometer. MS data were acquired in the data-dependent mode, detecting in the Orbitrap at 100 000 resolution. The eight most intense ions in the MS spectrum for each acquisition cycle, with a precursor charge state of +3 or greater, were isolated with a m/z window of 2 Th and fragmented in the linear ion trap with collision-induced dissociation (CID) at a normalized collision energy of 35. Subsequent (MS2) fragmentation spectra were then recorded in the Orbitrap at a resolution of 7500. Dynamic exclusion was enabled with single repeat count for 90 s.

Q Exactive Analysis. Mobile phase A consisted of water and 0.1% formic acid. Mobile phase B consisted of 80% v/v acetonitrile and 0.1% formic acid. Peptides were loaded at a flow rate of 0.5 μL/min and eluted at 0.2 μL/min, using a gradient increasing linearly from 2% B to 40% B in 169 min, with a subsequent linear increase to 95% B over 11 min. Eluted peptides were sprayed directly into the hybrid quadrupole-Orbitrap mass spectrometer. MS data (400−1600 m/z) were acquired in the data-dependent mode, detecting in the Orbitrap at 60 000 resolution. The ten most intense ions in the MS spectrum, with a precursor charge state of +3 or greater, were isolated with a m/z window of 2 Th and fragmented by higher-energy collision-induced dissociation (HCD) at a normalized collision energy of 28. Subsequent (MS2) fragmentation spectra were recorded in the Orbitrap at a resolution of 30 000. Dynamic exclusion was enabled with a single repeat count for 60 s.

In-Source Collision-Induced Dissociation Acquisitions. HSA, equine myoglobin, ovotransferrin from chicken (all from Sigma-Aldrich, St. Louis, MO), and creatine kinase from rabbit (Roche, Basel, Switzerland) were dissolved in 8 M urea with 50 mM ammonium bicarbonate to a concentration of 2 mg/mL each. The proteins were reduced by adding dithiothreitol at 2.5 mM followed by an incubation for 30 min at 20 °C. Subsequently, the samples were derivatized using iodoacetamide at 5 mM concentration for 20 min in the dark at 20 °C. The samples were diluted 1:5 with 50 mM ammonium bicarbonate and digested with trypsin (Pierce Biotechnology, Waltham, MA) at a protease-to-protein ratio of 1:100 (w/w) during a 16-h incubation period at 37 °C. Then the digestion was stopped by adding 10% TFA at a concentration of 0.5%. The digests were cleaned up using the StageTip protocol.[13] The samples were eluted from the C18 phase, partially evaporated using a vacuum concentrator, and resuspended in mobile phase A (0.1% formic acid). Two micrograms of tryptic digests were loaded directly onto a 50 cm EASY-Spray column (Thermo Fisher) packed with C18 stationary phase and equilibrated to 2% of mobile phase B (80% acetonitrile, 0.1% formic acid) running at a flow of 0.3 μL/min. Peptides were eluted by increasing mobile phase B content from 2 to 37.5% over 120 min, followed by ramping to 45% and to 95% within 5 min each. After a washing period of 5 min, the column was re-equilibrated to 2% B. The eluting peptides were sprayed into a Q Exactive High-field (HF) Hybrid Quadrupole-Orbitrap Mass Spectrometer (Thermo Fisher Scientific, Bremen, Germany). The mass spectrometric measurements in data-dependent mode were acquired as follows: a full scan from 400 to 1600 m/z with a resolution of 120 000 was recorded to find suitable peptide candidates which were subsequently quadrupole-isolated within a m/z window of 2 Th and fragmented by HCD at a normalized collision energy of 28, with fragmentation spectra recorded in the Orbitrap at a resolution of 30 000. Precursors with charge states from 3 to 6 were selected for isolation. Dynamic exclusion was set to 15 s. Each cycle allowed up to ten peptides to be fragmented before a new full scan was triggered. The effect of in-source collisional activation (ISCID) on the formation of noncovalently bound peptides was investigated by setting voltages from 0 to 20 eV in 5 eV increments for each individual run. Each value tested was probed in shuffled triplicates.

Data Processing. Raw files for cross-linking searches were processed using MaxQuant[15] (v. 1.6.1.0) to benefit from the implemented precursor m/z and charge correction. Resulting peak files in APL format were used to identify peptides in Xi[16] (v. 1.6.739). The database search with Xi used the following parameters: MS tolerance, 6 ppm; MS2 tolerance, 15 ppm; missed cleavages, 3; enzyme, trypsin; fixed modifications, carbamidomethylation (cm, +57.02 Da); variable modifications, oxidation methionine (ox, +15.99 Da). For sulfo-SDA, the cross-linker mass 82.04 Da and the modifications SDA-loop (+82.04 Da) and SDA-hyd (+100.05 Da) were used.[17] False discovery rate (FDR) estimation was done using xiFDR[18] (v. 1.1.26.58), using either 5% link FDR (without boosting) or a 5% peptide spectrum match (PSM) FDR. The Euclidean cross-link distances within HSA were estimated from mapping the peptide sequences to the three-dimensional structure when possible (PDB: 1AO6[19]).

DOI: 10.1021/acs.analchem.8b04037
Anal. Chem. 2019, 91, 2678−2685

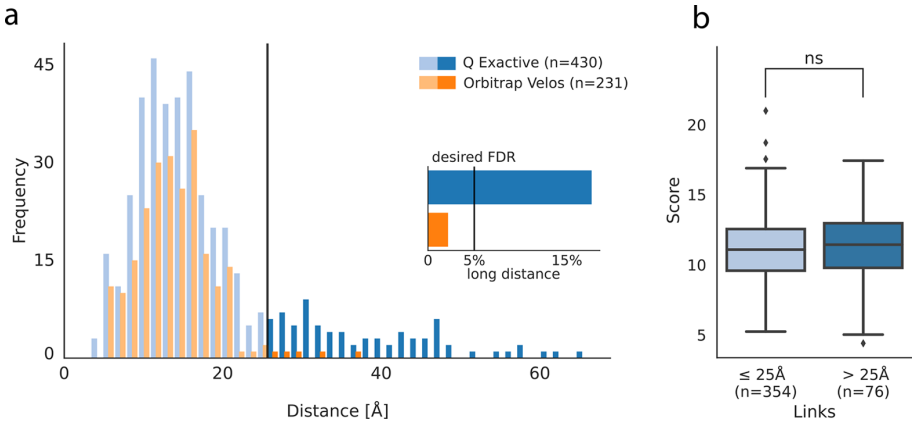


Figure 1. Quality control after cross-link identification at a 5% link FDR. (a) Results from cross-linking HSA with sulfo-SDA acquired on an Q Exactive and an LTQ Orbitrap Velos mass spectrometer. The line at 25 Å indicates the distance cutoff for links classified as long distance. The inlet shows the fraction of long-distance links (LDL) in each data set. (b) Score comparison between within-distance linkzs and LDL. LDL showed no significant (ns) deviation from the within-distance links (two-sided Mann−Whitney-U-test at $\alpha = 0.05$).

Searches for noncovalently associated peptides (NAP) in the absence of cross-linkers were also conducted using Xi with a feature to search for noncovalently associated peptides. FDR analysis was done at a 5% PSM level using the formula FDR = (TD − DD)/TT,[18] after removing all PSMs with a score less than 1. FDRs were then transformed to q-values, defined as the minimal FDR at which a PSM would pass the threshold.[20]

Linear peptide identifications from cross-linked acquisitions were done using MaxQuant. We added the above-defined SDA-loop and SDA-hyd modifications to the configuration file and allowed up to five modifications on a peptide together with a maximum of five missed cleavages. Resulting peptide identifications were filtered at the default FDR of 1%. Non-cross-linked acquisitions were searched with default settings treating each replicate as a different experiment in the experimental design.

RT profiles for a given m/z were extracted using the MS1 (peak picked) raw data after conversion to mzML using msconvert.[21] The postprocessing was done in Python using pyOpenMS.[22] RT profiles were defined as intensity values for a given m/z for the monoisotopic peak and two isotope peaks. During the developed look-up strategy, the precursor m/z of the identified cross-linked peptide, the m/z of the α peptide, and the m/z of the β peptide were searched in the MS1 data. The precursor mass matches only the sum of the individual peptides in a noncovalently associated peptide if one of the two peptides is SDA-loop-modified. Therefore, the MS1 data was screened for m/z traces of the individual peptides with and without an added SDA-loop modification. Similarly, all charge states up to the precursor charge were used. The m/z trace with the largest number of peaks was eventually selected for each individual peptide. The m/z seeds were all treated similarly; in a RT window of 180 s, the given m/z was searched with a 20 ppm tolerance. If the m/z was found, the intensity was extracted. Resulting RT profiles where smoothened by a moving average with 15 points. For further data processing and visualization, the RT profile with the most peaks (either monoisotopic, first isotope, or second isotope peak) was selected.

Statistical analysis and data processing were performed using Python and the scientific package SciPy.[23] Unless otherwise noted, we performed significant tests using one-sided Mann−Whitney-U-Tests with $\alpha = 0.05$ and continuity correction. We used the following encoding for p-values: ns, not significant; *, ≤0.05; **, ≤0.01; ***, ≤0.001. Along with the significance tests, we provided effect size estimates based on Cohen's d[24] with pooled standard deviations, which uses the following classification: small, $|d| \geq 0.2$; medium, $|d| \geq 0.5$; and large, $|d| \geq 0.8$.

The mass spectrometry raw files, peak lists, search engine results, MaxQuant parameter files, and FASTA files have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository[25] with the data set identifier PXD010895.

RESULTS AND DISCUSSION

The results are divided into 4 parts: (1) Describes the results from HSA cross-linking using sulfosuccinimidyl 4,4′-azipentanoate (sulfo-SDA) and then analysis with a Q Exactive (QE) and LTQ Orbitrap Velos (Velos) mass spectrometer; (2) describes the MS2 properties of the detected long-distance links (LDL) with the QE and introduces the hypothesis of noncovalently associated peptides (NAP) enduring ESI; (3) summarizes intensity and retention time (RT) properties of the identified PSMs; and (4) shows that noncovalently associated peptides also occur in the absence of cross-linking.

Instrument Comparison Revealing a High Number of Suspicious Cross-links in Q Exactive Data. We started by comparing the results from cross-linking HSA with sulfo-SDA using two different mass spectrometers: a Velos and a QE. Cross-linked peptides were identified using Xi with subsequent FDR filtering using xiFDR at a 5% link-level FDR. To independently assess the quality of the results, we evaluated how the identified cross-links matched to the available crystal structure of HSA. At 5% link FDR, we identified 449 (QE) and 240 (Velos) links, of which 430 and 231 could be mapped to the available sequence in the structural model, respectively. The distance distributions of the mapped cross-links looked similar for links below 22 Å (Figure 1a). However, for long distances, the link distributions looked different. The QE data shows a much higher percentage of links exceeding the 25 Å cutoff, which is the empirically defined distance limit of SDA cross-linking.[26] This leads to 18% long-distance links (LDL) for the QE data compared to 2% for the Velos data (Figure 1a

DOI: 10.1021/acs.analchem.8b04037
Anal. Chem. 2019, 91, 2678−2685

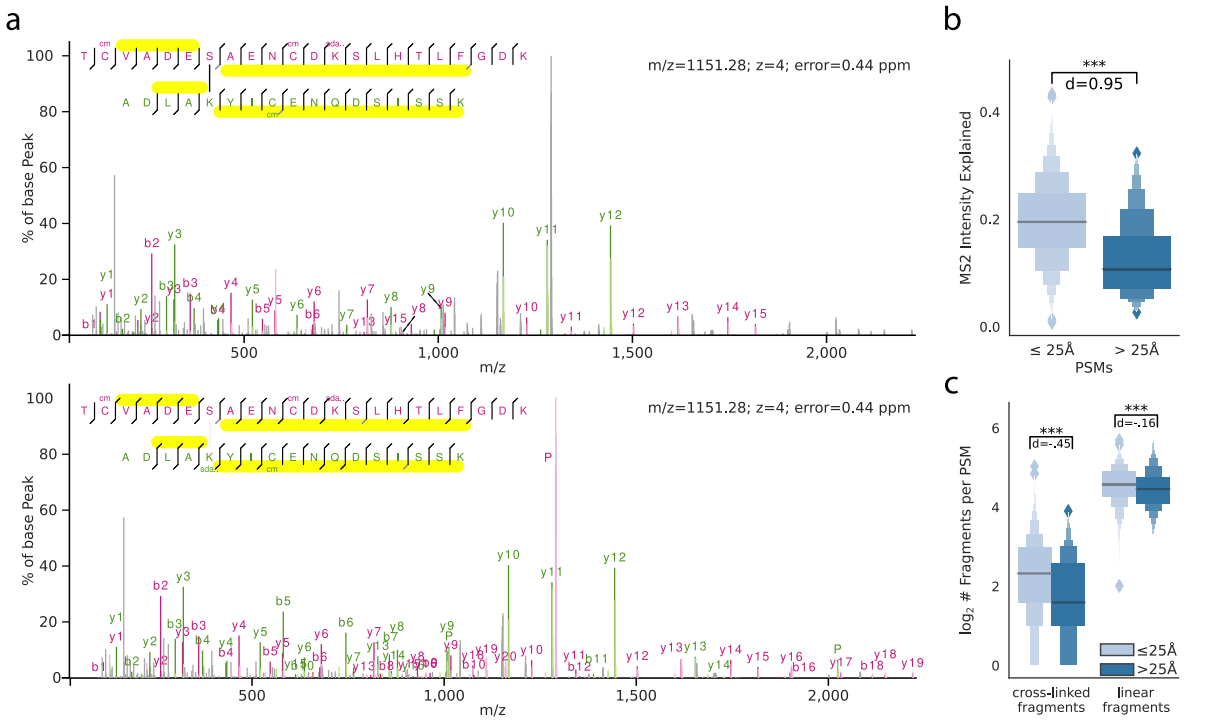


Figure 2. Spectral characteristic of noncovalently associated peptides. (a) Comparison of the same scan (scan 34887, raw file *V127_F*) searched with cross-link settings (upper panel) and searched with a noncovalent association setting (lower panel). (b) Comparison of the explained intensity in the MS2 spectrum from all PSMs that passed the 5% link-level FDR. (c) Comparison of cross-linker-containing fragments and linear fragments in the same set of PSMs as in (b). Number of observations for (b) and (c): ≤25 Å 2599 PSMs and >25 Å 326 PSMs.

inlet). Since the protein monomer band was analyzed, the possibility that the LDL were derived from cross-linked homo-oligomers can be largely neglected. One possibility is that the deeper analysis on the QE, which is faster and more sensitive than the Velos, detected a rare protein conformational state. However, a previous analysis of SDA-cross-linked HSA on the Velos yielded 500 identified links (5% link FDR), with comparatively few LDL (6%).[27] Also, data on the much faster and more sensitive Fusion Lumos did not return in our hands such proportion of LDLs (data not shown). This suggests that the QE data does not cover conformational flexibility of the protein. Instead, the QE data appears to suffer from a systematic error that leads to many false identifications. Importantly, this bias affects only target sequences as it is not controlled by the FDR estimation. If these LDL were indeed based on false identifications, one could suspect that they were identified based on weak data and thus derived from low-scoring PSMs. We therefore compared the highest scoring PSM for each link above and below 25 Å (Figure 1b). Remarkably, the LDL showed an even higher average score than the within-distance links. This difference was small and not significant, but it was still surprising that the two classes had a similar score distribution. Next, we manually investigated LDL PSMs to identify characteristics that might lead to a mechanistic explanation of these links.

Long-Distance Links Lacking Support for Being Cross-Linked. After suspecting a systematic identification error in QE data, we manually inspected annotated LDL spectra. We noticed that many spectra frequently contained unexplained fragment peaks of high intensity. For example, in the displayed spectrum (Figure 2a, upper panel), most of the high-intensity peaks are explained but not the base peak. This PSM was matched with a very low precursor error of 0.44 ppm and had a very good sequence coverage in general. However, while many of the linear fragments were identified, no cross-linked fragments were matched. While there is convincing evidence that the two identified peptides are correct, there is a lack of fragment evidence that these peptides were indeed cross-linked.

We tested our manual observations more systematically by comparing the explained intensity in the MS2 spectrum across all PSMs that passed the 5% link FDR (Figure 2b). There is already a twofold increase in the median explained intensity (EI) of the within-distance links (20% EI) and the LDL (10% EI). This trend is also supported by a significant MWU test (one-sided, α = 0.05) and a large Cohen's d effect size (d = 0.95). One possible explanation is that the spectra that yield LDL are simply of poor quality. This can happen when, for example, peptides of similar m/z were coisolated, the precursor was of low intensity, or the peptide simply did not fragment or ionize very well. But as shown in Figure 1b, the search engine scores of LDL were slightly higher than the scores from within-distance links. Therefore, poor spectral quality is not a likely reason for the large proportion of LDL. However, the number of matched cross-linked and linear fragments was significantly lower for the long-distance matches compared to that for the within-distance matches (Figure 2c).

Recently, it has been proposed that SDA-formed bonds are very susceptible to MS cleavage when involving a carboxylic acid functional group.[28] In these cases, the annotated spectra

DOI: 10.1021/acs.analchem.8b04037
Anal. Chem. 2019, 91, 2678−2685

Analytical Chemistry Article

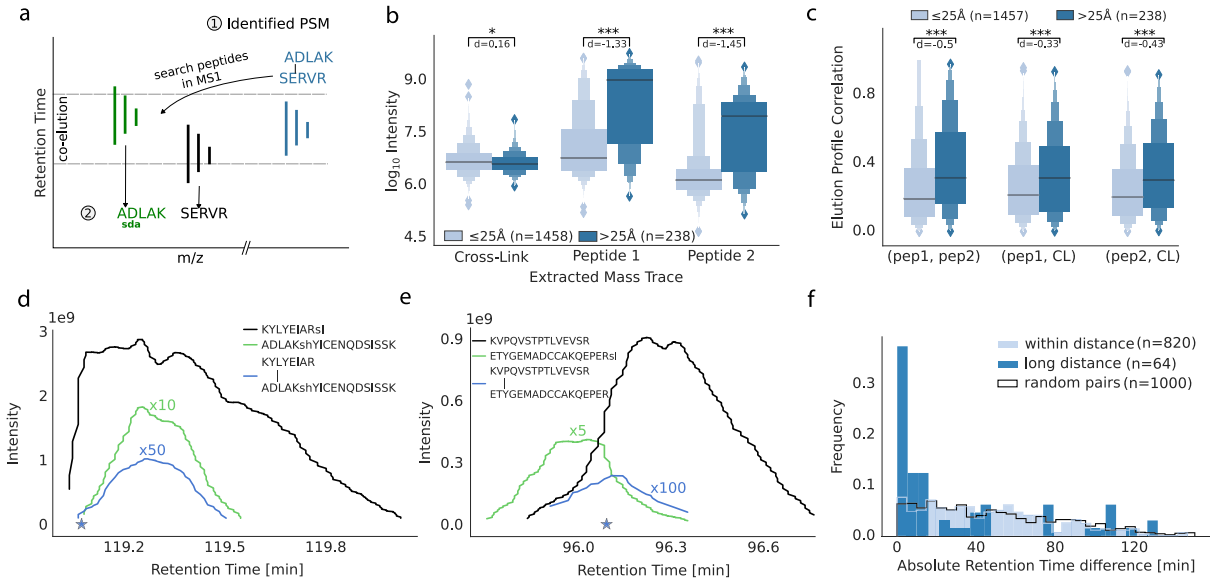


Figure 3. Analysis strategy and properties of LDL PSMs. (a) Noncovalent peptide search. On the basis of a cross-linked PSM (1), the individual peptide sequences are searched in the MS1 such that the summed mass equals the precursor mass of the identified cross-link (2). (b) Maximum intensity (along the m/z trace) for the identified cross-link and the m/z of the two individual peptides for links ≤25 Å and >25 Å. (c) Spearman correlation of intensity profiles of the cross-link and the two individual peptides based on m/z matching in a RT window. (d, e) Examples of intensity profiles of two LDL. Filled stars mark the isolation time point of the precursor that yielded the identified cross-link. Scaling factors for lower intensity curves are written above the respective curves (e.g., ×10 equals a factor of 10). Additional information about the PSMs can be retrieved through the uploaded results in PRIDE through the PSMIDs 7678478210 (d) and 7678602613 (e). (f) RT difference comparison of LDL and within-distance links.

would also show a low EI and a low number of cross-linked fragments with our search settings. However, it is unclear why such a reaction should preferentially lead to LDLs. Therefore, we hypothesized that the respective peptide pairs were not actually cross-linked but were noncovalently associated. Nevertheless, we investigated this in larger detail by following the approach of Iacobucci et al.[28] and performed a cleavable cross-linker search on the Velos and the QE acquisitions (Figure S2). A large portion of the identifications from the cleavable cross-linker search on the QE (38%) were long-distance links (presumably noncovalent peptide associations). However, the distribution of links that match the crystal structure revealed a preference for short distances, thereby indicating that MS cleavage of the cross-linker can indeed be observed. So, our data support both as parallel processes MS-cleavable SDA links and noncovalent peptide complexes.

It would be interesting to investigate sequence determinants of noncovalent association. Unfortunately, the lack of ground truth and the low number of observations make it difficult to investigate sequence-specific features that lead to noncovalent peptide complexes. While cross-links should preferentially fall below the distance cutoff, noncovalent peptide associations should distribute randomly across the distance histogram. Therefore, some links that match the crystal structure will also arise from noncovalent associations. Those links falling above the distance cutoff were too low in number for a statistical enrichment analysis.

Low Intense Noncovalently Associated Peptides Arising from Two Coeluting Peptides. As shown above, LDL frequently achieved high scores and there was good evidence based on the MS2 fragmentation that the peptides were correctly identified. Had the peptides paired non-covalently, this could happen either in solution or during the ESI process. In the latter case, one would expect the individual peptides to overlap in their chromatographic elution forming a noncovalent pair during their coelution. In contrast, for cross-linked peptides one would not expect any systematic coelution. Therefore, we investigated the elution of the individual peptides for all identifications (5% PSM FDR) following a look-up strategy that started from the MS2 trigger time of the cross-linked PSM (Figure 3a, for details see Materials and Methods).

We successfully extracted 1458 mass traces for PSMs of links within the distance cutoff and 238 mass traces for PSMs of LDLs. For these PSMs, we then compared the maximum intensity along the mass trace for the cross-link m/z and the two individual peptides m/z (Figure 3b) within a window of ±90 s. Interestingly, the MS1 signals of long-distance links had significantly lower intensities than links fitting to the crystal structure, albeit with small effect size. In contrast, the MS1 intensities attributed to the individual peptides of LDL were higher by almost 2 orders of magnitude within the elution window compared to the control (peptides observed in cross-links). This indicates a preference for coelution of individual peptides with linked peptide pairs in the case of LDL but not within-distance links.

The high signal intensity of individual peptides of LDL around the elution of the LDL peptide made us wonder if they coelute. We investigated the correlation of elution profiles more systematically by computing the Spearman correlation over the extracted ion chromatogram (XIC). While the absolute correlation is neither very high for the within-distance links nor for the LDL, the important feature is the difference between the two classes (Figure 3c). The correlations of two

DOI: 10.1021/acs.analchem.8b04037
Anal. Chem. 2019, 91, 2678−2685

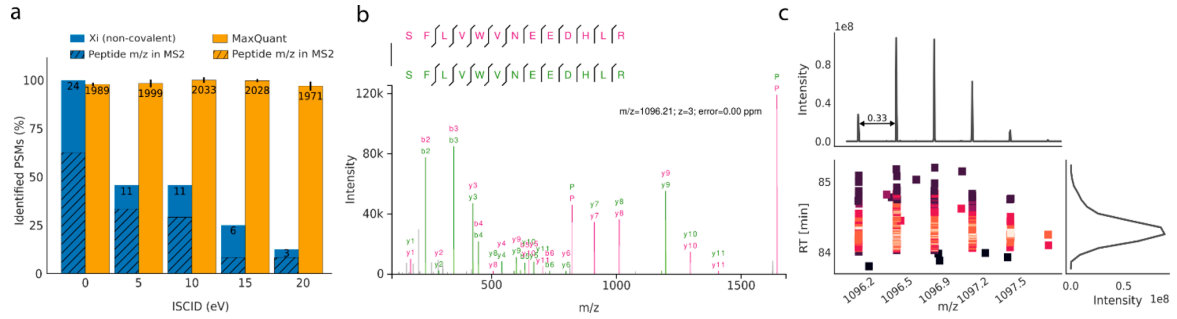


Figure 4. Noncovalently associated peptide identifications in non-cross-linked samples. (a) Number of PSMs after 5% PSM FDR in a noncovalent search and linear identifications (1% FDR). Peptide m/z fraction refers to occurrences where the individual peptide or precursor peaks are found in multiple charge states in the MS2 spectrum. (b) Noncovalent peptide identification with charge state 3, individual peptide peaks (P) were identified with charge 2 (822.41 m/z) and charge 1 (1643.82 m/z). (c) MS1-derived peptide feature for the PSM displayed in (b). Top panel shows the summed intensity over the m/z bins. Bottom panel shows the m/z over the RT color-coded by the intensity. Right panel shows the summed intensity over the RT.

single peptide m/z's with each other—but also individually with the cross-linked m/z—are all significantly larger for the long-distance links compared to those for the within-distance links (p-value ≤0.001). The fact that the absolute value of the correlation is moderate is not surprising as it would be a precondition of noncovalent association that the individual peptides elute at an overlapping but not necessarily identical time, as is also seen from two examples of coeluting and associating peptides (Figure 3d,e). In the first example, all three m/z species start eluting at a similar time point. One of them is very abundant (MS1 intensity 1e9), reaching saturation and showing a long elution tail. This covers the complete elution time of the second peptide. As expected for an association product of the two, the LDL peptide then coincides with the elution of the second peptide. In a second example, the two individual peptides partially coelute, and the LDL peptide is observed during the time of their overlapping elution.

To our surprise, some cross-links that match the protein structure showed correlating MS1 intensities with their linear counterparts, despite a narrow matching time window. Retention on a reversed phase is usually very sensitive such that even peptide pairs with different cross-link sites show a different elution time.[26] We therefore suspected the coeluting MS1 intensities to be the baseline signal of our look-up strategy, which is solely based on m/z values and lacks confirmation through identification data. Hence, we checked for the RTs from the individual linear peptides relative to the cross-links based on identifications instead of m/z matching alone. We compared the cross-link identifications with the closest RT from the linear identified peptides (with equal modifications and equal composition). The absolute difference of the individual RTs was mostly close to 0 min for the LDL PSMs and approximately uniformly distributed for within-distance PSMs (Figure 3f). The added control (random pairings of RTs from linear identified peptides that were also part of a cross-linked peptide) closely resembles the within-distance PSM distribution. However, only 50% of the PSMs have a RT difference smaller than 10 min. The remaining PSMs have a large RT difference which reduces the possibility of coelution. Interestingly, PSMs with a RT difference smaller than 10 min have an average score of 10.0 (n = 32), while the remaining PSMs (n = 32) have an average score of 6.7. Possibly, the lower score indicates imprecise peptide identifications and thus wrong RT times. In addition, matches with large RT differences can still originate from wrong identifications. Like target-decoy matches in a cross-link, in a NAP one of the peptides could be correct and the other might be a random match. In these cases, the RT difference would also be randomly distributed.

In-Source Fragmentation Reduction of the Number of Noncovalently Associated Peptides. On the basis of the results above, one would predict NAPs to form even without prior cross-linking. The phenomenon should depend on only peptide concentration and their affinities. We therefore investigated a four-protein mix without any cross-linker addition and wondered if noncovalently associated peptides could be identified. Note that here we changed to a Q Exactive high field. Indeed, we identified 24 noncovalent peptide associations (Figure 4). The formation of NAP is thus also observable in linear proteomics that do not involve any cross-linking chemistry. However, the number of NAP identifications is low and unlikely to affect linear proteomics.

Since the involved forces leading to an interaction are expected to be rather weak, employing in-source collision-induced dissociation (ISCID) should reduce the number of identified NAPs. Using an ISCID of 0, 5, 10, 15, and 20 eV, we find 24, 11, 11, 6, and 3 NAP identifications at 5% PSM FDR (Figure 4a). Increasing the ISCID from 0 to 20 results in a 90% decrease of NAPs identifications. As a control, we also investigated how linear peptide identifications were affected by these voltages for ISCID and observed only a minor detrimental effect. Predominantly, we saw self-associations of the same peptide with all ISCID settings (88%, 64%, 73%, 33%, and 67%) for 0, 5, 10, 15, and 20 eV ISCID). Also, in cross-linked HSA we saw many self-links of peptides, which initially perplexed us as these would indicate protein dimerization despite us having isolated and analyzed the monomer. These cross-linked peptides now pose strong candidates for NAPs as well. This indicates that special care must be taken when homomultimers are investigated via CLMS. Note that homomultimers are not necessarily identified through cross-links of the same peptide in both instances of the protein. Cross-links involving overlapping peptide sequences can also indicate homomultimerization (see Figure S4).

We noticed a feature of MS2 spectra of NAPs that may help identify them in the future. The intact peptide peaks in multiple charge states up to the NAP's precursor charge state are frequently observed and are of high intensity (Figure 4a,b). We

DOI: 10.1021/acs.analchem.8b04037
Anal. Chem. 2019, 91, 2678−2685

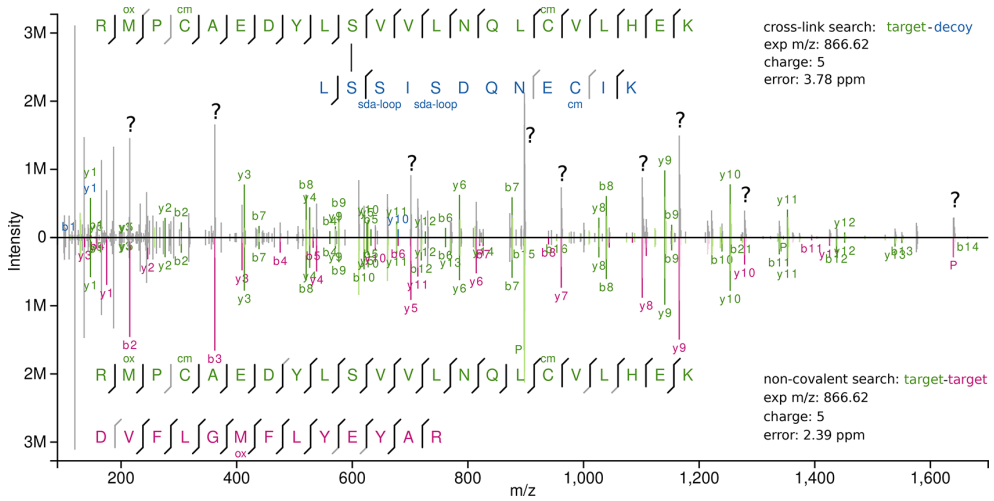


Figure 5. Butterfly plot of the same spectrum with different possible explanations. Upper panel shows the annotation from a cross-linking search (target−decoy identification). Lower panel shows the annotation from a noncovalent search (target−target identification). Q Exactive acquisition: raw file, *V127_K*; scan, 50038.

encountered this in 62% of cases for the ISCID data set of 0 eV. We are unaware of such charge-reduced precursor ions in HCD fragmentation spectra of linear peptides and do not see a single occurrence in our linear peptide data. This adds to NAPs being revealed at MS1 level through their overlapping elution with the individual linear peptides. It is unclear if NAP can be avoided altogether. However, critical assessment of the ionization settings appears to be advisable for CLMS analyses.

For the analysis of proteins via native MS, one should be aware that these unspecific associations might be possible too, even under "normal" LC conditions as we have used here. The exact conditions that support the formation of NAPs are not known.[29] However, previous studies found that electrostatic interactions lead to increased stability of noncovalent complexes,[30,31] but also solvent composition and ionization settings[29,32] are crucial. Likely, any parameter influencing the ionization such as instrument architecture and flow rates play a role. We therefore tested the influence of three flow rates on the formation of NAPs but found no differences within our experimental setup (Figure S3).

For cross-linking mass spectrometry experiments, NAPs pose a challenge. Cross-linking experiments using SDA or similar reagents are more susceptible to NAP identifications since the cross-linker can form loop-links on lysine residues, resulting in the same modification mass as a cross-linked peptide pair. However, the formation of NAPs does not depend on the cross-linker since we also observed their formation in non-cross-linked samples. Therefore, in theory, other cross-linkers will also lead to NAPs. A critical assessment of the specific instrument ionization settings is thus crucial for successful analysis of CLMS experiments. If the possible presence of NAPs is ignored, they will lead to wrong distance constraints. Even though structural-modeling approaches are to some extent robust to the number of false positives,[27] the influence of a systematic source of false positives is unknown. Experiments that aim to reconstruct the rough topology of protein complexes are at high risk of false conclusions being drawn from these false "cross-links". Wrong interprotein links and wrong intraprotein/loop-links might lead to inconclusive results. Therefore, we strongly suggest reducing the possibility of NAPs, either by optimizing acquisition settings or heuristic post-acquisition filters.

Significance of Noncovalently Associated Peptides. We observe NAPs here during the analysis of an SDA-cross-linked protein. While SDA is of central importance to high-density CLMS and the development of cross-linking for protein structure determination, this is a very young research area with currently few followers. Nevertheless, NAPs do not require the presence of SDA as we show by our analysis of a standard four-protein mix, without any cross-linking. The possible impact of NAPs goes into several directions, where few NAPs could make an impact. Self-association of loop-linked peptides would also occur with cross-linkers such as BS3 or DSS, leading to the possibility of misidentifying NAPs as cross-links. This would then lead to a false biological conclusion, namely, that a protein self-associates to form homodimers. Cleavable cross-linkers have the advantage that if a full set of signature peaks is observed, NAP formation can be ruled out. Unfortunately, the set of signature peaks is not always complete.[33] Second, our analysis showed that NAPs yield excellent spectra, often better than cross-linked peptides. When not considering NAPs, these good spectra can match only one of the associated peptides correctly, while for the second one the mass would be off by the assumed presence of a cross-link. This can lead only to a false target−target (TT) hit or target−decoy (TD) hit. Indeed, we found in our analysis an example (Figure 5) where a high-scoring TD from a cross-link search matched a TT during a NAP search with improved confidence. In routine analyses of protein complexes relatively few cross-links are being detected, so few high-scoring TDs may noticeably reduce the identified links. This was not the case in our analysis but should not be dismissed outright and warrants further attention. Finally, the presence of biologically not functional peptide−peptide complexes in the gas phase suggests that also the analysis of much larger proteins with many more interaction possibilities may lead to such nonbiological associations. Consequently, native mass spectrometry may require the development of appropriate controls as has been suggested before.[4]

DOI: 10.1021/acs.analchem.8b04037
Anal. Chem. 2019, 91, 2678−2685

CONCLUSION

Self-associations of peptides in solution has been shown to yield stable oligomers that endure the ionization process.[32] In addition, the preservation of noncovalent associations throughout ESI is exploited by native mass spectrometry. Here, we show that peptides with very similar chromatographic RT behavior can also remain together during the ionization process under normal liquid chromatography conditions as they are used in bottom-up proteomics. This implies that the association process can be unspecific and occur during normal LC-MS analysis. At the very least, the CLMS field should be aware of this. Pointing at ionization parameters and post-acquisition tests, we hope to assist the field in spotting and counteracting this effect.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.8b04037.

Conceptual drawings of cross-links and noncovalently associated peptides, results using cleavable cross-linking search software, and results from acquisitions with different flow rates (PDF)

AUTHOR INFORMATION

Corresponding Author

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

ORCID

Sven H. Giese: 0000-0002-9886-2447

Adam Belsom: 0000-0002-8442-4964

Lutz Fischer: 0000-0003-4978-0864

Juri Rappsilber: 0000-0001-5999-1310

Author Contributions

The manuscript was written through contributions of all authors.

Notes

The authors declare no competing financial interest.

The mass spectrometry raw files, peak lists, search engine results, MaxQuant parameter files, and FASTA files have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository[25] with the data set identifier PXD010895. In addition, the PSMs at 5% FDR are available online using xiVIEW:[34,35] Velos, HSA data (https://xiview.org/xi3/network.php?upload=34-08362-96692-34003-27750); QE, HSA data (https://xiview.org/xi3/network.php?upload=35-49786-17881-94522-79322); Q Exactive HF, protein mix (https://spectrumviewer.org/viewSpectrum.php?db=ISCID).

ACKNOWLEDGMENTS

This work was supported by the Einstein Foundation, the DFG [RA 2365/4-1, 25065445], and the Wellcome Trust through a Senior Research Fellowship to J.R. [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) Liko, I.; Allison, T. M.; Hopper, J. T.; Robinson, C. V. *Curr. Opin. Struct. Biol.* **2016**, *40*, 136−144.

(2) Boeri Erba, E.; Petosa, C. *Protein Sci.* **2015**, *24*, 1176−1192.

(3) Leney, A. C.; Heck, A. J. R. *J. Am. Soc. Mass Spectrom.* **2017**, *28*, 5−13.

(4) Smith, R. D.; Light-Wahl, K. J.; Winger, B. E.; Loo, J. A. *Org. Mass Spectrom.* **1992**, *27*, 811−821.

(5) Rappsilber, J. *J. Struct. Biol.* **2011**, *173*, 530−540.

(6) Yu, C.; Huang, L. *Anal. Chem.* **2018**, *90*, 144−165.

(7) Sinz, A. *Angew. Chem., Int. Ed.* **2018**, *57*, 6390−6396.

(8) Nguyen, H. T. H.; Andrikopoulos, P. C.; Rulíšek, L.; Shaffer, C. J.; Tureček, F. *J. Am. Soc. Mass Spectrom.* **2018**, *29*, 1706−1720.

(9) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551−3567.

(10) Maiolica, A.; Cittaro, D.; Borsotti, D.; Sennels, L.; Ciferri, C.; Tarricone, C.; Musacchio, A.; Rappsilber, J. *Mol. Cell. Proteomics* **2007**, *6*, 2200−2211.

(11) Chen, Z.; Fischer, L.; Tahir, S.; Bukowski-Wills, J.-C.; Barlow, P.; Rappsilber, J. *Wellcome Open Res.* **2016**, *1*, 5.

(12) Müller, F.; Fischer, L.; Chen, Z. A.; Auchynnikava, T.; Rappsilber, J. *J. Am. Soc. Mass Spectrom.* **2018**, *29*, 405−412.

(13) Rappsilber, J.; Ishihama, Y.; Mann, M. *Anal. Chem.* **2003**, *75*, 663−670.

(14) Ishihama, Y.; Rappsilber, J.; Andersen, J. S.; Mann, M. *J. Chromatogr. A* **2002**, *979*, 233−239.

(15) Cox, J.; Mann, M. *Nat. Biotechnol.* **2008**, *26*, 1367−1372.

(16) 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. *bioRxiv* **2018**, 355396.

(17) Giese, S. H.; Belsom, A.; Rappsilber, J. *Anal. Chem.* **2017**, *89*, 3802−3803.

(18) Fischer, L.; Rappsilber, J. *Anal. Chem.* **2017**, *89*, 3829−3833.

(19) Sugio, S.; Kashima, A.; Mochizuki, S.; Noda, M.; Kobayashi, K. *Protein Eng., Des. Sel.* **1999**, *12*, 439−446.

(20) Käll, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. *J. Proteome Res.* **2008**, *7*, 40−44.

(21) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. *Bioinformatics* **2008**, *24*, 2534−2536.

(22) Röst, H. L.; Schmitt, U.; Aebersold, R.; Malmström, L. *Proteomics* **2014**, *14*, 74−77.

(23) Oliphant, T. E. *Comput. Sci. Eng.* **2007**, *9*, 10−20.

(24) Cohen, J. *Statistical power analysis for the behavioral sciences*, 2nd ed.; Hillsdale, N., Ed.; Lawrence Erlbaum Associates: Hillsdale, NJ, 1988.

(25) 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. *Nucleic Acids Res.* **2016**, *44*, D447−D456.

(26) Belsom, A.; Schneider, M.; Brock, O.; Rappsilber, J. *Trends Biochem. Sci.* **2016**, *41*, 564−567.

(27) Belsom, A.; Schneider, M.; Fischer, L.; Brock, O.; Rappsilber, J. *Mol. Cell. Proteomics* **2016**, *15*, 1105−1116.

(28) Iacobucci, C.; Götze, M.; Piotrowski, C.; Arlt, C.; Rehkamp, A.; Ihling, C.; Hage, C.; Sinz, A. *Anal. Chem.* **2018**, *90*, 2805−2809.

(29) Chen, F.; Gülbakan, B.; Weidmann, S.; Fagerer, S. R.; Ibáñez, A. J.; Zenobi, R. *Mass Spectrom. Rev.* **2016**, *35*, 48−70.

(30) Jackson, S. N.; Moyer, S. C.; Woods, A. S. *J. Am. Soc. Mass Spectrom.* **2008**, *19*, 1535−1541.

(31) Woods, A. S.; Ferré, S. *J. Proteome Res.* **2005**, *4*, 1397−1402.

(32) Chitta, R. K.; Gross, M. L. *Biophys. J.* **2004**, *86*, 473−479.

(33) Liu, F.; Lössl, P.; Scheltema, R.; Viner, R.; Heck, A. J. R. *Nat. Commun.* **2017**, *8*, 15473.

(34) Combe, C. W.; Fischer, L.; Rappsilber, J. *Mol. Cell. Proteomics* **2015**, *14*, 1137−1147.

(35) Kolbowski, L.; Combe, C.; Rappsilber, J. *Nucleic Acids Res.* **2018**, *46*, W473−W478.

NOTE ADDED AFTER ASAP PUBLICATION

This paper was originally published ASAP on February 1, 2019. Due to a production error, "Affect" was incorrectly used in the title. The corrected version was reposted on February 5, 2019.

DOI: 10.1021/acs.analchem.8b04037
Anal. Chem. 2019, 91, 2678−2685

Chapter 5

Manuscript 4. Peptide Retention in Hydrophilic Strong Anion Exchange Chromatography Is Driven by Charged and Aromatic Residues

analytical chemistry

Article

Cite This: Anal. Chem. 2018, 90, 4635−4640

pubs.acs.org/ac

Peptide Retention in Hydrophilic Strong Anion Exchange Chromatography Is Driven by Charged and Aromatic Residues

Sven H. Giese,[†] Yasushi Ishihama,[‡] and Juri Rappsilber*[,†,‡,§]

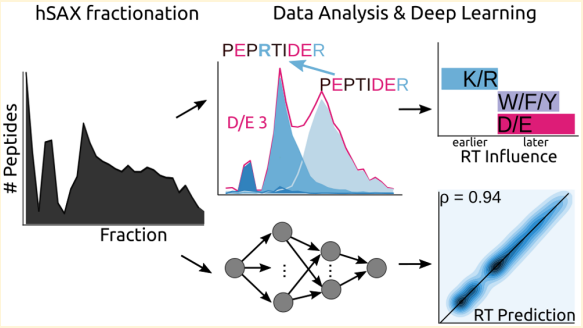
[†]Bioanalytics, Institute of Biotechnology, Technische Universität Berlin, 13355 Berlin, Germany

[‡]Graduate School of Pharmaceutical Sciences, Kyoto University, Kyoto 606-8501, Japan

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

Supporting Information

ABSTRACT: Hydrophilic strong anion exchange chromatography (hSAX) is becoming a popular method for the prefractionation of proteomic samples. However, the use and further development of this approach is affected by the limited understanding of its retention mechanism and the absence of elution time prediction. Using a set of 59 297 confidentially identified peptides, we performed an explorative analysis and built a predictive deep learning model. As expected, charged residues are the major contributors to the retention time through electrostatic interactions. Aspartic acid and glutamic acid have a strong retaining effect and lysine and arginine have a strong repulsion effect. In addition, we also find the involvement of aromatic amino acids. This suggests a substantial contribution of cation−π interactions to the retention mechanism. The deep learning approach was validated using 5-fold cross-validation (CV) yielding a mean prediction accuracy of 70% during CV and 68% on a hold-out validation set. The results of this study emphasize that not only electrostatic interactions but rather diverse types of interactions must be integrated to build a reliable hSAX retention time predictor.



Mass spectrometry (MS)-based proteomics is the driving technology for the characterization and quantification of complex protein samples.[1−3] With the current advancements in instrumentation and software solutions, the number of peptides and proteins that can be identified in a minimal amount of time have increased dramatically.[4] However, deep proteome coverage of higher eukaryotes, mammalian cell lines, or tissue is currently only feasible with extensive fractionation.[5,6] The wide dynamic range of all the expressed proteins in a cell remains a major challenge, leaving the least abundant proteins (and peptides) undiscovered. In these cases, online (1D) reverse phase liquid chromatography (RP-LC) does not yield the necessary separation of the proteome. Instead, prefractionation is commonly applied to further reduce the complexity. Ideally, the combined separation methods are as orthogonal as possible[5,7,8] to ensure the separation of similar analytes. Interestingly, high-pH RP is often used as prior fractionation method even though it is not truly orthogonal to standard RP (low pH). Importantly, there is no universal best prefractionation method. Rather, the optimal separation method needs to be chosen based on the analytes.[9,10]

While fractionation methods offer great possibilities to reduce the sample complexity, they usually require larger sample amounts and preparation time. Usually, most fractions are injected separately without pooling. Therefore, the peptide identification is fraction aware. This extra piece of information can be incorporated into the database search.[11−13] To fully utilize this information, a computational model needs to be developed that can confidently predict the retention time of a peptide based on its amino acid sequence. The proteomics community has successfully developed accurate models for the prediction of the retention time in low pH RP-LC, which typically is coupled directly to a mass spectrometer and therefore widely applied in proteomics.[14,15] Retention times have also been predicted for other chromatographic methods including high-pH RP-LC,[16,17] hydrophilic interaction liquid chromatography (HILIC),[18] and strong cation exchange chromatography (SCX).[19] Various algorithms have been applied for the described prediction task: simple linear regression models,[20] nonlinear models,[21] support vector regression models,[11,16] artificial neural networks,[22] or a physical model describing the chromatographic process.[23] For a comprehensive review, the reader is referred to Tarasova et al.[14] and Moruz and Käll.[15]

For standard shotgun proteomics, hydrophilic strong anion exchange chromatography (hSAX) is largely orthogonal to RP-LC.[5] Currently, there is no model to predict the retention time for hSAX. Moreover, the sequence specific features that

Received: December 11, 2017
Accepted: March 12, 2018
Published: March 12, 2018

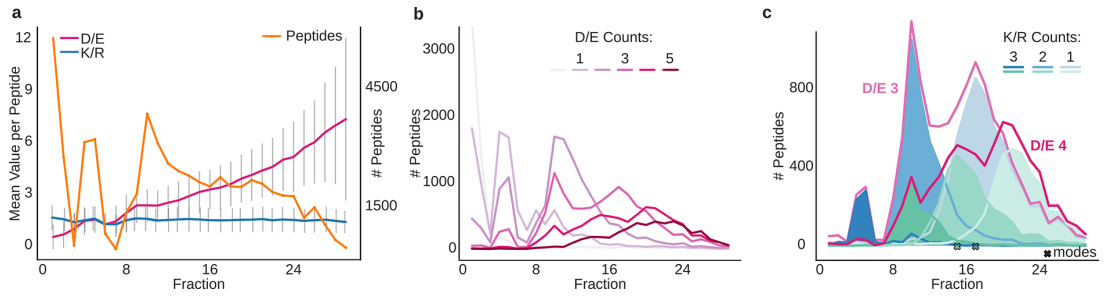


Figure 1. Effect of the charged residues on peptide retention in hSAX. (a) Mean residue count per peptide for D/E (red) and K/R (blue) over fraction. Error bars denote the standard deviation. Peptide count per fraction is shown in orange (total 59 297 unique peptides). (b) Effect of D/E count (range 0−5) on peptide retention. (c) Extracted chromatogram of peptides with three and four D/E (red). Subpopulations were defined according to the number of K/R residues (one to three, blue tones for peptides with three D/E residues and green tones for peptides with four D/E residues). Crosses mark the mode of the respective distributions.

influence the retention behavior of peptides during hSAX are still unknown. A common approach is to incorporate (limited) sequence information into the prediction model by creating position specific retention coefficients[18] or neighboring amino acid effects.[24] It would be desirable to (1) better understand the mechanisms governing the retention behavior of peptides during hSAX and (2) build a predictive machine learning model that confidently predicts the retention time of a peptide based on its sequence information.

In this study, we analyzed the chromatographic behavior of 59 297 peptides based on 29 hSAX fractions. We aim to contribute new insights into the interaction of peptides during hSAX and quantify how sequence features affect the retention behavior. To accomplish this, a machine learning workflow is applied and validated using 5-fold cross-validation. We developed a neural network model that predicts the retention time for peptides from an hSAX fractionation. The predictive model and the preprocessing are available in the Python package DePART (https://github.com/Rappsilber-Laboratory/DePART).

METHODS

Experimental Details. The experimental data taken for this study were published by Ritorto et al.[5] In brief, the authors performed hydrophilic strong anion exchange (hSAX) chromatography on macrophage cells from *Mus musculus* to test the peptide separation capabilities of hSAX followed by mass spectrometry. The tryptic digest of the cell lysate was analyzed with a LTQ Orbitrap Velos Pro (Thermo Fisher Scientific, West Palm Beach, FL). The fractionation was performed using an Ion Pac AS24[25] column (2 × 250 mm, 2000 Å pore size, Thermo Fisher Scientific, Part No.: 064153) with a 35 min gradient (0 to 1 M NaCl; solvent A, 20 mM Tris-HCl at pH 8.0; solvent B, 20 mM Tris-HCl at pH 8.0, 1 M NaCl). The functional group of the AS24 is an alkanol quaternary ammonium ion on a solid support that aims at minimal hydrophobicity. Details of the sample preparation protocols can be found in the original manuscript.[5]

Data Processing. For our study, Ritorto et al. made the results of their previous experiments available as MaxQuant result files. We postprocessed the MaxQuant evidence file. In total, 466 495 peptides were identified in 34 fractions. We applied stringent filtering to avoid ambiguity in the training data. This initial set of peptides was reduced by removing contaminants, decoys, "only by site" identifications, and modified peptides (other than carbamidomethylated cysteine). In addition, for peptides identified in two adjacent fractions, the identification with the lowest intensity was removed from the data set. Peptides identified in more than two fractions or in fractions that were not adjacent were also removed from the data. Finally, fractions with less than 300 unique peptide identifications were removed—leaving 59 297 unique peptide sequences distributed over 29 fractions for the data analysis. As an independent data set, we used PXD006188,[26] which was analyzed using MaxQuant[27] (v. 1.6.1.0) and filtered as described above, resulting in 93 372 peptides being identified in 32 fractions.

All processing was performed using Python 3.5 using the packages numpy, scipy, matplotlib, scikit-learn, pandas, and seaborn.

Machine Learning. For the computational modeling of the retention time we followed two separate strategies, a regression and a classification approach. In the regression case, a simple linear model (LM) with a length correction parameter (LCP) was used. The Python package pyteomics[20] with LCP optimization was used for the LM implementation. In the classification case, a logistic regression (LR) and a feedforward neural network (FNN) were used. In both cases, we evaluated (and trained) the model using the accuracy metric, defined as the proportion of correctly predicted fractions from all predictions. With the LM, such a metric is ill-defined since no discrete fraction is predicted. Therefore, we defined a forced accuracy metric by first rounding the predictions to the nearest integer and then computing the accuracy.

The FNN was implemented using Keras[28] with the Theano[29] backend. The network architecture consisted of four fully connected layers with 50, 40, 35, and 29 neurons. As final activation, the softmax function was used (Table S4). One strength of the simple additive model is the intuitive interpretation of the learned coefficients: a peptide's elution time increases (or decreases) by a certain factor based on the amino acid count. For neural networks, with nonlinear activation functions, the interpretation is not as straightforward. Therefore, we added peptide features (e.g., pI or aromaticity) based on the literature[11,30] and our initial exploratory data analysis to increase the predictive power in the classification task. The complete definition of features is available in Table S2.

The evaluation of the prediction performance was based on a 5-fold cross-validation (CV) strategy (including 75% of the data, 44 471 peptides). In addition, a hold-out validation set was used for the final model assessment (25% of the data,

DOI: 10.1021/acs.analchem.7b05157
Anal. Chem. 2018, 90, 4635−4640

14 825 peptides). In the CV setup, the training splits had 35 578 observations, and the validation splits had 8894 observations. We describe the machine learning workflow in more detail in the Supporting Information, including a performance comparison with other classifiers.

RESULTS

In the following section, we present our results and propose a model for the driving interactions in hydrophilic strong anion exchange chromatography (hSAX) for peptides. The result section is divided into four parts: (1) A general overview is given of the data and how the retention time during prefractionation is influenced by charged amino acids. (2) The influence of the charged amino acids is compared. (3) The influence of usually noncharged amino acids is compared, and finally, (4) a machine learning model is built to model peptide retention during hSAX.

Peptide Retention in hSAX Is Driven by the Charged Amino Acids. We first investigated the influence of acidic (E, D) and basic (K, R) amino acids on the retention behavior of peptides in an hSAX fractionation experiment. Note that histidine residues will be uncharged under the pH conditions used during fractionation. We used elution data of 59 297 tryptic peptides from murine macrophage cells separated into 29 fractions. Positively charged peptides elute early (fractions 1 and 2) and are separated from uncharged peptides (fractions 4 and 5) which in turn are separated from negatively charged peptides (fractions 7−29), where charge was calculated from the residues E, D, K, and R (Figure 1a).

While the mean count of D or E (D/E) residues in a peptide increases with the fraction number, the mean count of K/R residues stays constant (Figure 1a). In agreement with this, missed cleavages are not enriched in any of the fractions (Figure S1). The average retention behavior of tryptic peptides appears to be mainly influenced by the occurrences of D/E residues in the peptide sequence. These observations are also supported numerically by their Pearson correlation coefficients (PCC) of the summed residue charge per peptide and the observed fraction number: for D/E residues, the PCC is −0.75; for K/R, −0.03; and for D/E/K/R residues, the PCC is −0.83. The peptide length on the other hand has a much smaller overall influence across all fractions (PCC 0.33). Peptides with 0, 1, 2, 3, 4, and 5 D/E residues correspond on average to the fractions 3, 6, 10, 14, 18, and 20, respectively (Table S1), thus, leading to a mean increase per D/E residue of three fractions in retention time.

Even though the mean increase of fraction numbers highly correlates with the number of acidic residues, so does the D/E peak width (Figure 1b). In addition, the higher the number of D/E residues in the peptide, the more complex the distributions appear. Peptides with two D/E residues distribute on two peak fractions, while peptides with four D/E residues distribute on four to six peak fractions.

Therefore, we investigated the influence of basic residues on the retention time. Positively charged residues, lysine and arginine, should weaken peptide retention during hSAX. Indeed, K and R residues explain the multiple peak fractions of peptides with one D/E (Figure 1c). With an increasing number of K/R residues, the retaining effect of D/E diminishes, and thus peptides elute earlier. Since the effect is quite strong, in terms of retention shift by a single K/R residue, there is most likely a repulsion mechanism involved. Interestingly, the elution strength of K/R residues seems slightly stronger than the retaining effect of D/E residues: The mean fraction value of peptides with four D/E residues and two K/R residues (summed residue charges equal to 2) is 16.5, while for peptides with three D/E residues and one K/R (summed residue charge also equal to 2), the mean fraction is 18.1. However, this additional information on the K/R distribution does not fully explain the observed substructures; there are clearly peak tails visible, especially on the right side of the distributions (e.g., D/E, 4; K/R, 3 in Figure 1c).

Lysine Exhibits Stronger Electrostatic Repulsion than Arginine. We next evaluated if R and K differed in their effect on peptide retention (Figure 2a). Peptides with four D/E

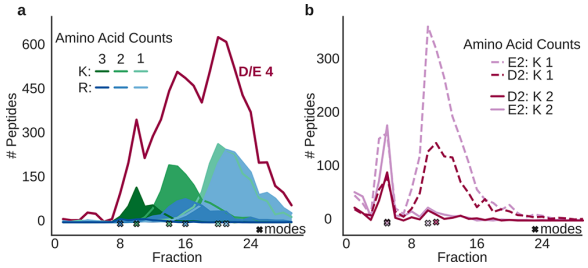


Figure 2. Detailed comparison of relative contributions of positively (K/R) and negatively (D/E) charged residues on peptide retention in hSAX. (a) Effect size of K/R residues. Peptides with four D/E residues were divided according to their K and R count (K, green tones; R, blue tones). (b) Effect size of E/D residues. Peptides with either two E or two D residues are shown, split according to their number of K residues (1 or 2).

residues were found in the factions 22, 17, and 11 (median fraction values) if they had one, two, or three arginines while they were found in the fractions 21, 15, and 10 if they had one, two, or three lysines. This means that lysines are more strongly repelled than arginines in hSAX (on average, 1.3 fractions). Statistical analysis using a Mann−Whitney−U (MWU) test supports this observation. However, since the observed effect size is rather small, the statistical significance should be interpreted with caution (Figure S2a).

Similarly, we investigated possible differences between aspartate and glutamate, peptides with either two D or two E residues and either one, two, or three lysines (Figure 2b shows data for up to two lysines). For this subset, the rounded median fraction number for peptides with two D or two E residues is 12, 11, and 5 and 12, 11, and 5, respectively. This leads to an average increase of 0.33 per fraction if there is an aspartate instead of a glutamate in the peptide sequence. For the negatively charged amino acids, we also conducted an MWU-test: although the observable effect was even smaller, the test still resulted in a significant difference between the retention behavior of D and E (Figure S2b).

Aromatic Amino Acids Play a Key Role in Peptide Retention during hSAX. As expected, peptide retention during hSAX is dominated by charged residues. However, peptides with one set of charged residues elute over many fractions. Therefore, charged amino acids do not suffice to explain peptide retention alone.

As a first step to search for additional contributions, a subset of peptides was selected (two D/E residues, one R/K residue). Then, the effect size of an amino acid on the retention time was estimated using the slope from a linear regression model. The response variable was set to the mean composition contribution

DOI: 10.1021/acs.analchem.7b05157
Anal. Chem. 2018, 90, 4635−4640

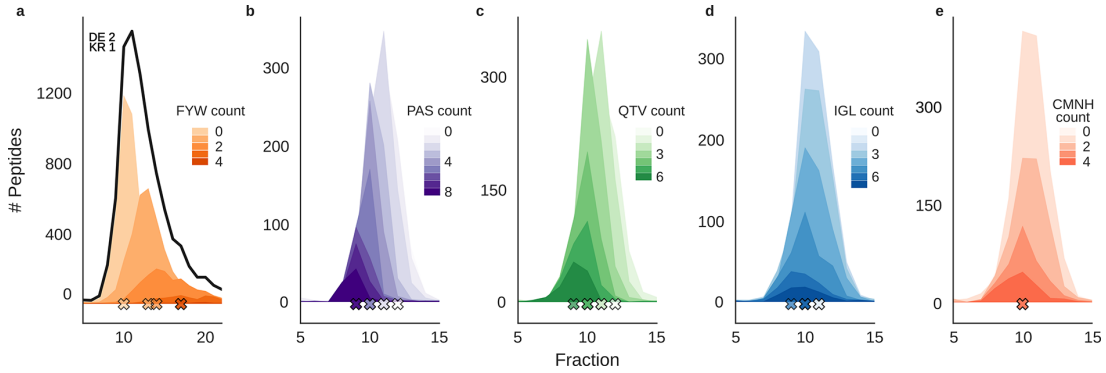


Figure 3. The effect of neutral amino acids on peptide retention in hSAX. Amino acids were grouped according to their influence on peptide retention in hSAX by linear regression (Supporting Information). (a) Elution behavior of peptides with different numbers of F/Y/W and two D/E, one K/R residues. (b–e) Elution behavior of peptides with different numbers of the indicated amino acids (b, P/A/S; c, Q/T/V; d, I/G/L; e, C/M/N/H) and two D/E, one K/R, zero F/Y/W. Crosses mark the mode of the subpopulations.

of an amino acid, while the explanatory variable was set to the fraction number. On the basis of the regression slope and the derived p-value (under the null hypothesis that the slope is equal to zero), the remaining amino acids can be divided into three categories: (1) retaining—if the slope is positive and the p-value is smaller than 0.05, (2) eluting—if the slope is negative and the p-value is smaller than 0.05, and (3) no (significant) effect—if the p-value is larger than 0.05. Accordingly, the (aromatic) amino acids F, Y, and W show the strongest retaining effect based on the regression slope (Figure 3, Figure S4). Interestingly, peptides with 0 aromatic residues are found in a sharp symmetrical distribution. With increasing aromatic amino acids in the peptide sequence, the distributions shift to later retention, become broader, and develop a right tail (Figure S6). In contrast, the amino acid contributions of A, P, and S and Q, T, and V show an eluting effect. For these amino acids, the subpopulation peaks look very sharp, even with increasing residues of the same group. The remaining amino acids C, I, N, G, L, V, H, and M do not show a clear trend and thus could be classified neither as eluting nor as retaining. Subtracting the weighted counts of the aromatic residues (0.8W + 0.6Y + 0.3F) to the residue charge increases the initial PCC from −0.83 to −0.86. Adding the weighted counts of the residues A, P, Q, S, T, and V (factor 0.1) further increases the retention PCC to −0.88.

A Neural Network Achieves the Highest Prediction Accuracy. As the final step in our analysis, we built a machine learning model to predict the retention time of a peptide based on its sequence features. After initial hyperparameter optimization for a set of classifiers and regressors (Supporting Information S3), we chose a linear regression model (LM), a logistic regression model (LR), and a feedforward neural network (FNN) for further analysis. The coefficients of the LM are shown in Figure 4a. As expected, the sign and magnitude of the coefficients largely match our manual analysis: First, the basic residues have a strong eluting effect on the retention time (large negative coefficient). Second, the acidic residues and the aromatic residues have a strong retaining effect on the retention time (large positive coefficient). In addition, the nuances regarding the effect size of the basic residues also fit our previous description that R is marginally stronger repelled than K. This is most likely due to the lower basicity of K. Similar to the coefficient representation from LM, FNNs can be used to estimate approximately the influence of the input features by

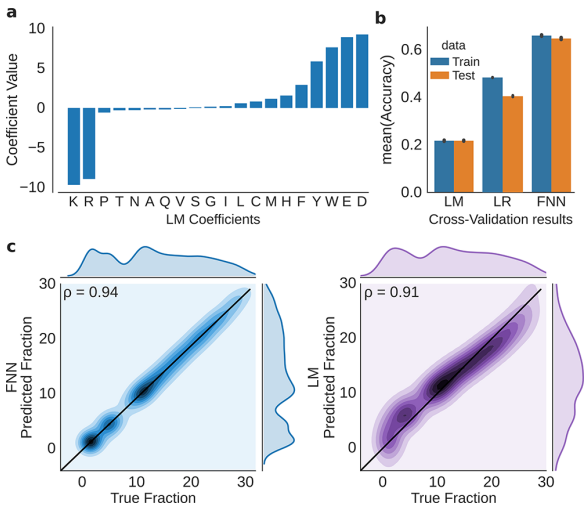


Figure 4. Peptide retention time prediction for hSAX using machine learning. (a) Residue retention coefficients from a linear model with length correction parameter. (b) Fraction of correct predictions (accuracy) of different prediction methods, estimated by 5-fold cross-validation based on 35 578 (train) and 8894 (test) peptides in each split. (c) Elution time prediction for the hold-out validation set, FNN classifier (left) and LM (right); ρ indicates the Pearson correlation. Linear Model (LM), Logistic Regression (LR), Feedforward Neural Network (FNN).

analyzing the input weights of the first layer. Since we also used position specific features in the machine learning workflow, the average of the input weights can be used to roughly measure these position dependent contributions to the retention in hSAX. Most importantly, it appears that the influence of D/E residues decreases with distance from the termini (Figure S7). Further, S/T/V/A/P/Q residues roughly follow a similar trend. In contrast, W/Y/F/H do not show decreasing weights for internal residues—the influence is rather stable across the positions. For the remaining amino acids (I/G/L/C/M/N), the weights are noisy and do not follow a clear pattern. This observation fits the estimation of their influence from the regression model. Therefore, the influence of these amino acids cannot be clearly defined.

DOI: 10.1021/acs.analchem.7b05157
Anal. Chem. 2018, 90, 4635−4640

A neural network was most successful in predicting the correct peptide fraction, as assessed by 5-fold cross-validation (Figure 4b). With an accuracy of 70 ± 0.81% (mean ± standard error of the mean), the classification algorithm outperformed the linear regression model (22 ± 0.13% accuracy) and the logistic regression model (48 ± 0.07% accuracy). With a lower prediction resolution, e.g., evaluating the accuracy in a window of ±1 fraction (1-off-accuracy), 92 ± 0.19% were correctly classified. Although optimization aimed for accuracy, the best performing FNN classifier also achieves a higher correlation coefficient on a hold-out validation set (never used for training) than the LM. The FNN achieves here a PCC of 0.94 where the LM achieves a PCC of 0.9 (Figure 4c). The accuracy on this validation set was comparable to the CV error with 68% accuracy and 92% one-off accuracy. As the accuracy metric already indicates, the LM performs much worse as seen in the marginal distributions (Figure 4c). The distribution of the predicted fractions does not appear similar to the observed fraction distribution. The FNN can better capture the nonlinear relationship and thus predicts the true fraction with a higher accuracy—which is supported by the similarity of the marginal distributions of the predicted and true fractions of the peptides in the validation set.

Finally, we wondered if the results obtained for data by Ritorto et al. would also be obtained with a different data set by independent investigators. We downloaded an hSAX data set from ProteomeXchange (PXD00618826) and repeated our analysis. For these data, the training set comprised 70 029 unique peptides and the validation set, 23 343 unique peptides. The accuracy during CV increased on the test data to 69 ± 0.21% and on the validation data to 72%. The one-off accuracy even increased to 96%, most likely due to higher number of training instances.

DISCUSSION

Fractionation methods such as ion exchange chromatography (IEX) are popular tools for enrichment of certain analytes and separation of complex samples. To perfect the separation process, a basic understanding of the underlying principles must be developed. For the principles behind the retention time of peptides in hSAX chromatography, a linear model is a useful starting point.

Our exploratory analysis as well as the modeling approach showed that electrostatic forces, as expected, are the most important interactions in hSAX. A previous study that compared several fractionation methods for phosphopeptides also reported a strong correlation of the acidic amino acids with the elution time of peptides.[9] The resolution based on simply counting the D/E/R/K residues is enough to roughly map the elution time of a peptide to ±5 fractions (on average). This simple approach is supported by a good PCC (−0.83) of the summed residue charge and the elution time. However, differentiating the repelling (K/R) and retaining (D/E) effect sizes should further improve the resolution. Additional improvements can be achieved by including the influence of the aromatic amino acids (W, Y, F; PCC −0.86).

The retaining effect of the aromatic amino acids could be explained through cation−π interactions: a well-known interaction from organic chemistry. Since aromatic amino acids have a delocalized π electron system, the flat face of the aromatic ring has a partial negative charge which attracts cations and thus enables strong electrostatic interactions.[31,32] Cation−π interactions are also essential for many biological processes and protein folding, in which K/R residues can also function as cations and thus reinforce bonds within a protein structure. Possibly, cation−π interactions also happen within a single peptide and therefore lead to a competition between the stationary phase and the side chains of K/R. Multiple aromatic amino acids in a peptide sequence lead to nonlinearity in the retention behavior, i.e., multiple aromatic amino acids support the interactions with the stationary phase more than expected from adding individual contributions, possibly by forming sandwich complexes of two aromatic amino acids and a cation.

For tryptic phosphopeptides, it has been shown that the peptide C-terminus is likely oriented toward the stationary phase[33] during the separation in anion exchange chromatography. Presumably, this also holds true for peptides in hSAX. However, comparing the neural network weights revealed that the influence of, e.g., D or E residue is not per se decreasing from the N-terminus to the C-terminus as has been observed for the SCX model.[33] Thus, it is possible that the peptide orientation in hSAX is bidirectional—or that D/E residues show a different elution behavior when near the termini. If the orientation of the peptide is indeed with the N-terminus toward the stationary phase, the decrease of the neural network weights is explainable with the limited accessibility of the acidic side chains when the residue is buried in the sequence. The same argumentation holds true for the orientation of the N-terminus toward the stationary phase. However, since we only analyzed tryptic peptides with basic side chains on the C-terminus, it seems unlikely that they would prefer this orientation. Another hypothesis is that the influence of C-terminal D/E residues is not directly through the interaction of the residues with the column but through intrapeptide interactions. For example, acidic side chains of D/E and basic side chains of K/R could form salt bridges. Thus, the closer the D/E residues are to the C-terminus, the larger is the contribution or effect in the determination of the retention time.

The retention time prediction field is fairly mature, and a selection of published tools achieved an $R^2 \geq 0.90$, according to a recent literature review.[14] While most solutions achieve a very high correlation (and R^2), the true accuracy (defined as true predictions/(true + false predictions)) is seldom evaluated. The models used to predict the fraction either do not provide an easily accessible probability or prefer to model the prediction task as a regression problem[19] allowing R^2 to be calculated. We modeled the prediction in a classification setup, using a feed-forward neural network (FNN). Here, accuracy is an appropriate evaluation metric. Accuracy is used to evaluate classification problems, and the algorithm was trained to optimize the accuracy and not R^2. With the current implementation, the FNN achieved an accuracy of 70 ± 0.81% during CV and 68% on the hold-out validation set. The accuracy is a stricter metric than the correlation coefficient or R^2; the one-off accuracy increases on the CV data set to 92 ± 0.19% and on the hold-out validation data set to 92%. One additional advantage of the FNN is that each prediction is associated with a probability. This is a useful feature since it allows selection of more confident predictions or incorporation of the uncertainty in postprocessing.

CONCLUSION

We presented a first description of the parameters that influence the retention of peptides during hSAX chromatography. As expected, the charged amino acids largely define the retention behavior of tryptic peptides. However, the aromatic

DOI: 10.1021/acs.analchem.7b05157
Anal. Chem. 2018, 90, 4635−4640

amino acids also have a large impact on the retention behavior presumably through cation−π interactions, which makes the retention mechanism of hydrophilic anion exchange chromatography more challenging to describe. Nevertheless, the proposed neural network model achieves a high accuracy of 68% on the hold-out validation set paired with a high correlation value of 0.94—which enables the usage of our model for statistical modeling of the confidence of peptide identifications based on prefractionation. In the future, we want to further improve our model with more training data, support for post-translational modifications, and incorporation into a robust scoring metric for peptide identification.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.7b05157.

Missed cleavage data, statistical comparison of the effect size of K/R and D/E residues, amino acid classification and details on the machine learning workflow (PDF)

■ AUTHOR INFORMATION

Corresponding Author

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

ORCID

Sven H. Giese: 0000-0002-9886-2447

Yasushi Ishihama: 0000-0001-7714-203X

Juri Rappsilber: 0000-0001-5999-1310

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Matthias Trost (Newcastle, United Kingdom) for providing MaxQuant result files and Michael Bohlke-Schneider for fruitful discussions. This work was supported by the Wellcome Trust through a Senior Research Fellowship to J.R. [103139], a JSPS Invitational Fellowship for Research in Japan No. L16568 to J.R. and Y.I., and JSPS Grants-in-Aid for Scientific Research No. 17H05667 and 16K15107 to Y.I. The Wellcome Centre for Cell Biology is supported by core funding from the Wellcome Trust [203149].

■ REFERENCES

(1) Aebersold, R.; Mann, M. *Nature* **2003**, *422*, 198−207.
(2) Ong, S.-E.; Mann, M. *Nat. Chem. Biol.* **2005**, *1*, 252−262.
(3) Yates, J. R.; Ruse, C. I.; Nakorchevsky, A. *Annu. Rev. Biomed. Eng.* **2009**, *11*, 49−79.
(4) Hebert, A. S.; Richards, A. L.; Bailey, D. J.; Ulbrich, A.; Coughlin, E. E.; Westphall, M. S.; Coon, J. J. *Mol. Cell. Proteomics* **2014**, *13*, 339−347.
(5) Ritorto, M. S.; Cook, K.; Tyagi, K.; Pedrioli, P. G. A.; Trost, M. *J. Proteome Res.* **2013**, *12*, 2449−2457.
(6) Manadas, B.; Mendes, V. M.; English, J.; Dunn, M. J. *Expert Rev. Proteomics* **2010**, *7*, 655−663.
(7) Dowell, J. A.; Frost, D. C.; Zhang, J.; Li, L. *Anal. Chem.* **2008**, *80*, 6715−6723.
(8) Yang, F.; Shen, Y.; Camp, D. G.; Smith, R. D. *Expert Rev. Proteomics* **2012**, *9*, 129−134.
(9) Alpert, A. J.; Hudecz, O.; Mechtler, K. *Anal. Chem.* **2015**, *87*, 4704−4711.
(10) Leitner, A.; Reischl, R.; Walzthoeni, T.; Herzog, F.; Bohn, S.; Förster, F.; Aebersold, R. *Mol. Cell. Proteomics* **2012**, *11*, M111.014126.
(11) Moruz, L.; Tomazela, D.; Käll, L. *J. Proteome Res.* **2010**, *9*, 5209−5216.
(12) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. *Nat. Methods* **2007**, *4*, 923−925.
(13) Klammer, A. A.; Yi, X.; MacCoss, M. J.; Noble, W. S. *Anal. Chem.* **2007**, *79*, 6111−6118.
(14) Tarasova, I. A.; Masselon, C. D.; Gorshkov, A. V.; Gorshkov, M. V. *Analyst* **2016**, *141*, 4816−4832.
(15) Moruz, L.; Käll, L. *Mass Spectrom. Rev.* **2017**, *36*, 615−623.
(16) Pfeifer, N.; Leinenbach, A.; Huber, C. G.; Kohlbacher, O. *J. Proteome Res.* **2009**, *8*, 4109−4115.
(17) Dwivedi, R. C.; Spicer, V.; Harder, M.; Antonovici, M.; Ens, W.; Standing, K. G.; Wilkins, J. A.; Krokhin, O. V. *Anal. Chem.* **2008**, *80*, 7036−7042.
(18) Krokhin, O. V.; Ezzati, P.; Spicer, V. *Anal. Chem.* **2017**, *89*, 5526−5533.
(19) Gussakovsky, D.; Neustaeter, H.; Spicer, V.; Krokhin, O. V. *Anal. Chem.* **2017**, *89*, 11795.
(20) Goloborodko, A. A.; Levitsky, L. I.; Ivanov, M. V.; Gorshkov, M. V. *J. Am. Soc. Mass Spectrom.* **2013**, *24*, 301−304.
(21) Krokhin, O. V. *Anal. Chem.* **2006**, *78*, 7785−7795.
(22) Petritis, K.; Kangas, L. J.; Yan, B.; Monroe, M. E.; Strittmatter, E. F.; Qian, W.-J.; Adkins, J. N.; Moore, R. J.; Xu, Y.; Lipton, M. S.; et al. *Anal. Chem.* **2006**, *78*, 5026−5039.
(23) Gorshkov, A. V.; Tarasova, I. A.; Evreinov, V. V.; Savitski, M. M.; Nielsen, M. L.; Zubarev, R. A.; Gorshkov, M. V. *Anal. Chem.* **2006**, *78*, 7770−7777.
(24) Moruz, L.; Staes, A.; Foster, J. M.; Hatzou, M.; Timmerman, E.; Martens, L.; Käll, L. *Proteomics* **2012**, *12*, 1151−1159.
(25) Pohl, C.; Saini, C. *J. Chromatogr. A* **2008**, *1213*, 37−44.
(26) Yu, P.; Petzoldt, S.; Wilhelm, M.; Zolg, D. P.; Zheng, R.; Sun, X.; Liu, X.; Schneider, G.; Huhmer, A.; Kuster, B. *Anal. Chem.* **2017**, *89*, 8884−8891.
(27) Cox, J.; Mann, M. *Nat. Biotechnol.* **2008**, *26*, 1367−1372.
(28) Chollet, F.; et al. *Keras*, 2015.
(29) Al-Rfou, R.; Alain, G.; Almahairi, A.; Angermueller, C.; Bahdanau, D.; Ballas, N.; Bastien, F.; Bayer, J.; Belikov, A.; Belopolsky, A.; et al. *arXiv e-prints*, 2016, abs/1605.0.
(30) Krokhin, O. V. *Anal. Chem.* **2006**, *78*, 7785−7795.
(31) Dougherty, D. A. *Science* **1996**, *271*, 163−168.
(32) Dougherty, D. a. *J. Nutr.* **2007**, *137*, 1504S−1508S discussion 1516S−1517S.
(33) Alpert, A. J.; Petritis, K.; Kangas, L.; Smith, R. D.; Mechtler, K.; Mitulović, G.; Mohammed, S.; Heck, A. J. R. *Anal. Chem.* **2010**, *82*, 5253−5259.

DOI: 10.1021/acs.analchem.7b05157
Anal. Chem. 2018, 90, 4635−4640

Chapter 6

Manuscript 5. Retention time prediction using neural networks increases identifications in crosslinking mass spectrometry

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-23441-0 OPEN

Retention time prediction using neural networks increases identifications in crosslinking mass spectrometry

Sven H. Giese [1,2,3,5], Ludwig R. Sinn [1,5], Fritz Wegner [1] & Juri Rappsilber [1,4]✉

Crosslinking mass spectrometry has developed into a robust technique that is increasingly used to investigate the interactomes of organelles and cells. However, the incomplete and noisy information in the mass spectra of crosslinked peptides limits the numbers of protein–protein interactions that can be confidently identified. Here, we leverage chromatographic retention time information to aid the identification of crosslinked peptides from mass spectra. Our Siamese machine learning model xiRT achieves highly accurate retention time predictions of crosslinked peptides in a multi-dimensional separation of crosslinked *E. coli* lysate. Importantly, supplementing the search engine score with retention time features leads to a substantial increase in protein–protein interactions without affecting confidence. This approach is not limited to cell lysates and multi-dimensional separation but also improves considerably the analysis of crosslinked multiprotein complexes with a single chromatographic dimension. Retention times are a powerful complement to mass spectrometric information to increase the sensitivity of crosslinking mass spectrometry analyses.

[1] Bioanalytics, Institute of Biotechnology, Technische Universität Berlin, Berlin, Germany. [2] Data Analytics and Computational Statistics, Hasso Plattner Institute for Digital Engineering, Potsdam, Germany. [3] Digital Engineering Faculty, University of Potsdam, Potsdam, Germany. [4] Wellcome Centre for Cell Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, UK. [5] These authors contributed equally: Sven H. Giese, Ludwig R. Sinn. ✉email: juri.rappsilber@tu-berlin.de

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

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

The identification of cross-linked peptides from spectra is however still challenged by the uneven fragmentation of the two peptides and the large search space that increase the odds of random matches. This is especially the case for heteromeric crosslinks as the size of their search space exceeds that of self-links, i.e., links falling within a protein or homomer[16]. Typically, database search tools use the precursor mass and fragmentation spectrum for the identification of peptides to compute a single final score for each CSM. For linear peptides, post-search methods such as Percolator[18] have been developed that train a machine learning predictor to discriminate correct from incorrect peptide identification. Percolator uses additional spectral information (features) such as charge, length, and other enzymatic descriptors of the peptide[19] to compute a final support vector machine (SVM) score. Similarly, the crosslink search engine Kojak[20] supports the use of PeptideProphet[21,22] and XlinkX[23] supports Percolator[18], while pLink2[24] and ProteinProspector[4] have a built-in SVM classifier to re-rank CSMs. Although RT data are readily available, none of these tools use the, often multi-dimensional, RT information for improved identification in crosslinking studies. A prerequisite for this would be that retention times could be predicted reliably.

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

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

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

Results and discussion

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

A substantial fraction of crosslinks below the confidence threshold are correct. Crosslinked peptides belonging to the high-molecular-weight *E. coli* proteome were deep-fractionated along three chromatographic dimensions (hSAX, SCX, and RP). This 3D fractionation approach led to 144 LC–MS runs as some of the 90 fractions contained enough material for repeated analysis. The resulting data were searched with an entrapment database approach (Fig. 1a) leading to 11,196 CSMs (11072 TT, 87 TD, 37 DD, Supplementary Fig. 3) at 1% CSM-FDR, separating self and heteromeric CSMs[16,39,40]. The human entrapment database allows to assess error, independently of the target-decoy approach. This will play a critical role here as *E. coli* decoys will be used for the machine learning-based rescoring (but not for the RT prediction). Judged by a set of peptide characteristic metrics (e.g., peptide length, pI, GRAVY) the human entrapment database resembles the properties of the *E. coli* target database (Supplementary Fig. 4).

Before attempting RT prediction and subsequent complementation of search scores, we investigated the extent of false negatives, approximated here by PPIs present in STRING[41] or APID[42] database. At 1% CSM-FDR, 110 such "validated" (val) protein–protein interactions were identified. 10%, 30%, and 50% CSM-FDR returned 226, 278, and 418 validated PPIs, respectively (Fig. 1b). When raising the CSM-FDR from 1% to 50% we thus saw a nearly 4-fold increase in the detectable number of validated PPIs. In contrast, using a pessimistic approach of semi-randomly

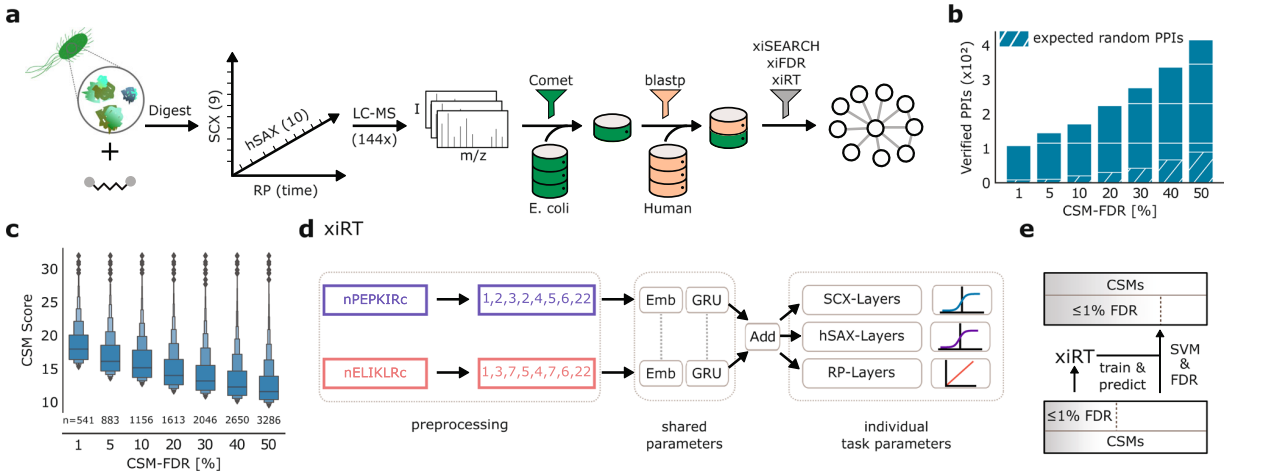


Fig. 1 Workflow overview. a Experimental and data analysis workflow. The soluble high-molecular-weight proteome of *E. coli* lysate was crosslinked and the digest sequentially fractionated by strong-cation exchange chromatography (SCX) (9 fractions collected), hydrophilic strong-anion exchange chromatography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

drawing pairs of *E. coli* proteins from the STRING/APID (first protein) and the search database (second protein) yielded purely by chance 10, 22, 44, and 91 overlapping PPIs with STRING or APID for 1%, 10%, 30%, and 50% CSM-FDR cutoffs, respectively. While this shows that loosening the FDR threshold increases validated PPIs also by chance, the actual observed number is much higher (418 versus 91 at 50% CSM-FDR). This means that there is a substantial number of valid PPIs with insufficient match confidence.

The underlying scoring challenge is essential to the identification of peptides in general. The plethora of search engines for linear[43] and crosslinked peptides[44] use spectral characteristics differently for their scoring. In xiSEARCH, the final score is a composite that incorporates spectral metrics such as explained intensity and matched number of fragments. Empirically, we observe a fast decrease in the search engine score (Fig. 1c) with increasing FDR. This indicates that at higher FDRs spectral matching metrics might be suboptimal. Poor spectral quality, inefficient peptide fragmentation, or random fragment matching all influence the search engine score negatively. RT information could complement MS information but this would require accurate RT prediction of cross-linked peptides.

Accurate multi-dimensional retention time prediction for crosslinked peptides. RT prediction for crosslinked peptides has not yet been achieved. One reason for this is the challenge of encoding a crosslinked pair of peptides for machine learning. We overcame this here using a Siamese neural network as part of a new machine learning application, xiRT (Fig. 1d), which allowed the incorporation of RTs into a rescoring workflow (Fig. 1e). The Siamese part of the network (embedding layer and recurrent layer) shares the same weights for both peptides. Practically, the sharing of weights leads to consistent predictions, independent of the peptide order. After the recurrent layer, the two outputs are combined and passed to three subnetworks consisting of dense layers with individual prediction layers (details on the architecture are available in Supplementary Fig. 1). In this multi-task learning setup, the network simultaneously learns to predict the hSAX, SCX and RP RT through a single training step. Multi-task learning can improve the overall performance of predictors by forcing the network to learn a robust representation of the input data[45].

The training and evaluation of xiRT followed a cross-validation (CV) strategy that avoided the simultaneous learning and prediction on overlapping parts of the data (see "Methods" section, Fig. 2a). We used a 3-fold CV strategy where two folds were used for training (excluding 10% for the validation throughout the training epochs) and one fold for testing/prediction. All CSMs with an FDR < 1% were used during the CV. For the remaining CSMs, the best predictor (with the lowest total loss) was used to predict the RTs.

To achieve the best possible prediction performance, hyper-parameters of the network were optimized. Since extensive hyper-parameter optimization on a small data set can lead to overfitting, we initially optimized a large part of hyper-parameters using 20,802 unique linear peptide identifications at 1% FDR. The final parameters for the Siamese network architecture for crosslinks were obtained by a small grid-search (6453 unique peptide-pairs at 1% CSM-FDR; Supplementary Fig. 5).

Using these parameters, we evaluated the learning behavior during the training time (epochs) across the CV folds. The training behavior on the three CV folds was similar and reached a stable trajectory after approximately 15 epochs (Fig. 2b). Based on very similar error trends on validation and training sets, we

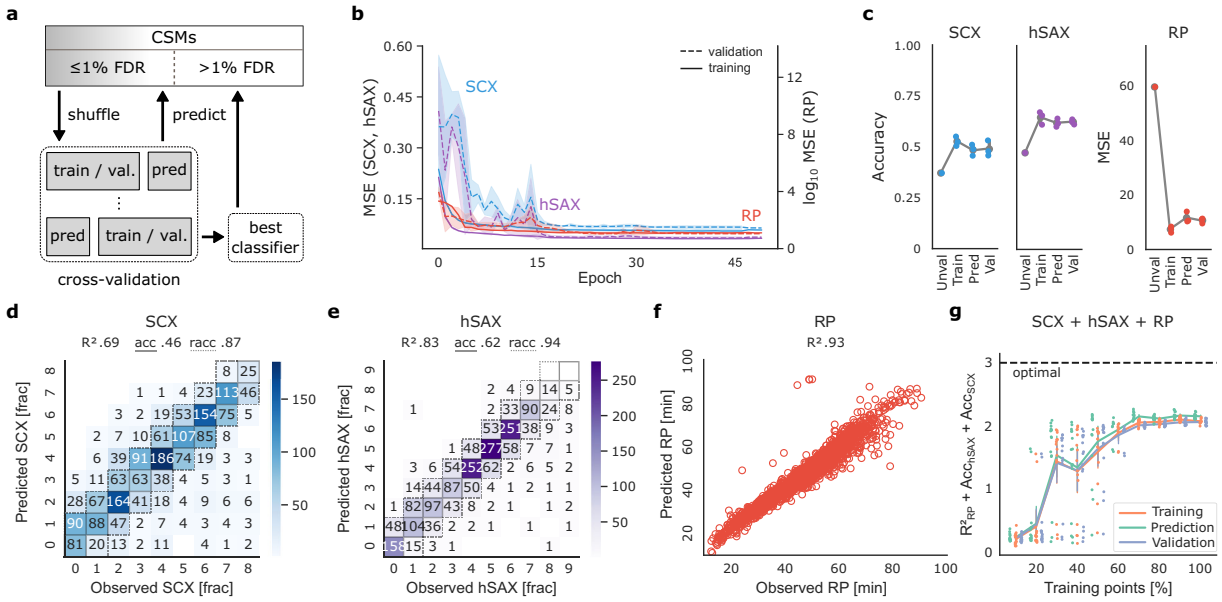


Fig. 2 Cross-validation of retention time prediction. a Applied cross-validation (CV) strategy in xiRT. To predict the retention times of CSMs excluded from training, the best CV classifier is used. **b** xiRT performance over training epochs for strong-cation exchange chromatography (SCX, blue), hydrophilic strong-anion exchange chromatography (hSAX, purple), and reversed-phase chromatography (RP, red) prediction with $k = 3$ CV-folds. Shaded areas show the estimated 95% confidence interval with the dashed/solid line representing the mean for the validation/training data, respectively. **c** xiRT performance across different metrics (error bars show standard deviation with the mean as center) for $k = 3$ CV folds. Prediction for the "unvalidated" data was only performed once. **d**–**f** Prediction results from a representative CV iteration for SCX, hSAX, and RP at 1% CSM-FDR. The achieved R^2, accuracy (acc) and relaxed accuracy (racc) are given at the top. **g** Learning curve with increasing number of CSMs, e.g., 10% (645 total CSMs, 387 for training, 43 for validation, 215 for prediction), 50% (3226, 1935, 216, 1075), 100% (6453, 3871, 431, 2151); bars indicate standard deviation with the line representing the mean for the training (red), prediction (green), and validation (blue) data. Source data are provided as a Source Data file.

concluded to have reached a state where neither overfitting nor underfitting occurred. The overall performance across the prediction folds was comparable in terms of accuracy (hSAX: 61% ± 1.1, SCX: 47% ± 1.7) and MSE (RP: 11.58 ± 2.0) (Fig. 2c). Comparing single-task and multi-task configurations of xiRT revealed no significant differences in the prediction accuracy but greatly reduced run times (Supplementary Figs. 6 and 7). Note that we estimated the theoretical boundaries given the ambiguous elution behavior (i.e., peptide elution across multiple chromatographic fractions) for SCX at 65% accuracy and for hSAX at 73% accuracy (Supplementary Table 4 and Supplementary Fig. 8). Most of the predictions showed only a small error, and thus a high relaxed accuracy: for hSAX 94% ± 0.0 and for SCX 87% ± 1.15 of the predictions were within a range of ± 1 fraction (Fig. 2d, e). The overall R^2_{RP} of 0.94 ± 0.01 also showed a predictable relationship for the RP dimension (Fig. 2f). The consistent accuracy and R^2 results across CV folds demonstrate reproducible training and prediction behavior which reduces unwanted biases from the different CV folds. In conclusion, RTs of crosslinked peptides can robustly be learned within a data set, making them available as features in a CSM rescoring framework.

It was difficult to compare our RT predictions to other studies which used SCX[46] or hSAX[29] for multiple reasons: (1) there is currently no other model that predicts the RT of crosslinked peptides, (2) the recent SSRCalc[46] study (SCX) for linear peptides used a much larger data set of 34,454 unique peptides and the fractionation was much more fine-grained (30–50 fractions). Similarly, the hSAX[29] study on linear peptides used a much finer fractionation (30 fractions) and a different methodology to encode the loss function during the machine learning. (3) Applied gradients and liquid chromatography conditions can change the elution behavior quite drastically. In our study, the number of observations was neither for hSAX nor for SCX equally distributed but varied between ~200 and ~2000 CSMs per fraction (Supplementary Fig. 3). Since we employed a partially exponential gradient during the chromatographic fractionation, the degree of peptide separation varied for earlier and later fractions.

Given that we had less data to train on than recent RT predictions of linear peptides, we evaluated how the numbers of observations influenced the prediction accuracy ($R^2_{RP} + Acc_{hsax} + Acc_{scx}$, Fig. 2g). The learning curve showed two important characteristics: first, the prediction performance over CV folds was very reproducible. This means that predictions were robust even with very moderate data quantity. Second, the maximal performance was achieved with ~70–100% of the data points (100% corresponding to 6453 total CSMs, 3871 for training, 431 for validation, 2151 for prediction). Given that a first plateau was reached with 30% of the data, it is unclear if the final prediction accuracy constitutes another local optimum or the limit of the prediction accuracy. The individual task metrics showed that the RP behavior seemed to be easier for the model to learn than the ordinal regression tasks (SCX, hSAX, Supplementary Fig. 9). The RP behavior could be accurately predicted from ~60% of the data points, while the maximum accuracy for hSAX and SCX dimensions was only achieved by using 80–100% of the data. In other words, while using even fewer CSMs might be possible when predicting RP RTs, one would expect a reduced accuracy in the hSAX/SCX dimensions.

An approach to reduce the number of required CSMs would be to leverage the abundantly available data on linear peptides for transfer learning. Indeed, a recent study showed that transfer learning across different peptide identification results works well for linear peptides[34]. We also implemented the option to

pre-train on linear data in xiRT. However, a robust and accurate RT prediction could be achieved on a multiprotein complex crosslinking study (FA-complex, see below) when first training on the *E. coli* CSMs (Supplementary Fig. 10). Another possibility to increase the training data size and robustness during CV is to increase the number of folds, e.g., 5- or 10-fold, at the cost of runtime. Increasing the expedience of xiRT, we also implemented transfer learning for cases when the number of fractions differs between the initial model and the new prediction task.

Explainable deep learning reveals amino acid contributions. Using the SHAP package, we set out to explain predictions made by xiRT. For instance, when a specific crosslinked peptide was analyzed, residue-specific contributions towards the predicted RT could be computed (Supplementary Fig. 11). The residues D, E, Y, and F displayed high SHAP values indicating a stronger retention during hSAX separation in a randomly chosen peptide. Looking at a specific crosslinked peptide in SCX (Supplementary Fig. 12), the SHAP values highlighted that K and R were the most important residues contributing towards later peptide elution. As one might expect, crosslinked K residues contributed much less towards later elution times than the stronger charged, unmodified K residues. Investigating the SHAP values for a collection of CSMs revealed additional contributions from W for hSAX and H for SCX while returning hydrophobic residues Y, F, W, I, L, V, and M for RP (Supplementary Fig. 13), revealing residue contributions in crosslinked peptides as seen in the respective analyses of linear peptides[29,46,47]. In summary, the SHAP values were good estimates for the individual RT contributions of the amino acid residues.

Next, we investigated the network architecture and the learned feature representations more closely (Supplementary Note 4). As first analysis, the dimensionality reduced embedding space across the network was analyzed (Supplementary Fig. 14). This revealed that the shared sequence-specific layer already captured the RP properties quite well, while the hSAX and SCX properties were not as clearly captured. As expected, the separation of CSMs according to RT increased the further the features propagated through the network. In the last layer, the RP and hSAX subnetworks reached a very good separation, while in the SCX subtask CSMs remained moderately separated in two dimensions.

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

To test the transferability of our findings, we also ran xiRT with unfiltered pLink2 results (Supplementary Note 4 and Supplementary Fig. 15). The prediction performance from *Q*-value-filtered CSMs was similar to the results with xiSEARCH (Supplementary Fig. 15a–c). A two-sided *t*-test between hSAX, SCX, and RP errors for TT and TDs revealed significant differences in the respective error distributions using pLink2 identifications for the RT predictions (Supplementary Fig. 15d). Importantly, the separation of true and false matches in two-dimensional space was also possible with pLink2 identifications (Supplementary Fig. 15e). In summary, xiRT can learn retention times irrespective of the used search engine and the learned chromatographic features alone carry substantial information to separate true from false matches.

To investigate the relevance of multi-dimensional RT predictions for the identification of cross-linked peptides, we first supplemented each CSM with RT features. Then, we performed a semi-supervised rescoring and evaluated the trained SVM model using the SHAP framework. We chose to analyze SHAP values for the 15 most important retention times features for TT observations (FDR > 1%) that were predicted to be a correct TT identification (Fig. 3c). This analysis revealed a similar magnitude for all 15 SHAP values implying that a single feature alone is insufficient to recognize false matches. Notably, the top 5 features contained features from RP, hSAX, and SCX predictions which indicates that each chromatographic dimension carried relevant information for the rescoring. Because 11 of the 15 features were predictions considering only one of the two peptides and not directly derived from peptide-pairs, the predicted RTs displayed a larger error. This analysis suggests that an RT prediction model for linear peptides can add valuable information for crosslink analyses. In general, the model learned mostly that low errors in the RT dimensions indicate true positive identifications. Thus, the model implicitly learned that the RT of a crosslinked peptide should differ from the RT of the individual peptides. This might become useful especially for distinguishing consecutive[48] from crosslinked peptides or when dealing with gas-phase associated peptides[36].

Rescoring crosslinked peptides enhances their identification. Before computing a combined score, we compared the CSM scores based on mass spectrometric information (xiSCORE) and RT features (SVM score, Fig. 4a). Both scores largely agreed. Heteromeric CSMs passing 1% CSM-FDR yielded high SVM scores. Also, most target-decoy CSMs achieved a low SVM score (Fig. 4a, right) and a low xiSCORE (Fig. 4a, top). The SVM score distribution of the TDs matched closely the distribution of TTs in the low scoring area, which indicated that they still modeled random TT matches and that overfitting was avoided. Interestingly, the TTs were overrepresented in the low scoring area for the xiSCORE but not for the SVM score, suggesting that true TTs remained hidden among the random matches when using xiSCORE alone. The broad SVM score distribution of TTs indicated that the rescoring process could be optimized. In conclusion, neither of the mass spectrometric information (xiSCORE) nor the RT information (SVM score) seem to reveal all true CSMs.

As a combination of both approaches should yield better results than either alone, we combined the SVM score with the xiSCORE. We evaluated the impact of rescoring CSMs on the number and quality of identified PPIs, as PPIs are typically the objective of large-scale cross-linking MS experiments.

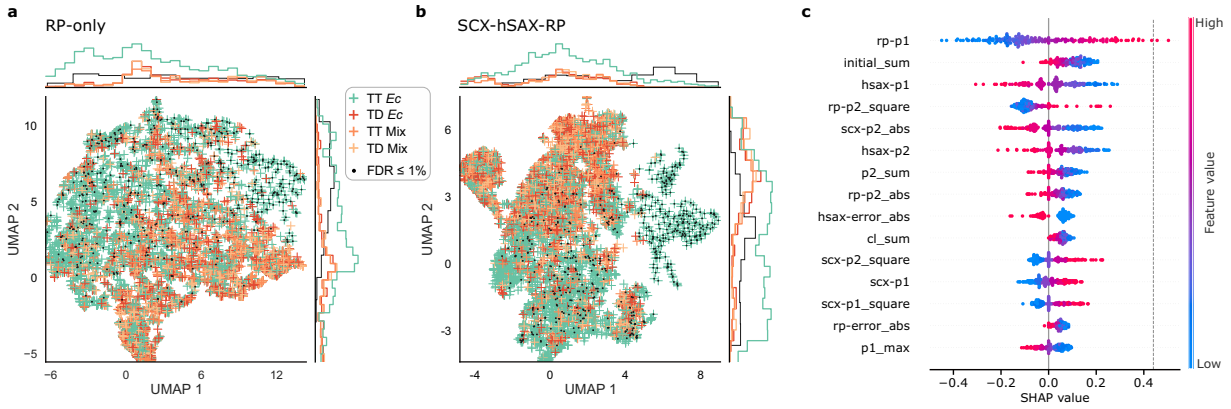


Fig. 3 Visualization of RT features. a xiRT-based features from reversed-phase chromatography (RP) dimension only (13 features) after dimensionality reduction with Uniform Manifold Approximation and Projection (UMAP). **b** xiRT-based feature from SCX-hSAX-RP (strong-cation exchange – hydrophilic strong-anion exchange – reversed-phase chromatography) dimensions (43 features) after dimensionality reduction with UMAP. Input data for **a** and **b** were CSMs of heteromeric links in the proteome-wide crosslinking dataset (*Ec* = *E. coli*; green = TT; red = TD, Mix = match between *E. coli* and human peptides; orange = TT; peach =TD), filtered to 50% CSM-FDR. Identifications passing 1% CSM-FDR are highlighted. Decoy-decoy identifications are not shown. **c** SHAP analysis of RT feature importance for CSM-rescoring (using a linear SVM) including SCX, hSAX and RP features (Supplementary Table 4). Each dot represents a previously identified CSM from 200 randomly chosen TTs that were excluded from training (i.e., CSM-FDR > 1%). The background data set consists of 100 TT and TD CSMs each. Dashed line indicates the base value for a prediction based on the background data alone (0.44). Source data are provided as a Source Data file.

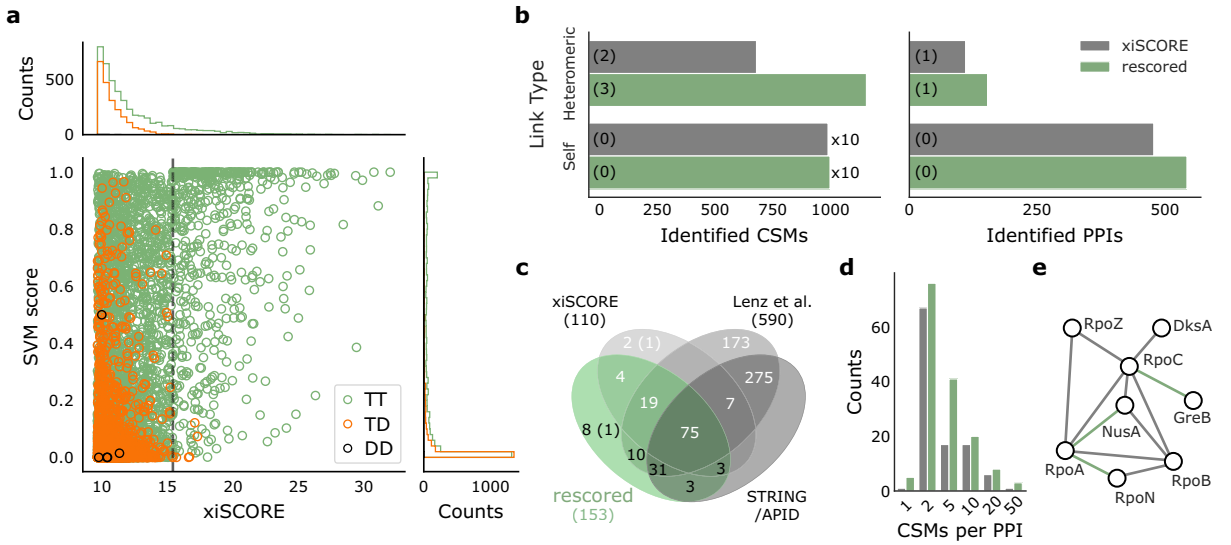


Fig. 4 Incorporation of RT prediction to CSM-scoring increases crosslink identification. a Score distributions of heteromeric CSMs based on mass spectrometric information (xiSCORE) and retention time features (SVM score). The dashed line indicates the xiSCORE-based CSM-FDR threshold of 1%. **b** Increase in the identification of TT-CSMs and PPIs at constant FDR. Numbers in brackets indicate identifications involving a human protein. **c** Overlap of observed PPIs (at 1% heteromeric PPI-FDR) to external references. Numbers in the Venn diagram represent the identified PPIs among *E. coli* proteins or PPIs involving human proteins (in brackets). Black numbers highlight the added benefit from combining xiSCORE with xiRT's SVM score for PPI identification. **d** Distribution of CSMs per PPI before (gray) and after CSM-rescoring (green). **e** Selected subnetwork of the RNA polymerase with PPIs only identified after the rescoring connected in green. Data in **b**–**e** correspond to a 1% PPI-FDR (prefiltered at a 5% CSM-FDR). Source data are provided as a Source Data file.

Heteromeric CSMs increased 1.7-fold and heteromeric PPIs increased 1.4-fold (Fig. 4b). Self-links increased only marginally in agreement with their smaller search space and accordingly lower random match frequency. Essentially, nearly all self-links were identified exhaustively based on mass spectrometric data alone. In contrast, RT information substantially improved the identification of heteromeric CSMs. Further gains might be possible by directly combining RT features with mass spectrometric features (and possibly also other) for supervised scoring.

Likely, the benefits of RT predictions for the rescoring depend on the data set and applied chromatographic separations. On the *E. coli* data, we, therefore, performed additional analyses where we limited the rescoring to only use a subset of the chromatographic dimensions (Supplementary Table 5). The number of identified CSMs for heteromeric links increased from 724 in the reference to 902 (RP only), 977 (SCX-RP), 1092 (hSAX-RP), and 1199 (SCX-hSAX-RP). Likewise, PPIs increased from 109 to 135, 131, 157, 152, respectively (Supplementary Table 5). As observed

above, gains can be expected from each chromatographic dimension. When having to choose one ion chromatography, the hSAX dimension seemed more useful than the SCX dimension which could arise from the better prediction performance or more complex separation mechanisms. Importantly, even using RP RT alone already led to a marked gain in heteromeric PPIs (see also next section).

To systematically evaluate the additionally identified PPIs from all three RT dimensions, we compared them to the originally identified PPIs based exclusively on xiSCORE. In addition, the STRING/APID databases and a set of PPIs from a larger study[16] served as extra references for validation. Almost all PPIs found in the original dataset by xiSCORE were also contained in the rescored data set (91%). 85% of the newly identified PPIs were either found in the data set from Lenz et al., in STRING/APID or both. Among the eight PPIs unique to the rescored data set, only one involved a human protein from the entrapment database (Fig. 4c), which we could manually resolve and match to *E. coli* (Supplementary Table 6). The remaining seven PPIs might constitute genuine PPIs. Note that the overall percentage of PPIs involving human proteins was reduced by rescoring. Since all human target proteins were included in the positive training data, this is an important indicator of a well-behaved model. Deepening trust further, almost all novel PPIs were identified with multiple CSMs (Fig. 4d). Finally, we selected the subnetwork of the RNA polymerase to investigate the additionally identified PPIs in a well-characterized interaction landscape (Fig. 4e). Indeed, all interactions added by RT-based rescoring were already reported in APID. In summary, all our evidence points at the successful complementation of MS information by RT, at least for a proteome-wide crosslinking analysis. It remained to be seen, however, if this could also be leveraged in more routine multiprotein complex analyses.

Multiprotein complex studies also benefit from the RT prediction. Many cross-linking MS studies investigate multiprotein complexes and rely on only a few chromatographic dimensions. We, therefore, evaluated the benefit of predicted RTs for the analysis of the FA-complex, an eight-membered multiprotein complex that was crosslinked using BS3. Here, the search engine score was supplemented exclusively with RP RT predictions during the rescoring. By using transfer learning, the small number of CSMs (692 unique CSMs, without considering charge states) found in this multiprotein complex analysis were sufficient to achieve accurate RP predictions (Supplementary Fig. 10). The resulting crosslinks at 1% residue-pair FDR (lower levels set to 5%) showed an increase of 36 (+10%) self- and 53 (+70%) heteromeric residue-pairs. Importantly, the rescored links showed no indication of increased hits to the entrapment database (Fig. 5a) indicating that no overfitting occurred during the rescoring. At the same time, heteromeric PPIs already identified before rescoring received additional support. For example, the number and sequence coverage of links increased between FAAP100 (100) and FANCB (B), FANCA (A) and FANCB, and FANCA and FANCG (G). Overall, the heteromeric links increased 1.7-fold with an even higher proportional increase in "verified" links, i.e., fitting the available structure, by 1.9-fold (Fig. 5b). The derived distance distribution of newly identified links is dissimilar from a random distribution and shows no indications of reduced quality (Fig. 5c). Applying this "structural validation" on its own might be optimistic[49], however, in summary, our rigorous quality control ensures trustworthy results. It is currently unclear how far even smaller data sets could benefit from xiRT. Generally, to improve prediction performance, pretraining on larger data sets will lead to better generalization abilities of the predictor. Subsequently, also smaller data sets can be used for accurate RT prediction. To additionally benefit from sample-specific information, increasing the cross-validation splits will utilize larger parts of the data during training. In any case, our successful analysis of a multiprotein complex supplemented with only RP features highlights the broad applicability of xiRT.

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

Methods

Sample preparation and multidimensional fractionation. Biomass was produced from a single clone of *Escherichia coli* K12 strain (BW25113 purchased from DSMZ, Germany; https://www.dsmz.de/) by fermentation in a Biostat A plus bioreactor (Sartorius, Göttingen, Germany) in LB medium with 0.5% (w/v) glucose at 37 °C while monitoring and adjusting pH and dissolved oxygen by the addition of sodium hydroxide/phosphoric acid or stir speed control, respectively. When the culture grew to an optical density$_{600}$ of 10 it was harvested by centrifugation at 5000×*g*, 4 °C for 15 min, then washed with 1× PBS, aliquoted, snap-frozen in liquid nitrogen, and stored at −80 °C. Cell pellets were resuspended in 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)) and lysed by sonication. Prior to sonication, cOmplete EDTA-free protease-inhibitors (Roche, Basel, Switzerland) were added according to the manufacturer's instructions. Then, Benzonase (Merck, Darmstadt, Germany) was added and the lysate cleared from cellular debris by centrifugation for 15 min at 4 °C and 15.000×*g*. Fresh DTT was supplied to 2 mM. The obtained supernatant was treated further by ultracentrifugation using a 70 Ti fixed-angle rotor for 1 h at 106,000×*g* and 4 °C. Subsequently, the protein solution was concentrated using Amicon spin filters (15 kDa molecular weight cut-off; Merck, Darmstadt, Germany) to reach a total protein concentration of 10 mg/ml, as judged by microBCA assay (ThermoFisher Scientific, Waltham, MA, USA) and aggregates removed by centrifugation for 5 min at 16,900×*g* and 4 °C. Then, 2 mg of this soluble high molecular weight proteome was separated on a BioSep SEC-S4000 column (600 × 7.8 mm, pore size 500 Å, particle size 5 µm, Phenomenex, CA, USA) at 200 µl/min flow rate and 4 °C with fraction collection of 200 µl over the separation range from ~3 MDa to 150 kDa (as judged by Gel filtration calibration kit (HMW), GE Healthcare) to give 44 fractions. The proteins of each fraction were crosslinked using 0.75 mM disuccinimidyl suberate (DSS; Sigma, St. Louis, MO, USA). The cross-linked samples were pooled and precipitated using acetone. Upon resuspending in 6 M urea, 2 M thiourea, 100 mM ammonium bicarbonate (ABC), the samples were derivatized by incubating 30 minutes at room temperature with 10 mM dithiothreitol followed by 20 mM iodoacetamide in the dark. Proteolysis was accomplished using LysC protease (1:100 protease-to-substrate mass ratio; Pierce Biotechnology, Rockford, IL, USA) for 4.5 h at 37 °C, followed by 1:5 dilution with 100 mM ABC and additional digestion with and Trypsin (1:25 protease-to-substrate mass ratio; Pierce Biotechnology, Rockford, IL, USA). Digestions were quenched by adding trifluoroacetic acid (TFA) and cleaned up using Stage-tips. The sample was fractionated in the first dimension on a PolySulfoethyl A strong cation exchange chromatography (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 on an Äkta pure system (GE Healthcare, Chicago, IL, USA) at 21 °C. Mobile phase A was 10 mM monopotassium phosphate pH 3.0, 30% acetonitrile; mobile phase B additionally contained 1 M potassium chloride (KCl). About 0.4 mg peptides dissolved in mobile phase A were loaded and eluted isocratically over 2 min, followed by an exponential gradient up to 700 mM KCl with the 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. We collected nine high-salt fractions of 0.2 ml size during several replica SCX runs. Identical fractions were pooled and desalted using Stage-tips followed by separation in the

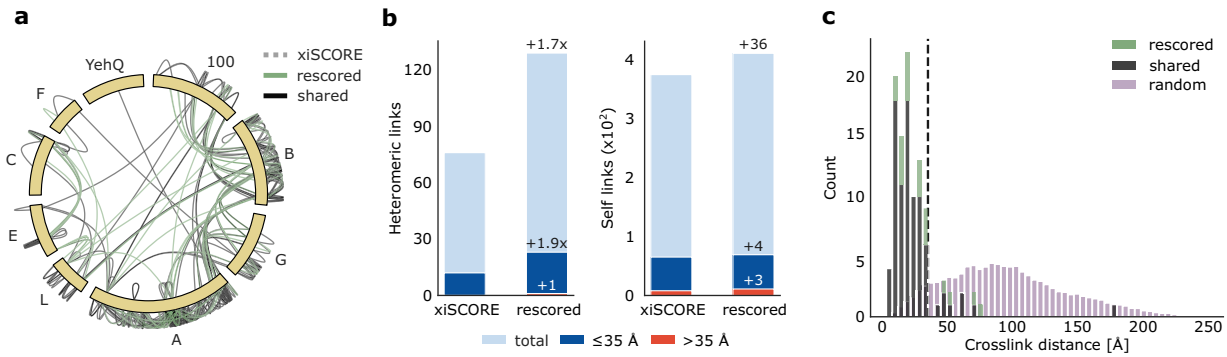


Fig. 5 Benefit of RT prediction for multiprotein complex crosslink analysis. a Crosslink network from the Fanconi anemia complex analysis, shown in the circular view. Unique residue pairs from xiSCORE (gray), after rescoring (green), and shared (black) between these analyses are depicted (1% residue-pair FDR). Proteins associated to the Fanconi anemia core complex are indicated with their gene name suffix. The *E. coli* protein YehQ represents a match from the entrapment database. **b** Quantitative assessment of residue-pairs with and without rescoring, and including calculated distances in the model (all, light blue; ≤35 Å, blue; >35 Å red). **c** Distribution of crosslink distances from identified residue-pairs ($n = 105$) following rescoring (green), shared (black) between rescoring and xiSCORE (since no crosslinks unique to xiSCORE), and theoretically possible residue-pairs (random) that could be mapped to the model. Source data are provided as a Source Data file.

second chromatographic dimension by hydrophilic strong anion exchange chromatography (hSAX). Here, we used a Dionex IonPac AS-24 hSAX column (250 × 2.0 mm) with an AG-24 guard column (Thermo Fisher Scientific, Dreieich, Germany) running at 0.15 ml/min on an Äkta pure system (see above) and at 15 °C. Mobile phases A and B were 20 mM Tris*HCl pH 8.0 with B additionally containing 1 M sodium chloride. Samples were loaded in mobile phase A and separated under isocratic conditions for 3 min, followed by elution using an exponential gradient: 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, each step lasting for one minute. Fractions of 0.15 ml size were collected along the gradient. Ten pools were prepared (fractions 3-6/7-14/15-17/18-19/20-21/22-23/24-25/26-27/29-29/30-35) and desalted using Stage-tips.

LC-MS for crosslink identification. Analysis of crosslinked peptides by LC-MS was conducted on a Q Exactive HF mass spectrometer (ThermoFisher 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. Solvents A and B were 0.1% (v/v) formic acid and 80% (v/v) acetonitrile, 0.1% (v/v) formic acid, respectively. Peptide fractions were dissolved and loaded in 1.6% acetonitrile, 0.1% formic acid onto an Easy-Spray column (C18, 50 cm, 75 µm ID, 2 µm particle size, 100 Å pore size) operated at 300 nl/min flow and 45 °C. Peptide elution used the following gradient: 2 to 7.5% buffer B within 5 min, from 7.5 to 42.5% over 80 min, to 50% B over 2.5 min, and then to 95% buffer B within 2.5 min and flushed for another 5 min before re-equilibration at 2% B. Survey scans were acquired at a resolution of 120,000, automated gain control of $3*10^6$, maximum injection time of 50 ms while scanning from 400–1450 m/z in profile mode. The top 10 intense precursor ions with $z = 3$-6 and passing the peptide match filter (preferred) were isolated using a 1.4 m/z window and fragmented by higher-energy collisional dissociation using stepped normalized collision energies of 24, 30, and 36. Fragment ion scans were recorded at a resolution of 60,000, with automated gain control set to $5*10^4$, maximum injection time of 120 ms, underfill ratio of 1%, and scanning from 200–2000 m/z. Dynamic exclusion for previously fragmented precursors and their isotopes was enabled for 30 s. To minimize the non-covalent gas-phase association of peptides, in-source-CID was enabled at 15 eV[36]. Each LC-MS run lasted for 120 min.

Spectra and peptide spectrum match processing. All raw spectra were converted to Mascot generic format (MGF) using msConvert[50] (3.0.20175.cbf82d022). The database search with Comet[51] (v. 2019010) was done with the following settings: peptide mass tolerance 3 ppm; isotope_error 3; fragment bin 0.02; fragment offset 0.0; decoy_search 1; fixed modification on C (carbamidomethylation, +57.021 Da); variable modifications on M (oxidation, +15.99 Da). False discovery rate (FDR) estimation was performed for each acquisition. First, the highest-scoring PSM for a modified peptide sequence was selected, then the FDR was computed based on Comet's *e*-value. Spectra were searched using xiSEARCH (v. 1.6.753)[12], after recalibration of precursor and fragment m/z values, with the following settings: precursor tolerance, 3 ppm; fragment tolerance, 5 ppm; missed cleavages, 2; missed monoisotopic peaks[52], 2; minimum peptide length, 7; variable modifications: oxidation on M, mono-links for linear peptides on K, S, T, Y, fixed modifications: carbamidomethylated C. The specificity of the crosslinker DSS was configured to link K, S, T, Y, and the protein N terminus with a mass of 138.06807 Da. The searches were run with the workflow system snakemake[53]. The FDR on CSM-level was defined as FDR = TD − DD/TT[40], where TD indicates the number of target-decoy matches, DD the number of decoy–decoy matches, and TT the number of target-target matches. Crosslinked peptide spectrum matches (CSMs) with non-consecutive peptide sequences were kept for processing[48]. PPI level FDR computation was done using xiFDR[40] (v. 2.1.3 and 2.1.5 for writing mzIdentML) to an estimated PPI-FDR of 1%, disabling the boosting and filtering options. CSM, peptide, and residue-level FDR were fixed at 5%, protein group FDR was set to 100%. FDR estimations for self and heteromeric links were done separately. In xiFDR a unique CSM is defined as a combination of the two peptide sequences including modifications, link sites, and precursor charge state. For the assessment of identified CSMs an entrapment database (described in the next section), as well as decoy identifications, were used on both, CSM and PPI levels. PPI results were also compared against the APID[42] and STRING[41] databases (v11, minimal combined confidence of 0.15).

Database creation. The database of potentially true crosslinks was defined as *Escherichia coli* proteome (reviewed entries from Uniprot release 2019-08). This database was filtered further to proteins identified with at least a single linear peptide at a q-value[54] threshold of 0.01, $q(t) = \min_{s \le t} \mathrm{FDR}(s)$, with the threshold t and score s. This resulted in 2850 proteins. In addition to the FDR estimation through a decoy database, we used an entrapment database. The proteins from the entrapment database represent the search space of false-positive CSMs independent of *E. coli* decoys and were sampled from human proteins (UP000005640, retrieved 2019-05). *E. coli* decoys might fail in this task after machine learning if overfitting should have taken place. So, entrapment targets allow control for overfitting. For this, human target peptides were treated as targets and human decoy peptides as decoys. To avoid complications through false spectrum matches due to homology, we used blastp[55] (BLAST 2.9.0+, blastp-short mode, word size 2, *e*-value cutoff 100) and aligned all *E. coli* tryptic peptides (1 missed cleavage, maximum length 100) to the human reference. All proteins that showed peptide alignments with a sequence identity of 100% were removed from the human database. Only the remaining 9990 sequences were used as candidates in the entrapment database. For each of the 2850 *E. coli* proteins, a human protein was added to the database. To reduce search space biases from protein length and thus different number of peptides for the two organisms, we followed a special sampling strategy. The human proteins were selected by a greedy nearest neighbor approach based on the K/R counts and the sequence length. The final number of proteins in the combined database (*E. coli* and human) was 5700 (2850*2).

Fanconi anemia monoubiquitin ligase complex data processing. The publicly available raw files from an analysis of the BS3-crosslinked Fanconi anemia monoubiquitin ligase complex[56] (FA-Complex) were downloaded from PRIDE together with the original FASTA file (PXD014282). The raw files were processed as described for the *E. coli* data (m/z recalibration and searched with xiSEARCH), followed by an initial 80% CSM-FDR filter for further processing. Due to the much smaller FASTA database (8 proteins), the entrapment database was constructed more conservative than for the proteome-wide *E. coli* experiment, i.e., for each of the target proteins, the amino acid composition was used to retrieve the nearest neighbor in an *E. coli* database. The FDR settings to evaluate the rescoring were set to 5% CSM- and peptide-pair level FDR, 1% residue-pair- and 100% PPI-FDR using xiFDR without boosting or additional filters. The resulting links were visualized (circular view) and mapped to an available 3D structure (final refinement model "sm.pdb")[57,58] using xiVIEW[59]. To ease the comparison of identified and random distances, a random Euclidean distance distribution was derived in three steps: first, all possible cross-linkable residue-pair distances in the 3D

structure were computed. Second, 300 random "bootstrap" samples with n distances were drawn (n = the number of identified residue-pairs at a given FDR) and third, the mean per distance bin was computed across all 300 samples.

xiRT—3D Retention Time Prediction. The machine learning workflow was implemented in python (v. >3.7) and is freely available from https://github.com/Rappsilber-Laboratory/xiRT. xiRT is the successor of DePART[29], which was developed for the retention time (RT) prediction of hSAX fractionated peptides based on pre-computed features. xiRT makes use of modern neural network architectures and does not require feature engineering. We used the popular python packages sklearn[60] (0.24.1) and TensorFlow[61] (v. 1.15 and >2) for processing (Supplementary Note 1 for more details). xiRT consists of five components (Fig. 1d and Supplementary Fig. 1, Supplementary Note 1): (1) The input for xiRT are amino acid sequences with arbitrary modifications in text format (e.g., Mox for oxidized Methionine). xiRT uses a similar architecture for linear and crosslinked peptide RT prediction. Before the sequences can be used as input for the network, the sequences are label encoded by replacing every amino acid by an integer and further 0-padded to guarantee that all input sequences have the same length. Modified amino acids, as well as crosslinked residues, are encoded differently than their unmodified counterparts. (2) The padded sequences were then forwarded into an embedding layer that was trained to find a continuous vector representation for the input. (3) To account for the sequential structure of the input sequences, a recurrent layer was used (either GRU or LSTM). Optionally, the GRU/LSTM layers were followed by batch normalization layers. For cross-linked peptide input, the respective outputs from the recurrent layers were then combined through an additive layer (default setting). (4) Task-wise subnetworks were added for hSAX, SCX, and RP retention time prediction. All three subnetworks had the same architecture: three fully connected layers, with dropout and batch normalization layers between them. The shape of the subnetworks is pyramid-like, i.e., the size of the layers decreased with network depth. (5) Each subnetwork had its own activation function. For the RP prediction, a linear activation function was used and mean squared error (MSE) as loss function. For the prediction of SCX and hSAX fractions, we followed a different approach. The fraction variables were encoded for ordinal regression in neural networks[62]. For example, in a three-fraction setup, the fractions (f) were encoded as $f_1 = [0, 0, 0], f_2 = [1, 0, 0]\ and f_3 = [1, 1, 0]$. Subsequently, we chose sigmoid activation functions for the prediction layers and defined binary cross-entropy (BC) as loss function. To convert predictions from the neural network back to fractions, the index of the first entry with a predicted probability of <0.5 was chosen as the predicted fraction. The overall loss was computed by a weighted sum of the MSE_{RP}, BC_{SCX}, and BC_{hSAX}. The weight parameters are only necessary when xiRT is used to predict multiple RT dimensions at the same time (multi-task). To predict a single dimension (single-task, e.g., RP only), the weight can be set to 1. The number of neurons, dropout rate, intermediate activation functions, the weights for the combined loss, number of epochs, and other parameters in xiRT were optimized on linear peptide identification data. Reasonable default values are provided within the xiRT package. For optimal performance, further optimization might be necessary for a given task.

Cross-validation and prediction strategy. Cross-validation (CV) is a technique to estimate the generalization ability of a machine learning predictor[63] and is often used for hyper-parameter optimization. We performed a 3-fold CV for the hyper-parameter optimization on the linear peptide identification data from xiSEARCH, excluding all identifications to the entrapment database (Supplementary Note 2 and Supplementary Fig. 2 for details). We defined a coarse grid of parameters (Supplementary Table 1) and chose the best performing parameters based on the average total (unweighted) loss, R^2_{RP} and accuracy across the CV folds. Further, we define the relaxed accuracy (racc) to measure how many predictions show a lower prediction error than |1| fraction. We then repeated the process with an adapted set of parameters (Supplementary Table 2). In addition to the standard CV strategy, we used a small adjustment: per default, in k-fold cross-validation, the training split consists of $k - 1$ parts of the data (folds) and a single testing fold. However, we additionally used a fraction (10%) from the training folds as extra validation set during training. The validation set was used to select the best performing classifier over all epochs. The model assessment was strictly limited to the testing folds. This separation into training, validation, and testing was also used for the semi-supervised learning and prediction of RTs, i.e., when xiRT was used to generate features to rescore CSMs previously identified from mass spectrometric information. In this scenario, the CV strategy was employed to avoid the training and prediction on the same set of CSMs. In xiRT, a unique CSM is defined as a combination of the two peptide sequences, ignoring link sites and precursor charge.

Supervised peptide spectrum match rescoring. To assess the benefits of RT predictions, we used a semi-supervised support vector (SVM) machine model. The implementation is based on the python package scikit-learn[60] in which optimal parameters are determined via cross-validation. The input features were based on the initial search score (for FA-complex only) and differences between predicted and observed RTs. For each cross-linked peptide, three predictions were made per chromatographic dimension: for the crosslinked peptide, for the alpha peptide, and the beta peptide. Additional features were engineered depending on the number of chromatographic dimensions and included the summed, absolute, or squared values of the initial features (Supplementary Table 3 for all features). For example, for three RT dimensions, the total number of features was 43. The data for the training included all CSMs that passed the 1% CSM-FDR cutoff (self, heteromeric/TT, TD, DDs) and TD/DD identifications that did not pass this cutoff. TTs were labeled as positive training examples, TD and DDs (DXs) were labeled as negative training examples.

To stratify the k-folds during CV, the CSMs were binned into k xiSCORE percentiles. Afterward, they were sampled such that each score range was equally represented across all CV folds. When the positive class was limited to the TT identifications at 1% CSM-FDR, the number of negative observations was usually larger than the number of positive observations. To circumvent this, for each CV split, a synthetic minority over-sampling technique (SMOTE)[64] was used to generate a balanced number of positive and negative training samples (here only used for the FA-complex data). SMOTE was applied within each CV fold to avoid information leakage. A 3-fold CV was performed for the rescoring. In each iteration during the CV, two folds were used for the training of the classifier, and the third fold was used to compute an SVM score. During this CV step, a total of three classifiers were trained. The scores for all TT-CSMs that did not pass the initial FDR cutoff were computed by averaging the score predictions from the three predictors. For all CSMs passing the initial FDR cutoff, rescoring was performed when the CSM occurred in the test set during the CV. The final score was defined as: $xi_{rescored} = xi_{SCORE} + xi_{SCORE} \times SVM_{score}$, where SVM_{score} was the output from the SVM classifier and xi_{SCORE} the initial search engine score.

Feature analysis. The KernelExplainer from SHAP[65] (Shapley Additive exPlanations, v.0.36.0) was used to analyze the importance of features derived from the SVM classifier. SHAP estimates the importance of a feature by setting its value to "missing" for an observation in the testing set while monitoring the prediction outcome. We used a background distribution of 200 samples (100 TT, 100 TD) from the training data to simulate the "missing" status for a feature. SHAP values were then computed for 200 randomly selected TT (predicted to be TT) that were not used during the SVM training. SHAP values allow to directly estimate the contributions of individual features towards a prediction, i.e., the expected value plus the SHAP values for a single CSM sums to the predicted outcome. For a selected CSM, a positive SHAP value contributes towards a true match prediction. For the interpretability analysis (SHAP) of the learned features in xiRT, the DeepExplainer was used (Supplementary Note 3).

In addition, we performed dimensionality reduction using UMAP[66] on the RT feature space for visualization purposes (excluding the search engine score). UMAP was run with default parameters (n_neighbors = 15, min_dist = 0.1) on the standardized feature values. The list of used features for the multi-task learning setup is available in Supplementary Table 3.

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

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

Data availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the jPOST partner repository[67] with the data set identifier PXD020407 and at https://doi.org/10.6019/PXD020407. Raw data of the FA-Complex are available via the previously published PRIDE identifier PXD014282. Additional files and intermediate results are available via Zenodo at https://doi.org/10.5281/zenodo.4270323. PPI data were retrieved from STRING (https://string-db.org/, v11) and APID (http://cicblade.dep.usal.es:8080/APID/init.action, downloaded 09/2019). Source data are provided with this paper.

Code availability

The developed python package is available on the python package index, on GitHub (https://github.com/Rappsilber-Laboratory/xiRT) and via Zenodo (https://doi.org/10.5281/zenodo.4270323).

Received: 20 July 2020; Accepted: 26 April 2021;
Published online: 28 May 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**, 1 (2018).
2. Yu, C. & Huang, L. Cross-linking mass spectrometry: an emerging technology for interactomics and structural biology. *Anal. Chem.* **90**, 144–165 (2018).

3. Leitner, A., Faini, M., Stengel, F. & Aebersold, R. Crosslinking and mass spectrometry: an integrated technology to understand the structure and function of molecular machines. *Trends Biochem. Sci.* https://doi.org/10.1016/j.tibs.2015.10.008 (2016).
4. Trnka, M. J., Baker, P. R., Robinson, P. J. J., Burlingame, A. L. & Chalkley, R. J. Matching cross-linked peptide spectra: only as good as the worse identification. *Mol. Cell. Proteom.* **13**, 420–434 (2014).
5. Giese, S. H., Fischer, L. & Rappsilber, J. A study into the collision-induced dissociation (CID) behavior of cross-linked peptides. *Mol. Cell. Proteom.* **15**, 1094–1104 (2016).
6. Barysz, H. M. & Malmström, J. Development of large-scale cross-linking mass spectrometry. *Mol. Cell. Proteomics* https://doi.org/10.1074/mcp.R116.061663 (2018).
7. Rinner, O. et al. Identification of cross-linked peptides from large sequence databases. *Nat. Methods* **5**, 315–318 (2008).
8. Chen, Z. A. et al. Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry. *EMBO J.* **29**, 717–726 (2010).
9. 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).
10. Schweppe, D. K. et al. Mitochondrial protein interactome elucidated by chemical cross-linking mass spectrometry. *Proc. Natl Acad. Sci. USA* **114**, 1732–1737 (2017).
11. 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).
12. Mendes, M. L. et al. An integrated workflow for crosslinking mass spectrometry. *Mol. Syst. Biol.* **15**, e8994 (2019).
13. 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).
14. Ryl, P. S. J. et al. In situ structural restraints from cross-linking mass spectrometry in human mitochondria. *J. Proteome Res.* **19**, 327–336 (2020).
15. O'Reilly, F. J. et al. In-cell architecture of an actively transcribing-translating expressome. *Science* **369**, 554–557 (2020).
16. Lenz, S. et al. Reliable identification of protein-protein interactions by crosslinking mass spectrometry. *Nat. Commun.* https://doi.org/10.1038/s41467-021-23666-z (2021).
17. Gonzalez-Lozano, M. A. et al. Stitching the synapse: Cross-linking mass spectrometry into resolving synaptic protein interactions. *Sci. Adv.* **6**, eaax5783 (2020).
18. The, M., MacCoss, M. J., Noble, W. S. & Käll, L. Fast and accurate protein false discovery rates on large-scale proteomics data sets with percolator 3.0. *J. Am. Soc. Mass Spectrom.* **27**, 1719–1727 (2016).
19. Granholm, V., Noble, W. S. & Käll, L. A cross-validation scheme for machine learning algorithms in shotgun proteomics. *BMC Bioinformatics* **13**, S3 (2012).
20. Hoopmann, M. R. et al. Kojak: efficient analysis of chemically cross-linked protein complexes. *J. Proteome Res.* **14**, 2190–2198 (2015).
21. Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74**, 5383–5392 (2002).
22. Ma, K., Vitek, O. & Nesvizhskii, A. I. A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet. *BMC Bioinformatics* **13**, S1 (2012).
23. Liu, F., Lössl, P., Scheltema, R., Viner, R. & Heck, A. J. R. Optimized fragmentation schemes and data analysis strategies for proteome-wide cross-link identification. Nat. Commun. https://doi.org/10.1038/ncomms15473 (2017).
24. Chen, Z.-L. et al. A high-speed search engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides. *Nat. Commun.* **10**, 3404 (2019).
25. Klammer, A. A., Yi, X., MacCoss, M. J. & Noble, W. S. Improving tandem mass spectrum identification using peptide retention time prediction across diverse chromatography conditions. *Anal. Chem.* **79**, 6111–6118 (2007).
26. Dwivedi, R. C. et al. Practical implementation of 2D HPLC scheme with accurate peptide retention prediction in both dimensions for high-throughput bottom-up proteomics. *Anal. Chem.* **80**, 7036–7042 (2008).
27. Krokhin, O. V. Sequence-specific retention calculator. algorithm for peptide retention prediction in ion-pair RP-HPLC: application to 300- and 100-Å pore size C18 sorbents. *Anal. Chem.* **78**, 7785–7795 (2006).
28. Pfeifer, N., Leinenbach, A., Huber, C. G. & Kohlbacher, O. Improving peptide identification in proteome analysis by a two-dimensional retention time filtering approach. *J. Proteome Res.* **8**, 4109–4115 (2009).
29. Giese, S. H., Ishihama, Y. & Rappsilber, J. Peptide retention in hydrophilic strong anion exchange chromatography is driven by charged and aromatic residues. *Anal. Chem.* https://doi.org/10.1021/acs.analchem.7b05157 (2018).
30. Alpert, A. J. et al. Peptide orientation affects selectivity in ion-exchange chromatography. *Anal. Chem.* **82**, 5253–5259 (2010).
31. Yeung, D., Klaassen, N., Mizero, B., Spicer, V. & Krokhin, O. V. Peptide retention time prediction in hydrophilic interaction liquid chromatography: zwitter-ionic sulfoalkylbetaine and phosphorylcholine stationary phases. *J. Chromatogr. A* https://doi.org/10.1016/j.chroma.2020.460909 (2020).
32. Ba, L. J. & Caruana, R. Do deep nets really need to be deep? *Nature* **521**, 436–444 (2013).
33. Tran, N. H., Zhang, X., Xin, L., Shan, B. & Li, M. De novo peptide sequencing by deep learning. *Proc. Natl. Acad. Sci. USA* https://doi.org/10.1073/pnas.1705691114 (2017).
34. Ma, C. et al. Improved peptide retention time prediction in liquid chromatography through deep learning. *Anal. Chem.* **90**, 10881–10888 (2018).
35. Gessulat, S. et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* **16**, 509–518 (2019).
36. Giese, S. H., Belsom, A., Sinn, L., Fischer, L. & Rappsilber, J. Noncovalently associated peptides observed during liquid chromatography-mass spectrometry and their affect on cross-link analyses. *Anal. Chem.* **91**, 2678–2685 (2019).
37. Giese, S. H., Belsom, A. & Rappsilber, J. Optimized fragmentation regime for diazirine photo-cross-linked peptides. *Anal. Chem.* **88**, 8239–8247 (2016).
38. 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.* **8**, 15473 (2017).
39. Walzthoeni, T. et al. False discovery rate estimation for cross-linked peptides identified by mass spectrometry. *Nat. Methods* **9**, 901–903 (2012).
40. Fischer, L. & Rappsilber, J. Quirks of error estimation in cross-linking/mass spectrometry. *Anal. Chem.* **89**, 3829–3833 (2017).
41. Szklarczyk, D. et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gky1131 (2019).
42. Alonso-López, Di. et al. APID database: Redefining protein-protein interaction experimental evidences and binary interactomes. *Database* **2019**, 1–8 (2019).
43. Xu, C. & Ma, B. Software for computational peptide identification from MS-MS data. *Drug Discov. Today* https://doi.org/10.1016/j.drudis.2006.05.011 (2006).
44. Yılmaz, Ş. et al. Cross-linked peptide identification: A computational forest of algorithms. *Mass Spectrom. Rev.* **37**, 738–749 (2018).
45. Ruder, S. An overview of multi-task learning in deep neural networks. Preprint at https://arxiv.org/abs/1706.05098 (2017).
46. Gussakovsky, D., Neustaeter, H., Spicer, V. & Krokhin, O. V. Sequence-specific model for peptide retention time prediction in strong cation exchange chromatography. *Anal. Chem.* **89**, 11795–11802 (2017).
47. Guo, D., Mant, C. T., Taneja, A. K., Parker, J. M. R. & Rodges, R. S. Prediction of peptide retention times in reversed-phase high-performance liquid chromatography I. Determination of retention coefficients of amino acid residues of model synthetic peptides. *J. Chromatogr. A* https://doi.org/10.1016/0021-9673(86)80102-9 (1986).
48. Iacobucci, C. & Sinz, A. To be or not to be? Five guidelines to avoid misassignments in cross-linking/mass spectrometry. *Anal. Chem.* **89**, 7832–7835 (2017).
49. 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* https://doi.org/10.1038/s41592-020-0959-9 (2020).
50. Kessner, D., Chambers, M., Burke, R., Agus, D. & Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **24**, 2534–2536 (2008).
51. Eng, J. K. et al. A deeper look into comet - implementation and features. *J. Am. Soc. Mass Spectrom.* https://doi.org/10.1007/s13361-015-1179-x (2015).
52. 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).
53. Koster, J. & Rahmann, S. Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).
54. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* https://doi.org/10.1073/pnas.1530509100 (2003).
55. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* https://doi.org/10.1016/S0022-2836(05)80360-2 (1990).
56. Shakeel, S. et al. Structure of the Fanconi anaemia monoubiquitin ligase complex. *Nature* **575**, 234–237 (2019).
57. Farrell, D. P. et al. Deep learning enables the atomic structure determination of the Fanconi Anemia core complex from cryoEM. *IUCrJ* **7**, 881–892 (2020).
58. farrell, daniel. Deep learning enables the atomic structure determination of the Fanconi Anemia core complex from cryoEM. https://doi.org/10.5281/ZENODO.3998806 (2020).
59. Graham, M. J., Combe, C., Kolbowski, L. & Rappsilber, J. xiView: a common platform for the downstream analysis of crosslinking mass spectrometry data. Preprint at *bioRxiv* https://doi.org/10.1101/561829 (2019).

60. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
61. Abadi, M. et al. TensorFlow: a system for large-scale machine learning. In *Proc. 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016* (2016).
62. Cheng, J., Wang, Z. & Pollastri, G. A neural network approach to ordinal regression. In *Proc. International Joint Conference on Neural Networks*. https://doi.org/10.1109/IJCNN.2008.4633963 (2008).
63. Berrar, D. in *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*. https://doi.org/10.1016/B978-0-12-809633-8.20349-X (2018).
64. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* https://doi.org/10.1613/jair.953 (2002).
65. Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions. *Nips* **16**, 426–430 (2017).
66. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: uniform manifold approximation and projection. *J. Open Source Softw.* **3**, 861 (2018).
67. Okuda, S. et al. JPOSTrepo: an international standard data repository for proteomes. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkw1080 (2017).

Acknowledgements

We thank Edward Rullmann, Andrea Graziadei, and Francis J. O'Reilly for the critical reading of the manuscript, and Jakub Bartoszewicz (RKI / HPI) for fruitful discussions. We are grateful to Tabea Schütze for help with fermenting *E. coli*. This work was supported by NVIDIA with hardware from the grant "Artificial Intelligence for Deep Structural Proteomics", by the Wellcome Trust through a Senior Research Fellowship to J.R. (103139) and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2008 - 390540038 – UniSysCat, and by grant no. 392923329/GRK2473. The Wellcome Centre for Cell Biology is supported by core funding from the Wellcome Trust (203149).

Author contributions

Study design: S.H.G., L.R.S., and J.R. Software implementation: S.H.G. Sample preparation and mass spectrometry acquisition: L.R.S. and F.W. Data analysis: S.H.G., L.R.S., and J.R. All authors critically evaluated and approved the manuscript.

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/s41467-021-23441-0.

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

Peer review information *Nature Communications* thanks Wen-Feng Zeng and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

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

Chapter 7

Outlook

The analysis of single proteins and protein complexes via crosslinking mass spectrometry is becoming an important part of many integrated structural biology approaches. The advancements in the field of structural biology will also increase the demand for structural data under physiological conditions to complement traditional structure determination methods. The necessity to choose the best suiting crosslinking chemistry, acquisition schema, identification workflows, and proper error estimation based on the biological question makes the field not easily accessible to a large audience. For example, the use of promiscuous crosslinkers can increase the distance constraints from crosslinking of a single protein or small complex, by at least an order of magnitude. At the same time, the promiscuous crosslinking chemistry could overwhelm current database search paradigms in proteome-wide studies, without proper enrichment.
The ambitious goal to analyze snapshots of the entire interactome of a cell in a time-resolved manner thus faces several challenges; even if crosslinking mass spectrometry is undergoing rapid changes with the development of new enrichment strategies, instruments, fragmentation techniques, and (faster) crosslinking chemistry. These advancements bring the field stepwise closer to *true* proteome-wide studies that cover a large part of the proteome. Computational approaches will also continuously play a vital role in the success of CLMS. So far, a sophisticated sensitivity analysis comparing heuristic and exhaustive search strategies remains to be done. For example, with the availability of powerful graphics processing units, more costly search strategies are feasible. In addition to the ever-increasing computing capacities, machine learning methods become more powerful with more data at hand. With the ability to predict each step of the liquid chromatography-mass spectrometry workflow for crosslinks, interesting possibilities open. For instance, access to the application programming interfaces of the mass spectrometer vendors would allow to target crosslinks more reliably or enable live searching of spectra. But also, traditional database searches can benefit from accurate retention time prediction. Either through post-search rescoring or through a more targeted database search, where only peptides are considered that are predicted to elute with the measured precursor.
For CLMS, the next years will trigger a similar revolution as linear proteomics, with further improvements in post-search validation algorithms, seamless identification and quantitation, and more evolved data interpretation techniques. Lastly, it is worthwhile to think outside the mass spectrometry field for the analysis of proteins. An interesting development is the possibility to sequence proteins with the nanopore. Even though this approach is still in its infancy it will certainly extend the toolbox for studying proteomics.

Bibliography

Acuner Ozbabacan, Saliha Ece et al. (2011). "Transient protein-protein interactions". In: *Protein Eng. Des. Sel.* 24.9, pp. 635–648. ISSN: 1741-0126. DOI: 10.1093/protein/gzr025. URL: https://academic.oup.com/peds/article-lookup/doi/10.1093/protein/gzr025.

Belsom, Adam et al. (2016). "Serum Albumin Domain Structures in Human Blood Serum by Mass Spectrometry and Computational Biology". In: *Mol. Cell. Proteomics* 15.3, pp. 1105–1116. ISSN: 1535-9476. DOI: 10.1074/mcp.M115.048504. URL: http://www.mcponline.org/lookup/doi/10.1074/mcp.M115.048504.

Bouwmeester, Robbin et al. (2020). "The Age of Data-Driven Proteomics: How Machine Learning Enables Novel Workflows". In: *Proteomics*, p. 1900351. DOI: 10.1002/pmic.201900351.

C. Silva, Ana S et al. (2019). "Accurate peptide fragmentation predictions allow data driven approaches to replace and improve upon proteomics search engine scoring functions". In: *Bioinformatics* 35.24. Ed. by Jonathan Wren, pp. 5243–5248. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz383. URL: https://academic.oup.com/bioinformatics/article/35/24/5243/5488123.

Cerofolini, Linda et al. (2019). "Integrative Approaches in Structural Biology: A More Complete Picture from the Combination of Individual Techniques". In: *Biomolecules* 9.8, p. 370. ISSN: 2218-273X. DOI: 10.3390/biom9080370. URL: https://www.mdpi.com/2218-273X/9/8/370.

Chen, Ke, Lukasz A. Kurgan, and Jishou Ruan (2007). "Prediction of flexible/rigid regions from protein sequences using k-spaced amino acid pairs". In: *BMC Struct. Biol.* 7.1, p. 25. ISSN: 14726807. DOI: 10.1186/1472-6807-7-25. URL: http://bmcstructbiol.biomedcentral.com/articles/10.1186/1472-6807-7-25.

Chen, Zhen-Lin et al. (2019). "A high-speed search engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides". In: *Nat. Commun.* 10.1, p. 3404. ISSN: 2041-1723. DOI: 10.1038/s41467-019-11337-z. URL: http://www.nature.com/articles/s41467-019-11337-z.

Chen, Zhuo Angel et al. (2010). "Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry." In: *EMBO J.* 29.4, pp. 717–26. ISSN: 1460-2075. DOI: 10.1038/emboj.2009.401. URL: http://www.ncbi.nlm.nih.gov/pubmed/20094031.

Cheng, Yifan (2015). "Single-Particle Cryo-EM at Crystallographic Resolution". In: *Cell* 161.3, pp. 450–457. ISSN: 00928674. DOI: 10.1016/j.cell.2015.03.049. URL: http://dx.doi.org/10.1016/j.cell.2015.03.049https://linkinghub.elsevier.com/retrieve/pii/S0092867415003694.

Chu, Feixia, Daniel T. Thornton, and Hieu T. Nguyen (2018). *Chemical cross-linking in the structural analysis of protein assemblies*. DOI: 10.1016/j.ymeth.2018.05.023.

Craveur, Pierrick et al. (2015). "Protein flexibility in the light of structural alphabets". In: *Front. Mol. Biosci.* 2.MAY, pp. 1–20. ISSN: 2296-889X. DOI: 10.3389/fmolb.2015.00020. URL: http://www.frontiersin.org/Structural{_}Biology/10.3389/fmolb.2015.00020/abstract.

Diedrich, Jolene K., Antonio F. M. Pinto, and John R. Yates (2013). "Energy Dependence of HCD on Peptide Fragmentation: Stepped Collisional Energy Finds the Sweet Spot". In: *J. Am. Soc. Mass Spectrom.* 24.11, pp. 1690–1699. ISSN: 1044-0305. DOI: 10.1007/s13361-013-0709-7. arXiv: NIHMS150003. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3624763/pdf/nihms412728.pdf.

Fang, Tiantian (2018). "A Novel Computer-Aided Lung Cancer Detection Method Based on Transfer Learning from GoogLeNet and Median Intensity Projections". In: *2018 IEEE Int. Conf. Comput. Commun. Eng. Technol. CCET 2018*. ISBN: 9781538674376. DOI: 10.1109/CCET.2018.8542189.

Gessulat, Siegfried et al. (2019). "Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning". In: *Nat. Methods* 16.6, pp. 509–518. ISSN: 15487105. DOI: 10.1038/s41592-019-0426-7. URL: http://dx.doi.org/10.1038/s41592-019-0426-7.

Giese, Sven H., Adam Belsom, and Juri Rappsilber (2016). "Optimized fragmentation regime for diazirine photo-cross-linked peptides". In: *Anal. Chem.* 88.16, pp. 8239–8247. ISSN: 15206882. DOI: 10.1021/acs.analchem.6b02082. URL: https://www.ncbi.nlm.nih.gov/pubmed/27454319.

Giese, Sven H, Lutz Fischer, and Juri Rappsilber (2015). "A study into the CID behavior of cross-linked peptides." In: *Mol. Cell. Proteomics*, pp. 1094–1104. ISSN: 1535-9484. DOI: 10.1074/mcp.M115.049296. URL: http://www.ncbi.nlm.nih.gov/pubmed/26719564.

Giese, Sven H., Lutz Fischer, and Juri Rappsilber (2016). "A Study into the Collision-induced Dissociation (CID) Behavior of Cross-Linked Peptides". In: *Mol. Cell. Proteomics* 15.3, pp. 1094–1104. ISSN: 1535-9476. DOI: 10.1074/mcp.M115.049296. URL: http://www.mcponline.org/lookup/doi/10.1074/mcp.M115.049296.

Giese, Sven H., Yasushi Ishihama, and Juri Rappsilber (2018). "Peptide Retention in Hydrophilic Strong Anion Exchange Chromatography Is Driven by Charged and Aromatic Residues". In: *Anal. Chem.*, acs.analchem.7b05157. ISSN: 0003-2700. DOI: 10.1021/acs.analchem.7b05157. URL: http://pubs.acs.org/doi/10.1021/acs.analchem.7b05157.

Giese, Sven H. et al. (2019). "Noncovalently Associated Peptides Observed during Liquid Chromatography-Mass Spectrometry and Their Affect on Cross-Link Analyses". In: *Anal. Chem.* 91, pp. 2678–2685. ISSN: 15206882. DOI: 10.1021/acs.analchem.8b04037.

Gonzalez-Lozano, M. A. et al. (2020). "Stitching the synapse: Cross-linking mass spectrometry into resolving synaptic protein interactions". In: *Sci. Adv.* 6.8, eaax5783. ISSN: 2375-2548. DOI: 10.1126/sciadv.aax5783. URL: https://advances.sciencemag.org/lookup/doi/10.1126/sciadv.aax5783.

Götze, Michael et al. (2019). "A Simple Cross-Linking/Mass Spectrometry Workflow for Studying System-wide Protein Interactions". In: *Anal. Chem.* 91.15, pp. 10236–10244. ISSN: 0003-2700. DOI: 10.1021/acs.analchem.9b02372. URL: https://pubs.acs.org/doi/10.1021/acs.analchem.9b02372.

Hoopmann, Michael R. et al. (2015). "Kojak: efficient analysis of chemically cross-linked protein complexes." In: *J. Proteome Res.* 14.5, pp. 2190–8. ISSN: 1535-3907. DOI: 10.1021/pr501321h. URL: http://www.ncbi.nlm.nih.gov/pubmed/25812159.

Iacobucci, Claudio and Andrea Sinz (2017). "To Be or Not to Be? Five Guidelines to Avoid Misassignments in Cross-Linking/Mass Spectrometry". In: *Anal. Chem.* 89.15, pp. 7832–7835. ISSN: 0003-2700. DOI: 10.1021/acs.analchem.7b02316. URL: https://pubs.acs.org/doi/10.1021/acs.analchem.7b02316.

Ikeya, Teppei, Peter Güntert, and Yutaka Ito (2019). "Protein structure determination in living cells". In: *Int. J. Mol. Sci.* 20.10, pp. 1–13. ISSN: 14220067. DOI: 10.3390/ijms20102442.

Joice, Regina et al. (2014). "characterization of protein crosslinks via MS and an open-modification search strategy". In: 6.244, pp. 1–16. ISSN: 15378276. DOI: 10.1126/scitranslmed.3008882.Plasmodium. arXiv: NIHMS150003.

Käll, Lukas et al. (2007). "Semi-supervised learning for peptide identification from shotgun proteomics datasets." In: *Nat. Methods* 4.11, pp. 923–5. ISSN: 1548-7091. DOI: 10.1038/nmeth1113. URL: http://www.ncbi.nlm.nih.gov/pubmed/17952086.

Kao, Athit et al. (2011). "Development of a novel cross-linking strategy for fast and accurate identification of cross-linked peptides of protein complexes." In: *Mol. Cell. Proteomics* 10.1, p. M110.002212. ISSN: 1535-9484. DOI: 10.1074/mcp.M110.002212. URL: http://www.ncbi.nlm.nih.gov/pubmed/20736410.

Kao, Athit et al. (2012). "Mapping the Structural Topology of the Yeast 19S Proteasomal Regulatory Particle Using Chemical Cross-linking and Probabilistic Modeling". In: *Mol. Cell. Proteomics* 11.12, pp. 1566–1577. ISSN: 1535-9476. DOI: 10.1074/mcp.M112.018374. URL: http://www.mcponline.org/lookup/doi/10.1074/mcp.M112.018374.

Keskin, Ozlem, Nurcan Tuncbag, and Attila Gursoy (2016). "Predicting Protein-Protein Interactions from the Molecular to the Proteome Level". In: *Chem. Rev.* 116.8, pp. 4884–4909. ISSN: 15206890. DOI: 10.1021/acs.chemrev.5b00683.

Klammer, Aaron A et al. (2007). "Improving tandem mass spectrum identification using peptide retention time prediction across diverse chromatography conditions." In: *Anal. Chem.* 79.16, pp. 6111–8. ISSN: 0003-2700. DOI: 10.1021/ac070262k. URL: http://www.ncbi.nlm.nih.gov/pubmed/17622186.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep learning". In: *Nature* 521.7553, pp. 436–444. ISSN: 0028-0836. DOI: 10.1038/nature14539. arXiv: arXiv:1312.6184v5. URL: http://www.nature.com/articles/nature14539.

Leitner, Alexander (2016). "Cross-linking and other structural proteomics techniques: How chemistry is enabling mass spectrometry applications in structural biology". In: *Chem. Sci.* 7.8, pp. 4792–4803. ISSN: 20416539. DOI: 10.1039/c5sc04196a. URL: http://xlink.rsc.org/?DOI=C5SC04196A.

Lenz, Swantje et al. (2018). "In-Search Assignment of Monoisotopic Peaks Improves the Identification of Cross-Linked Peptides". In: *J. Proteome Res.* 17.11, pp. 3923–3931. ISSN: 1535-3893. DOI: 10.1021/acs.jproteome.8b00600. URL: https://pubs.acs.org/doi/10.1021/acs.jproteome.8b00600.

Liu, Fan et al. (2015). "Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry." In: *Nat. Methods* 12.12, pp. 1179–84. ISSN: 1548-7105. DOI: 10.1038/nmeth.3603. URL: http://www.nature.com/doifinder/10.1038/nmeth.3603http://www.ncbi.nlm.nih.gov/pubmed/26414014.

Liu, Fan et al. (2017). "Optimized fragmentation schemes and data analysis strategies for proteome-wide cross-link identification." In: *Nat. Commun.* 8.May, p. 15473. ISSN: 2041-1723. DOI: 10.1038/ncomms15473. arXiv: 0507126 [physics]. URL: http://www.ncbi.nlm.nih.gov/pubmed/28524877.

Liu, Fan et al. (2018). "The interactome of intact mitochondria by cross-linking mass spectrometry provides evidence for coexisting respiratory supercomplexes". In: *Mol. Cell. Proteomics* 17.2, pp. 216–232. ISSN: 1535-9476. DOI: 10.1074/mcp.RA117.000470. URL: http://www.mcponline.org/lookup/doi/10.1074/mcp.RA117.000470.

Liu, Hsuan-Liang and Jyh-Ping Hsu (2005). "Recent developments in structural proteomics for protein structure determination". In: *Proteomics* 5.8, pp. 2056–2068. ISSN: 1615-9853. DOI: 10.1002/pmic.200401104. URL: http://doi.wiley.com/10.1002/pmic.200401104.

Lu, Lei et al. (2018). "Identification of MS-Cleavable and Noncleavable Chemically Cross-Linked Peptides with MetaMorpheus". In: *J. Proteome Res.* 17.7, pp. 2370–2376. ISSN: 1535-3893. DOI: 10.1021/acs.jproteome.8b00141. URL: https://pubs.acs.org/doi/10.1021/acs.jproteome.8b00141.

Luchinat, Enrico and Lucia Banci (2017). "In-cell NMR: a topical review". In: *IUCrJ* 4.2, pp. 108–118. ISSN: 2052-2525. DOI: 10.1107/S2052252516020625. URL: http://scripts.iucr.org/cgi-bin/paper?S2052252516020625.

Ma, Kelvin, Olga Vitek, and Alexey I. Nesvizhskii (2012). "A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet". In: *BMC Bioinformatics* 13.Suppl 16, S1. ISSN: 1471-2105. DOI: 10.1186/1471-2105-13-S16-S1. URL: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-S16-S1.

MacLean, Brendan et al. (2010). "Skyline: An open source document editor for creating and analyzing targeted proteomics experiments". In: *Bioinformatics*. ISSN: 13674803. DOI: 10.1093/bioinformatics/btq054.

Mehta, Virja and Laura Trinkle-Mulcahy (2016). "Recent advances in large-scale protein interactome mapping". In: *F1000Research* 5.0, p. 782. ISSN: 2046-1402. DOI: 10.12688/f1000research.7629.1. URL: https://f1000research.com/articles/5-782/v1.

Mendes, Marta L et al. (2019). "An integrated workflow for crosslinking mass spectrometry". In: *Mol. Syst. Biol.* 15.9, e8994. ISSN: 1744-4292. DOI: 10.15252/msb.20198994. URL: https://onlinelibrary.wiley.com/doi/abs/10.15252/msb.20198994.

Merkley, Eric D. et al. (2014). "Distance restraints from crosslinking mass spectrometry: Mining a molecular dynamics simulation database to evaluate lysine-lysine distances". In: *Protein Sci.* 23.6, pp. 747–759. ISSN: 09618368. DOI: 10.1002/pro.2458. URL: http://doi.wiley.com/10.1002/pro.2458.

Miao, Zhichao and Yang Cao (2016). "Quantifying side-chain conformational variations in protein structure". In: *Sci. Rep.* 6.1, p. 37024. ISSN: 2045-2322. DOI: 10.1038/srep37024. URL: http://www.nature.com/articles/srep37024.

Miura, Kenji (2018). "An Overview of Current Methods to Confirm Protein- Protein Interactions". In: *Protein Pept. Lett.* 25.8, pp. 728–733. ISSN: 09298665. DOI: 10.2174/0929866525666180821122240. URL: http://www.eurekaselect.com/164836/article.

Müller, Fränze et al. (2018). "On the Reproducibility of Label-Free Quantitative Cross-Linking/Mass Spectrometry". In: *J. Am. Soc. Mass Spectrom.* 29.2, pp. 405–412. ISSN: 1044-0305. DOI: 10.1007/s13361-017-1837-2. URL: http://link.springer.com/10.1007/s13361-017-1837-2.

Müller, Fränze et al. (2019). "Data-independent Acquisition Improves Quantitative Cross-linking Mass Spectrometry". In: *Mol. Cell. Proteomics* 18.4, pp. 786–795. ISSN: 1535-9476. DOI: 10.1074/mcp.TIR118.001276. URL: http://www.mcponline.org/lookup/doi/10.1074/mcp.TIR118.001276.

Nooren, Irene M.A. (2003). "NEW EMBO MEMBER'S REVIEW: Diversity of protein-protein interactions". In: *EMBO J.* 22.14, pp. 3486–3492. ISSN: 1460-2075. DOI: 10.1093/emboj/cdg359. URL: http://emboj.embopress.org/cgi/doi/10.1093/emboj/cdg359.

O'Reilly, Francis J. and Juri Rappsilber (2018). "Cross-linking mass spectrometry: methods and applications in structural, molecular and systems biology". In: *Nat. Struct. Mol. Biol.* 25.11, p. 1. ISSN: 1545-9993. DOI: 10.1038/s41594-018-0147-0. URL: http://dx.doi.org/10.1038/s41594-018-0147-0http://www.nature.com/articles/s41594-018-0147-0.

O'Reilly, Francis J et al. (2020). "In-cell architecture of an actively transcribing-translating expressome". In: *bioRxiv*, p. 2020.02.28.970111. DOI: 10.1101/2020.02.28.970111. URL: http://biorxiv.org/content/early/2020/02/28/2020.02.28.970111.abstract.

Rappsilber, Juri (2011). "The beginning of a beautiful friendship: Cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes". In: *J. Struct. Biol.* 173.3, pp. 530–540. ISSN: 10478477. DOI: 10.1016/j.jsb.2010.10.014. URL: http://www.ncbi.nlm.nih.gov/pubmed/21029779.

Renard, Bernhard Y. et al. (2010). "Estimating the confidence of peptide identifications without decoy databases." In: *Anal. Chem.* 82.11, pp. 4314–8. ISSN: 1520-6882. DOI: 10.1021/ac902892j. URL: http://www.ncbi.nlm.nih.gov/pubmed/20455556.

Révész, Ágnes et al. (2018). "Selection of Collision Energies in Proteomics Mass Spectrometry Experiments for Best Peptide Identification: Study of Mascot Score Energy Dependence Reveals Double Optimum". In: *J. Proteome Res.* 17.5, pp. 1898–1906. ISSN: 1535-3893. DOI: 10.1021/acs.jproteome.7b00912. URL: https://pubs.acs.org/doi/10.1021/acs.jproteome.7b00912.

Ritorto, Maria Stella et al. (2013). "Hydrophilic Strong Anion Exchange (hSAX) Chromatography for Highly Orthogonal Peptide Separation of Complex Proteomes". In: *J. Proteome Res.* 12.6, pp. 2449–2457. ISSN: 1535-3893. DOI: 10.1021/pr301011r. URL: http://pubs.acs.org/doi/abs/10.1021/pr301011r.

Robinson, Philip J. et al. (2015). "Molecular architecture of the yeast Mediator complex". In: *Elife* 4.September2015, pp. 1–29. ISSN: 2050084X. DOI: 10.7554/eLife.08719.

Ruprecht, Benjamin et al. (2017). "Hydrophilic Strong Anion Exchange (hSAX) Chromatography Enables Deep Fractionation of Tissue Proteomes." In: *Methods Mol. Biol.* Methods in Molecular Biology 1550. Ed. by Lucio Comai, Jonathan E. Katz, and Parag Mallick, pp. 69–82. ISSN: 1940-6029. DOI: 10.1007/978-1-4939-6747-6_7. URL: http://www.ncbi.nlm.nih.gov/pubmed/28188524.

Ryl, Petra S.J. et al. (2020). "In Situ Structural Restraints from Cross-Linking Mass Spectrometry in Human Mitochondria". In: *J. Proteome Res.* 19.1, pp. 327–336. ISSN: 15353907. DOI: 10.1021/acs.jproteome.9b00541.

Sali, Andrej et al. (2003). "From words to literature in structural proteomics". In: *Nature* 422.6928, pp. 216–225. ISSN: 0028-0836. DOI: 10.1038/nature01513. URL: http://www.nature.com/articles/nature01513.

Schmidt, Carla and Henning Urlaub (2017). *Combining cryo-electron microscopy (cryo-EM) and cross-linking mass spectrometry (CX-MS) for structural elucidation of large protein assemblies*. DOI: 10.1016/j.sbi.2017.10.005.

Sinz, Andrea (2017). "Divide and conquer: cleavable cross-linkers to study protein conformation and protein–protein interactions". In: *Anal. Bioanal. Chem.* 409.1, pp. 33–44. ISSN: 16182650. DOI: 10.1007/s00216-016-9941-x. URL: http://dx.doi.org/10.1007/s00216-016-9941-x.

— (2018). "Cross-Linking/Mass Spectrometry for Studying Protein Structures and Protein-Protein Interactions: Where Are We Now and Where Should We Go from Here?" In: *Angew. Chem. Int. Ed. Engl.* 57.22, pp. 6390–6396. ISSN: 1521-3773. DOI:

10.1002/anie.201709559. URL: http://www.ncbi.nlm.nih.gov/pubmed/29334167.

Smits, Arne H. and Michiel Vermeulen (2016). "Characterizing Protein-Protein Interactions Using Mass Spectrometry: Challenges and Opportunities". In: *Trends Biotechnol.* xx, pp. 1–10. ISSN: 18793096. DOI: 10.1016/j.tibtech.2016.02.014. URL: http://dx.doi.org/10.1016/j.tibtech.2016.02.014.

Steen, Hanno and Matthias Mann (2004). "The ABC's (and XYZ's) of peptide sequencing." In: *Nat. Rev. Mol. Cell Biol.* 5.9, pp. 699–711. ISSN: 1471-0072. DOI: 10.1038/nrm1468. URL: http://www.ncbi.nlm.nih.gov/pubmed/15340378.

Steigenberger, B. et al. (2020). "To Cleave or Not To Cleave in XL-MS?" In: *J. Am. Soc. Mass Spectrom.* 31.2, pp. 196–206. ISSN: 1044-0305. DOI: 10.1021/jasms.9b00085. URL: https://pubs.acs.org/doi/10.1021/jasms.9b00085.

Stieger, Christian E., Philipp Doppler, and Karl Mechtler (2019). "Optimized Fragmentation Improves the Identification of Peptides Cross-Linked by MS-Cleavable Reagents". In: *J. Proteome Res.* 18.3, pp. 1363–1370. ISSN: 1535-3893. DOI: 10.1021/acs.jproteome.8b00947. URL: https://pubs.acs.org/doi/10.1021/acs.jproteome.8b00947.

Swaney, Danielle L, Graeme C McAlister, and Joshua J Coon (2008). "Decision tree-driven tandem mass spectrometry for shotgun proteomics." In: *Nat. Methods* 5.11, pp. 959–64. ISSN: 1548-7105. DOI: 10.1038/nmeth.1260. arXiv: NIHMS150003. URL: http://www.ncbi.nlm.nih.gov/pubmed/18931669.

Trnka, Michael J et al. (2014). "Matching Cross-linked Peptide Spectra: Only as Good as the Worse Identification". In: *Mol. Cell. Proteomics* 13.2, pp. 420–434. ISSN: 1535-9476. DOI: 10.1074/mcp.M113.034009. URL: http://www.ncbi.nlm.nih.gov/pubmed/24335475.

Walzthoeni, Thomas et al. (2013). "Mass spectrometry supported determination of protein complex structure." In: *Curr. Opin. Struct. Biol.* 23.2, pp. 252–60. ISSN: 1879-033X. DOI: 10.1016/j.sbi.2013.02.008. URL: http://www.ncbi.nlm.nih.gov/pubmed/23522702.

Yang, Bing et al. (2012). "Identification of cross-linked peptides from complex samples." In: *Nat. Methods* 9.9, pp. 904–6. ISSN: 1548-7105. DOI: 10.1038/nmeth.2099. URL: /Users/yurikoharigaya/Documents/ReadCubeMedia/yang2012.pdf{\%}5Cnhttp://dx.doi.org/10.1038/nmeth.2099http://www.ncbi.nlm.nih.gov/pubmed/22772728.

Yılmaz, Şule et al. (2018). "Cross-linked peptide identification: A computational forest of algorithms". In: *Mass Spectrom. Rev.* 37.6, pp. 738–749. ISSN: 02777037. DOI: 10.1002/mas.21559. URL: http://doi.wiley.com/10.1002/mas.21559.

Yu, Clinton and Lan Huang (2018). "Cross-Linking Mass Spectrometry: An Emerging Technology for Interactomics and Structural Biology". In: *Anal. Chem.* 90.1, pp. 144–165. ISSN: 0003-2700. DOI: 10.1021/acs.analchem.7b04431. URL: http://pubs.acs.org/doi/10.1021/acs.analchem.7b04431.

Zhou, Xie Xuan et al. (2017). "PDeep: Predicting MS/MS Spectra of Peptides with Deep Learning". In: *Anal. Chem.* 89.23, pp. 12690–12697. ISSN: 15206882. DOI: 10.1021/acs.analchem.7b02566.

Supplementary Material: A study into the CID behavior of cross-linked peptides

Sven H. Giese[1, 2], Lutz Fischer[1], and Juri Rappsilber*[1, 2]

[1]Department of Bioanalytics, Institute of Biotechnology, Technische Universität Berlin, 13355 Berlin, Germany
[2]Wellcome Trust Centre for Cell Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3BF, United Kingdom

1 List of Supplementary Material

S1: Output from XiFDR containing tab-delimited information about each PSM

S2: Annotated spectra for all cross-link identifications from S1

S3: Distribution of residues involved in cross-link bonds

S4: Factors influencing the fragmentation similarity

S5: Sensitivity and specificity for transforming the highest intense ions

S6: Single spectra that are used in the manuscript (Fig. 3A)

S7: Linear spectrum corresponding to the alpha peptide in the manuscript (Fig. 4D)

S8: Cross-linker fragmentation observations

S9: Comparison of Figure 1 results based on 910 PSMs wtih results based on 8,301 PSMs

S10: Comparison of Figure 2 results based on 910 PSMs wtih results based on 8,301 PSMs

S11: Comparison of Figure 3 results based on 910 PSMs wtih results based on 8,301 PSMs

S12: Comparison of Figure 4 results based on 910 PSMs wtih results based on 8,301 PSMs

S1: XiFDR output

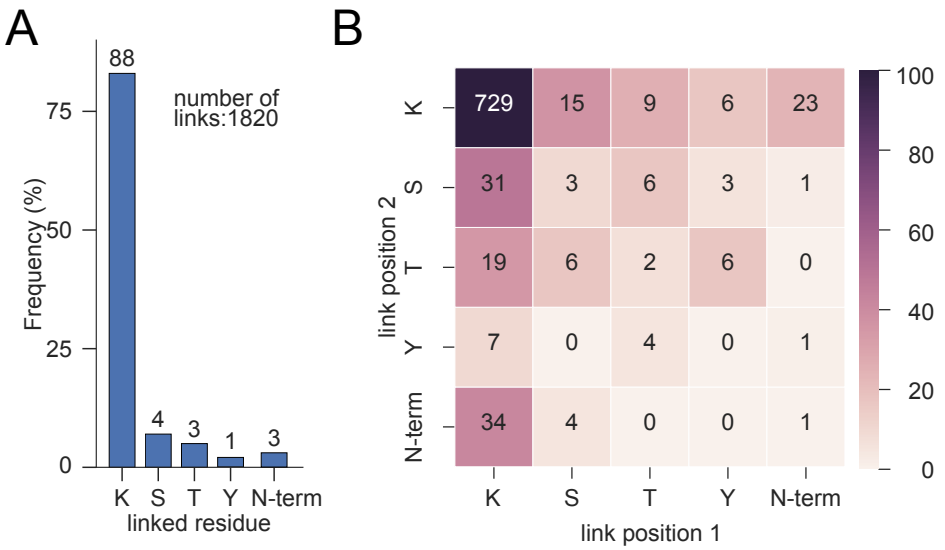
Available online as separate Excel file.

S2: Annotated spectra

Available online as separate PDF. Peptides at the top are alpha peptides (in red); peptides at the bottom are beta peptides (in blue). Annotation: Loss ions (water, ammonia) are not annotated but the peaks are highlighted with light red/blue colors. Cross-linked fragments are annotated as +P. Precursor ions as P. P+i(P) denotes ions that have a modified lysine rest on one end of the cross-link and on the other the complete second peptide. P(+P) denotes the individual peptide fragments without the cross-linker mass and P+(P) refers to the individual peptides with the cross-linker attached. Mean(ppmerror) and std(ppmerror) refer to the measurement error on the fragment peak matching.

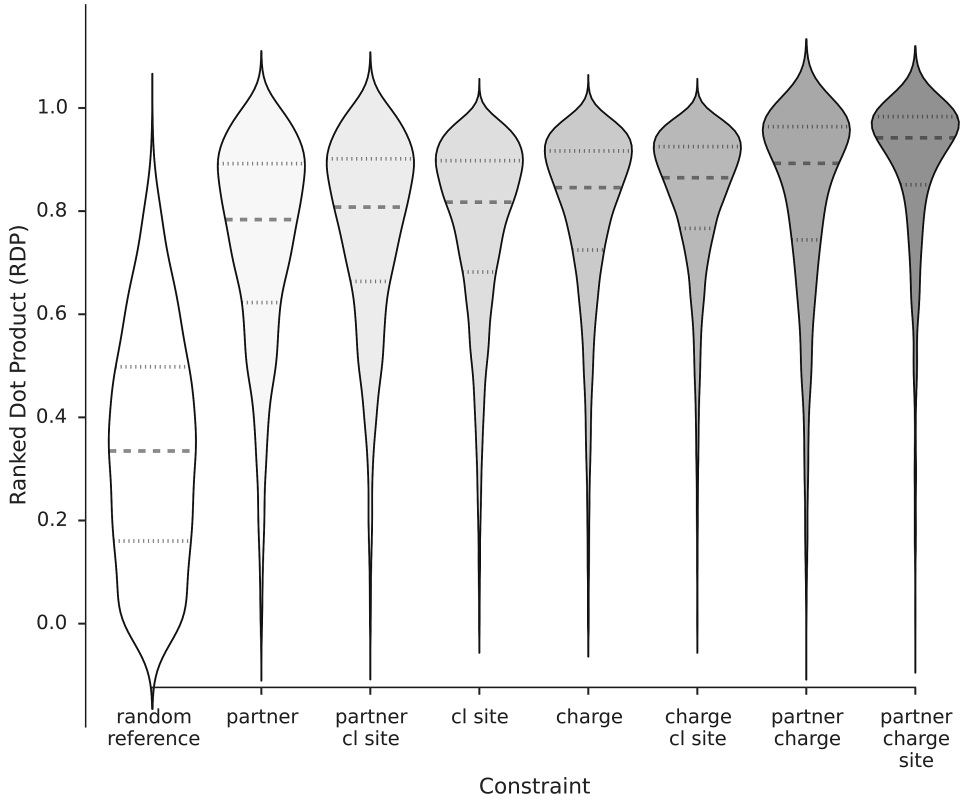
*juri.rappsilber@tu-berlin.de

S3: Cross-linked residue distribution



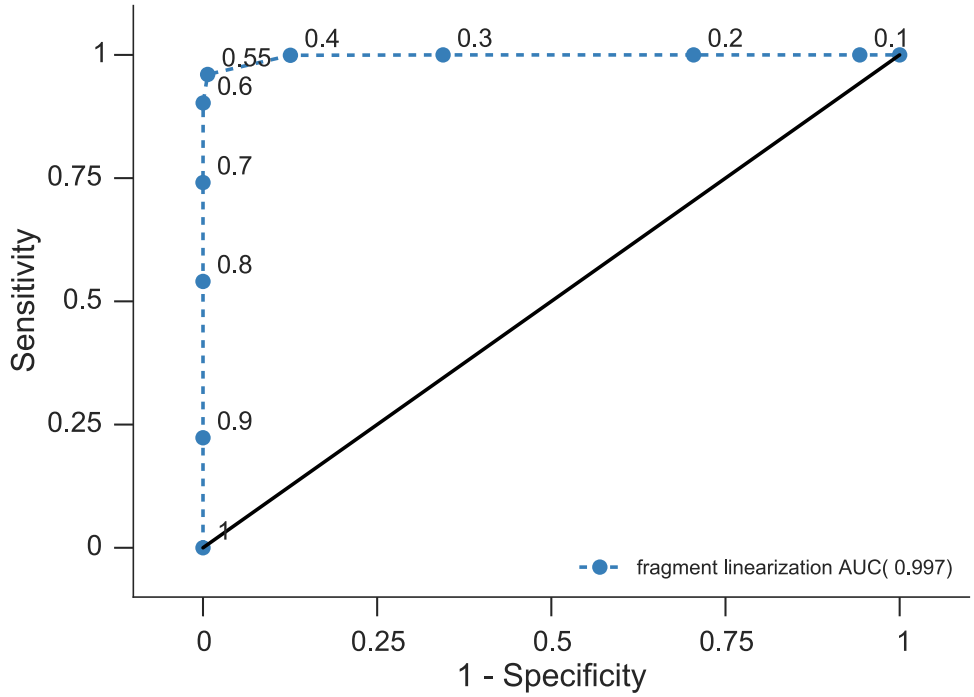
Supplementary Figure 3: Cross-linked residue distribution. Frequency distribution of all observed cross-linked residues from 910 PSMs (**A**). Heatmap showing all observed linkage combinations based on BS3 cross-linking. Annotations in the cells refer to the absolute counts of an observed linkage pair. Cross-links involving serine, threonine or tyrosine in at least one linkage site account for roughly 14% of the cross-links (**B**).

S4: Cross-link factors that influence fragmentation



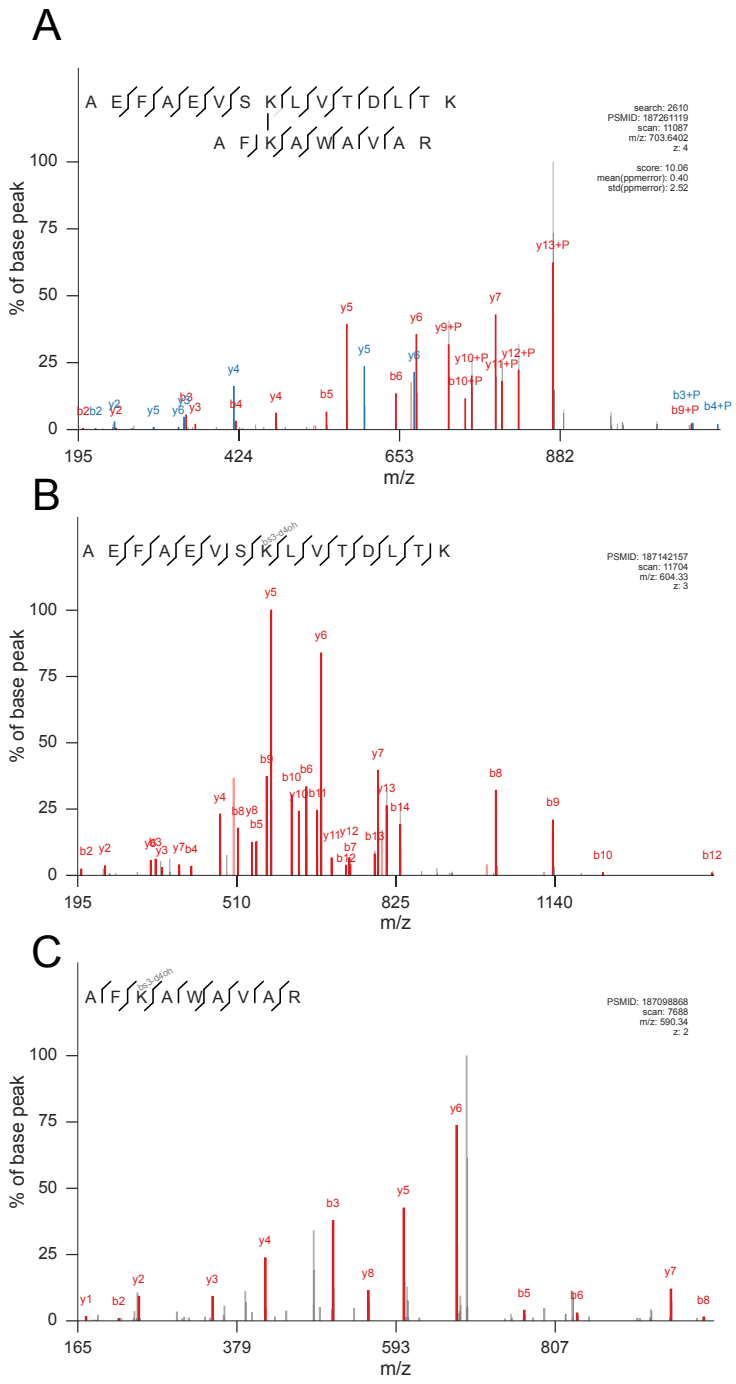
Supplementary Figure 4: Influence on the cross-linking context on the spectral similarity. The x-axis describes the similarity class constraints for the comparison. For example, in the 'partner' class the cross-linked peptides that were compared against each other had to be identified with the same partner. Multiple constraints are simply combinations of individual classes. A reference distribution is derived by randomly comparing spectra of cross-linked peptides. The data was derived from 8,301 high-confidence identifications by XiFDR with a 5% false discovery rate (FDR). In addition, each peptide in a cross-link was required to be six amino acids or longer. Note that this Figure includes data from unpublished acquisitions. Abbreviations: partner - partner peptide in a cross-link, cl site - cross-linking site position. Data were derived from additional unpublished data from our in-house database.

S5: Fragment linearization among the top ten ions



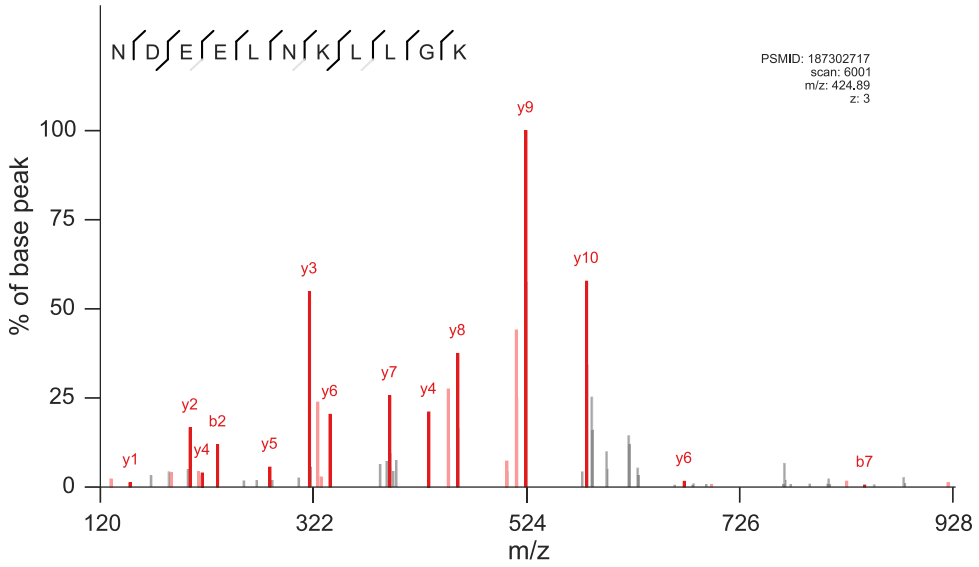
Supplementary Figure 5: ROC curve evaluating the linearization of the ten highest intense (and identified) ions. Sensitivity and specificity are defined as in the manuscript taking only fragments from the alpha peptide into account. Annotations in the plot refer to the relative mass cut-off that is used to decide whether or not to linearize a fragment.

S6: Individual spectra



Supplementary Figure 6: Peak annotation for the individual spectra used in the overlay spectrum shown in the manuscript (Fig 3A). The cross-linked PSM **(A)**, with the alpha peptide in red (upper peptide sequence) and the beta peptide in blue (lower peptide sequence), the alpha peptide **(B)** and the beta peptide **(C)** are shown. Additional information such as precursor mass (m/z), precursor charge (z), scan number, and PSMID (unique identifier) are annotated.

S7: Linear spectrum



Supplementary Figure 7: Linear spectrum corresponding to the alpha peptide shown in the manuscript (Fig. 4D). Additional information such as precursor mass (m/z), precursor charge (z), scan number, and PSMID (unique identifier) are annotated.

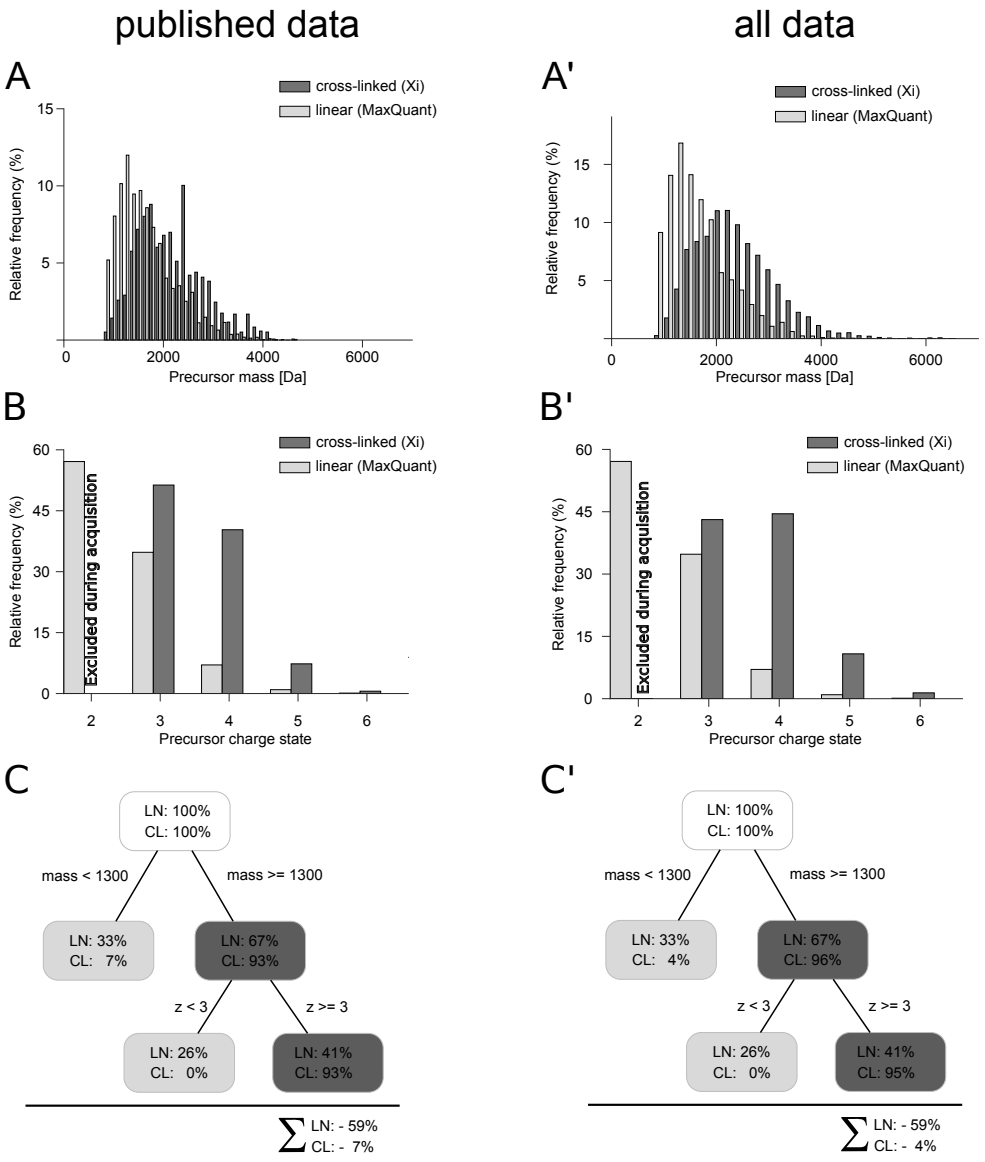
S8: Cross-linker fragmentation

count type / ion type	immonium type (P+i(P))	peptide + cross-linker loss (P(+P))	peptide loss (P+(P))
per cross-link	3%	16%	7%
per peptide	2%	10%	4%

Supplementary Table 1: Cross-linker fragmentation frequencies. P+i(P)) type ions have a modified lysine rest on one end of the cross-link and on the other the complete second peptide. P(+P) refers to the individual peptide fragments without the cross-linker mass and P+(P) refers to the individual peptides with the cross-linker attached.

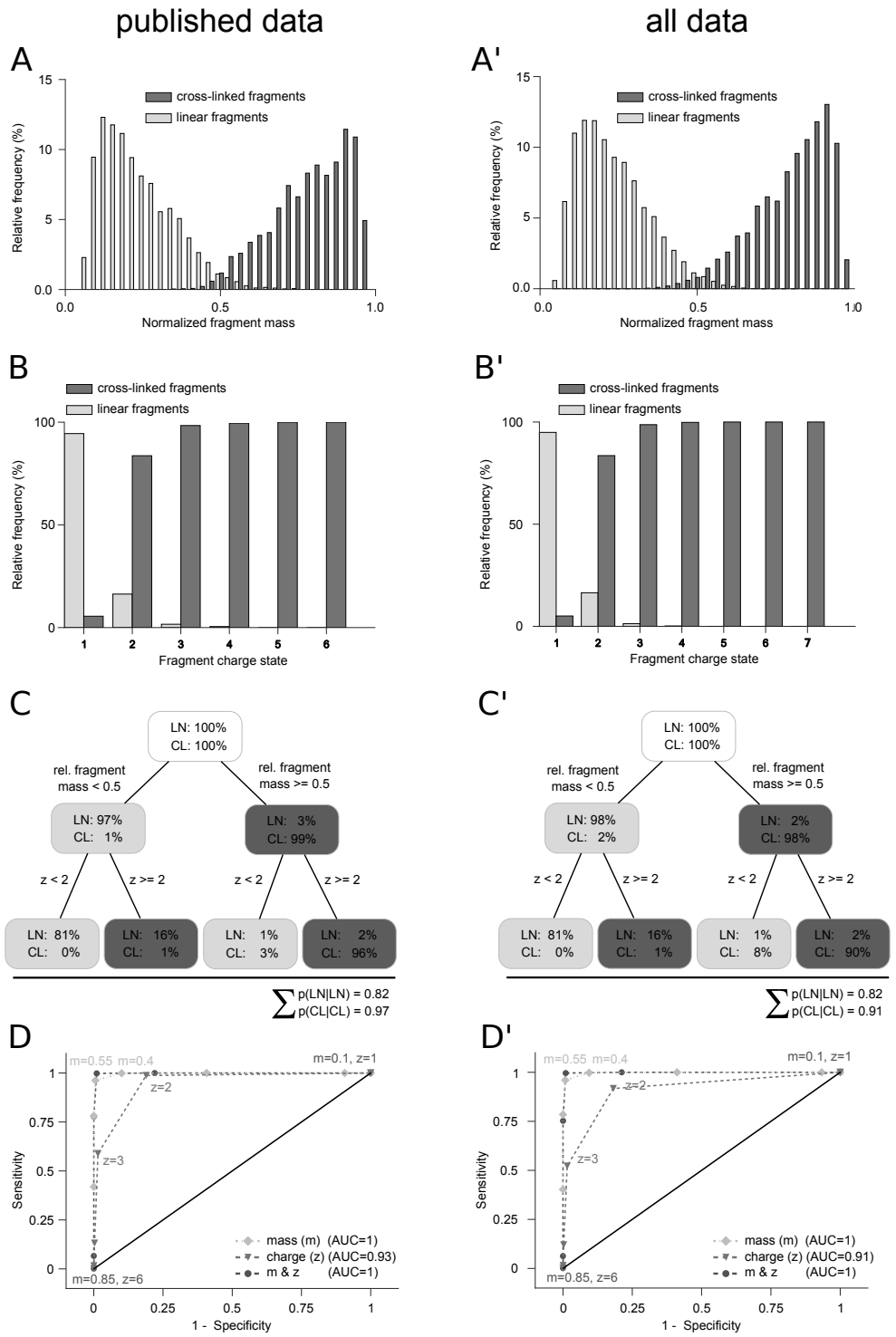
S9: Figure 1 comparison to a larger sample

Section S9-S12 compare the results form the manuscript with another larger sample from our local database including unpublished data.



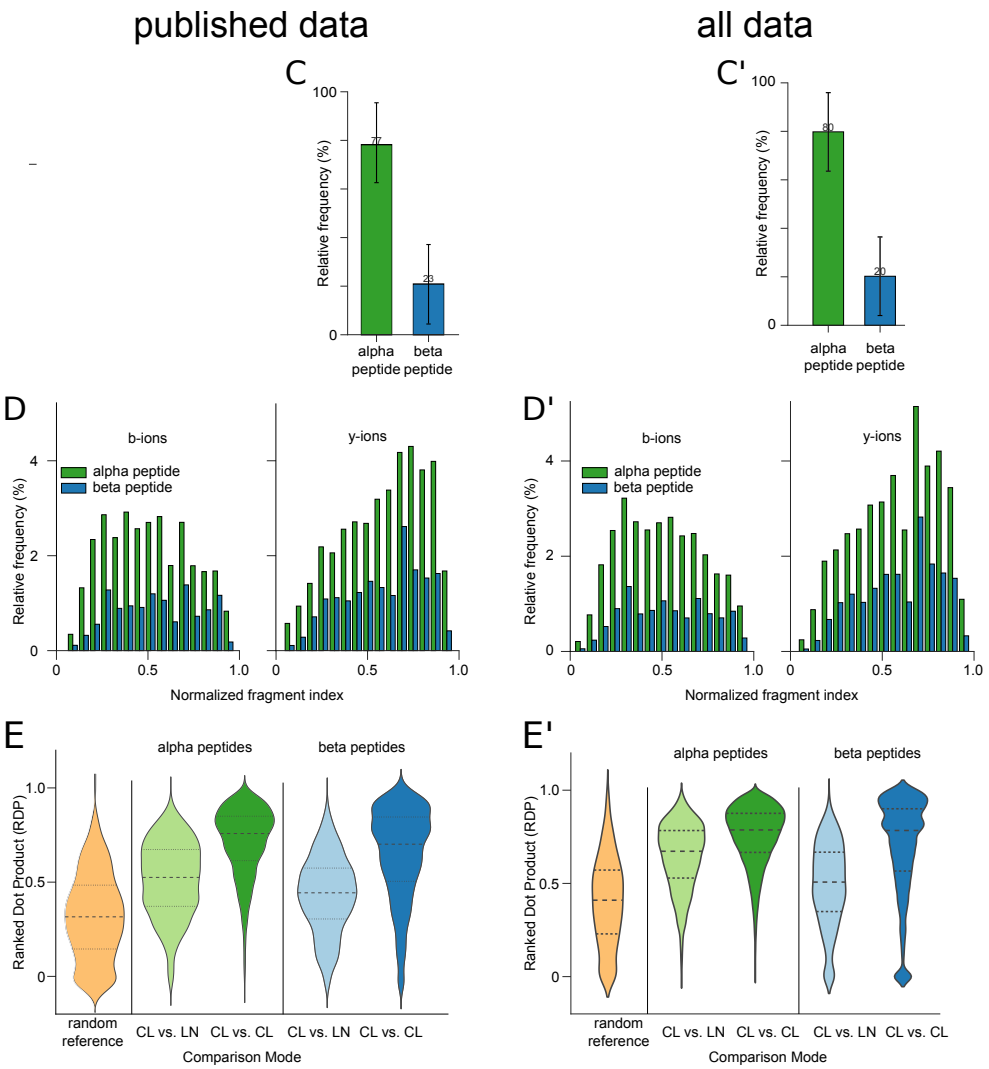
Supplementary Figure 9: Influence of different sample sizes on the results of Figure 1. The panel names refer to the same names in the manuscript. The left column panels (A-C) are the same plots as in the manuscript (published samples). The right column panels (A'-C') are one-to-one copies of the plots from the manuscript but with a larger sample size.

S10: Figure 2 comparison to a larger sample



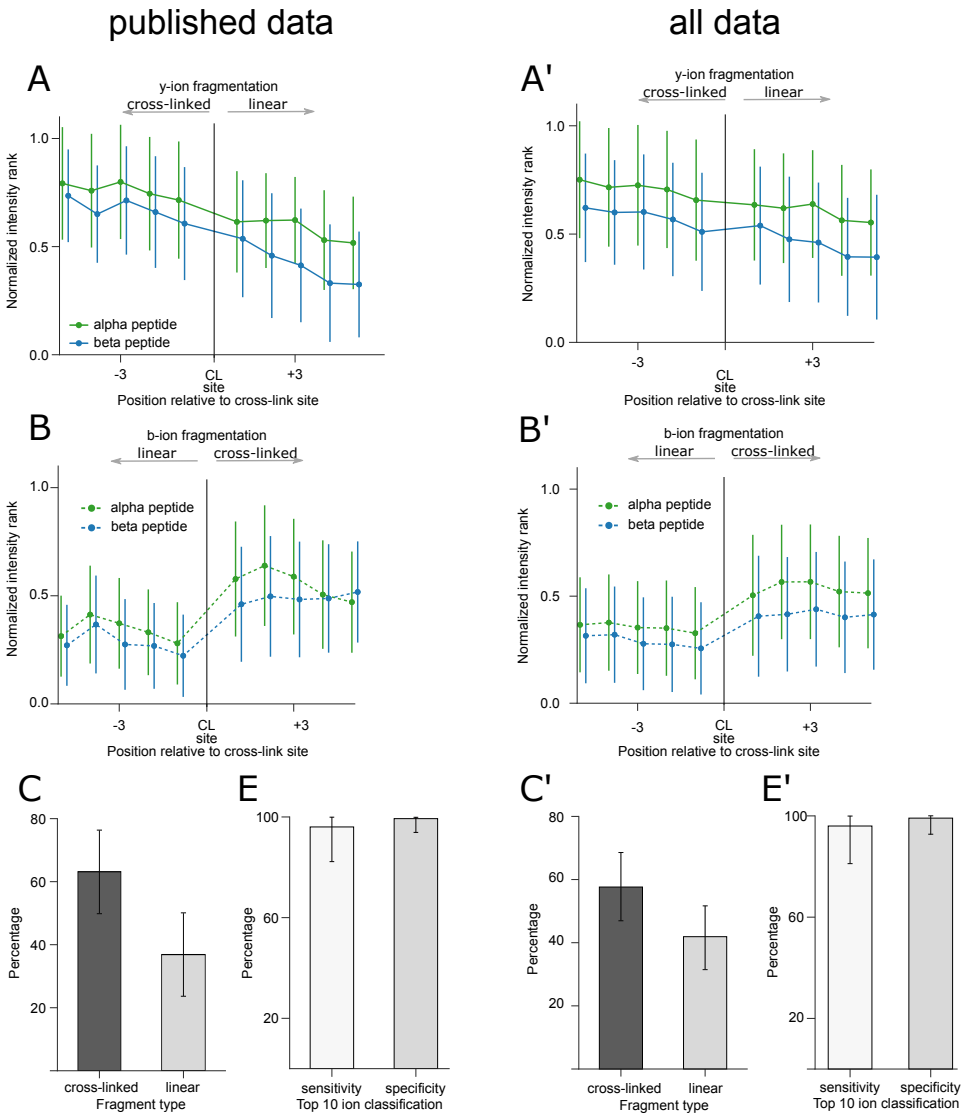
Supplementary Figure 10: Influence of different sample sizes on the results of Figure 2. The panel names refer to the same names in the manuscript. The left column panels (A-D) are the same plots as in the manuscript (published samples). The right column panels (A'-D') are one-to-one copies of the plots from the manuscript but with a larger sample size (all samples).

S11: Figure 3 comparison to a larger sample



Supplementary Figure 11: Influence of different sample sizes on the results of Figure 3. The panel names refer to the same names in the manuscript. The left column panels (C-E) are the same plots as in the manuscript (published samples). The right column panels (C'-E') are one-to-one copies of the plots from the manuscript but with a larger sample size (all samples).

S12: Figure 4 comparison to a larger sample



Supplementary Figure 12: Influence of different sample sizes on the results of Figure 4. The panel names refer to the same names in the manuscript. The left column panels (A, B, C, E) are the same plots as in the manuscript (published samples). The right column panels (A', B', C', E') are one-to-one copies of the plots from the manuscript but with a larger sample size (all samples).

Supporting Information: Optimized Fragmentation Regime for Diazirine Photo-Cross-Linked Peptides

Sven H. Giese[1, 2], Adam Belsom[2], and Juri Rappsilber[*1, 2]

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

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

List of Supporting Information

Note:

Raw files, reference FASTA file and the peptide-spectrum-match results (PSMs) are available via the PRIDE repository (PDX: PXD003737)[1].

Annotation description for Figure S4-S8

Peptides at the top are alpha peptides (in red); peptides at the bottom are beta peptides (in blue). Annotation: Loss ions (water, ammonia) are not annotated but the peaks are highlighted with light red/blue colors. Cross-linked fragments are annotated as +P. Mean(ppm error) and std(ppm error) refer to the measurement error on the fragment peak matching.

*juri.rappsilber@tu-berlin.de
[1]https://www.ebi.ac.uk/pride/archive/projects/PXD003737

Table S1: PSMs, links and peptide pairs overview

In cross-linking mass spectrometry the well-known false discovery rate (FDR) can be viewed at from different levels: First, the commonly known PSM level, i.e. with an 5% PSM FDR we expect 5% of the PSMs to be false positives. However, since the information we are after (spatial constraints) is only available on link level it is beneficial to compute the FDR on the link level. On the link level for a 5% FDR, 5% of the identified cross-links are expected to be false positives. Since multiple PSMs can lead to the same link it is often more sensible to compute the FDR on the link level. The details of the FDR for cross-links are discussed by Fischer *et al.* (submitted). Throughout the manuscript PSMs are evaluated at a 5% PSM FDR and links are evaluated on a 5% link FDR. For the special case where we look at peptide sequences we start with a 5% PSM FDR and then remove the information about cross-link site, charge state and the associated spectrum.

Table S1: Identification overview

Fragmentation	PSMs (5% PSM FDR)	Links (5% Link FDR)	Unique Links (% from 1,390)	Peptide Sequences (5% PSM FDR)
HCD	**958**	**446**	**46%**	**356**
CID	604	297	31%	205
EThcD	433	240	25%	169
ETD	311	202	21%	141
ETciD	296	205	21%	130

Note: PSMs and links were extracted using their respective 5% FDR level. A PSM is defined by the two peptides in a cross-link match, the charge state and the cross-link sites. A link is defined by the two residues that are cross-linked regardless of the peptide sequences.'Unique Links' refers to the number of unique links that the individual method contributes to the total number of unique links (1,390). For comparing the different peptide species that were identified the peptide sequences from the PSMs were extracted. Thus, omitting the charge state and the cross-link site information. Cells in bold refer to the maximum value in each column.

Figure S1: Evaluation of the identified cross-links against the crystal structure of HSA

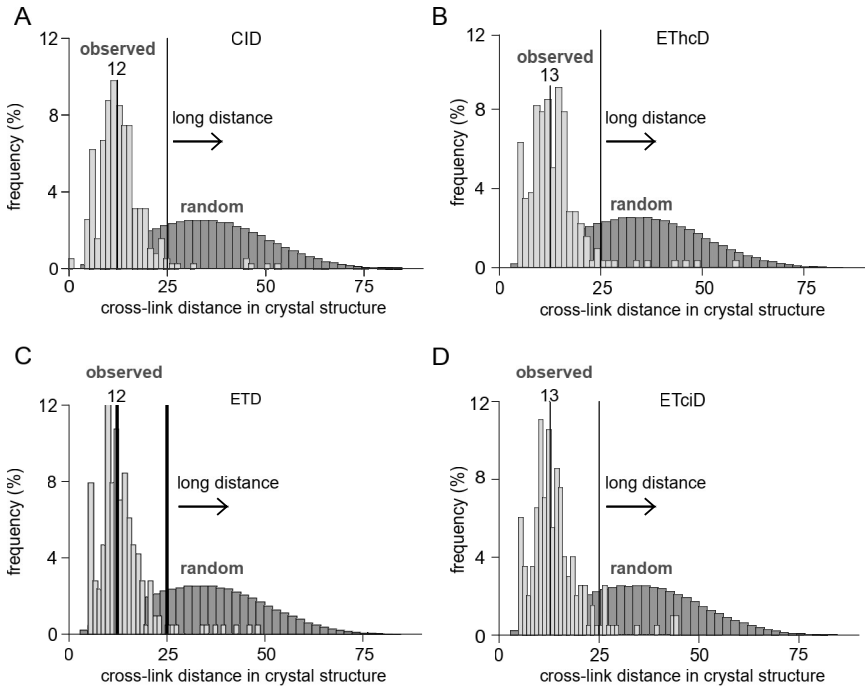


Figure S1: Evaluation of the identified cross-links against the crystal structure of HSA (A-D) . The light grey distribution reflects the distance measurement between identified residues in a cross-link mapped to the crystal structure. Distances were measured by computing the Euclidean distance between CA-atoms of the identified residues. Results for the four fragmentation techniques: CID, EThcD, ETciD and ETD are shown. The dark grey distribution reflects all pairwise combinations of cross-linkable residues in the crystal structure. The median for each of the light grey distributions is shown as vertical line in the distribution with the actual value on top. The black vertical line at 25 $\AA$ is used to classify cross-links as long-distance or not.

Table S2: Number of acquired MS2 spectra

Table S2: Number of MS2 spectra recorded per fragmentation scheme.

fragmentation scheme	number of MS2 spectra	Spectra increase	Identification Rate
HCD	109,004	39%	0.88
CID	99,356	26%	0.61
EThcD	83,093	6%	0.52
ETD	81,406	4%	0.38
ETciD	78,615	0%	0.38

Note: Number of MS2 spectra refers to the sum of all acquired MS2 spectra in the three replicates. Spectra increase is measured relative to the number of spectra acquired with ETciD and all other fragmentation schemes. Identification Rate refers to the fraction of unique PSMs (identified at 5% FDR) and total number of acquired spectra.

Figure S2: Decision surface with additional Information

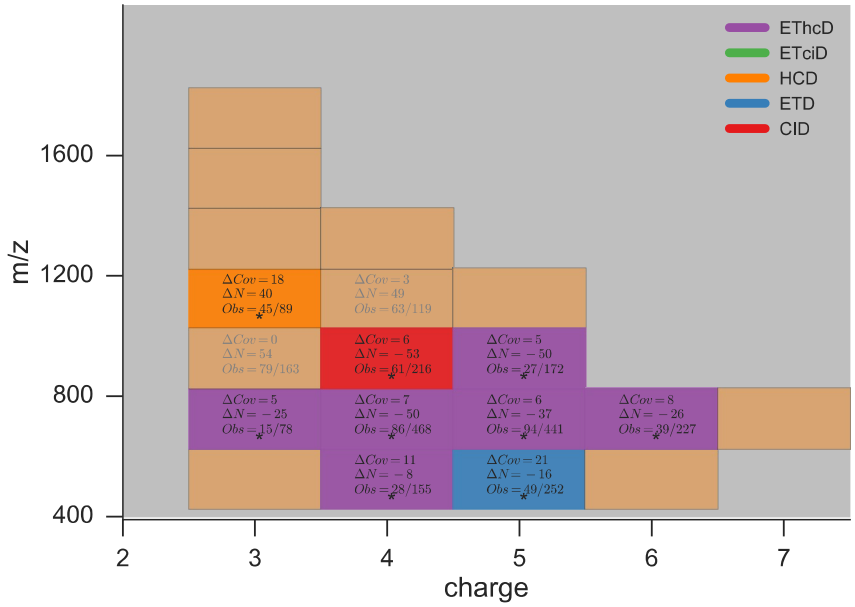


Figure S2: Pseudo-decision surface. This Figure refers to Figure 4D in the manuscript. In addition, information such as ΔN - the difference in the number of identifications, ΔC - the difference in coverage, between the best and second best method in that respective m/z - charge bin is annotated. *Obs* refers to fraction of identifications for the best method compared to all methods.

Figure S3: Absolute and relative number of identifications per m/z bin

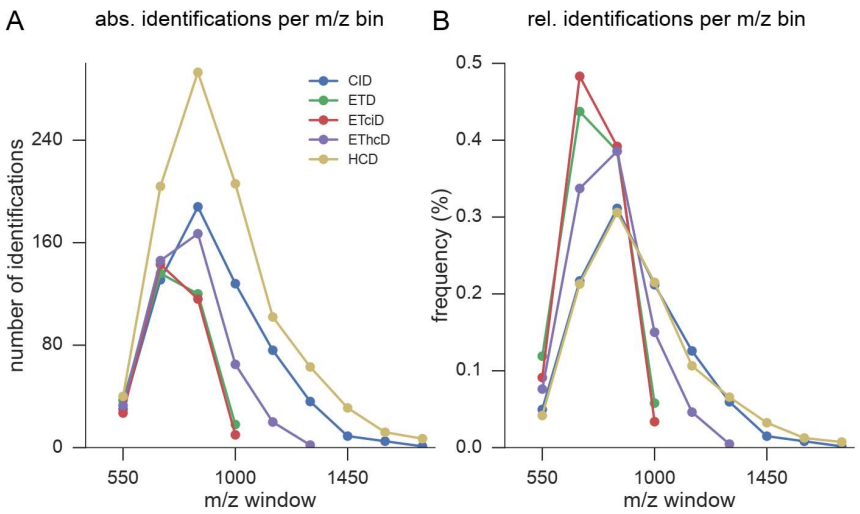


Figure S3: Absolute and relative number of identifications per m/z bin. (A) For each m/z bin the number of peptide-spectrum-matches is counted. For (B) each count is divided by the sum of PSMs over all m/z bins for each fragmentation method.

Figure S4: Spectra are available online.

Supporting Information: Noncovalently Associated Peptides Observed during Liquid Chromatography-Mass Spectrometry and Their Effect on Cross-Link Analyses

Sven H. Giese[1], Adam Belsom[1,2], Ludwig Sinn[1], Lutz Fischer[1], and Juri Rappsilber*[1,2]

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

[2]Wellcome Centre for Cell Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH93BF, United Kingdom

Contents

List of Figures

*juri.rappsilber@tu-berlin.de

S1 Visualization of cross-links and noncovalently associated peptides

Fig. S1 visually describes the different types of peptides relevant for the main manuscript. Importantly, only the cross-linked peptide (a) and the noncovalently associated peptide (c) have the same mass because of the loop-link in c). This mass ambiguity is the reason that noncovalently associated peptides can be misidentified as cross-links. More details on general cross-linking nomenclature can for example be found in *Rappsilber* [1]. Note that the two peptides in c) do not need to have the same sequence.

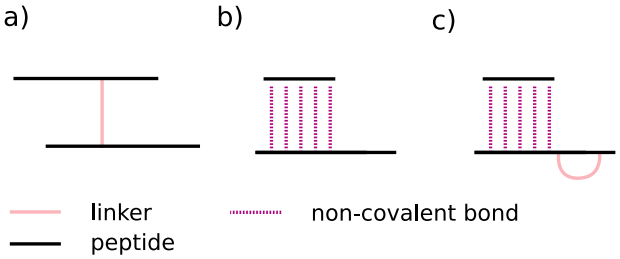


Figure S1: Visualization of different peptide definitions. (a) a typical cross-link between two peptides. (b) a non-covalent peptide association between two peptides. (c) same as (b) but one of the peptides is loop-linked, the mass of this species is the same as that of a cross-linked peptide.

S2 CLMS identifications assuming cleavability of SDA

In this section we used MeroX [2] to identify MS-cleavable cross-linker products from the Q Exactive (QE) data set. We used the same settings for MeroX (v. 1.6.6.6) as in [3]. The SDA reaction product is assumed to be cleavable when involving a carboxylic acid functional group [3]. The individual search results were combined using the MeroX Merger (v. 1.2) with the -P 5 setting to set the desired FDR cut-off to 5%. From the merged results we extracted the unique links and computed their distance in the crystal structure of HSA (PDB: 1AO6). For the QE data, 184 unique links were identified of which 160 could be mapped to the crystal structure. 38% (61 links) were long-distance links ($C_\alpha distance \geq 25$Å), while 62% (99 links) matched the distance constraint. For the Velos data, 34 unique links were identified of which 29 could be mapped to the crystal structure. 21% (6 links) were long-distance links, while 79% (23 links) matched the distance constraint. The results are consistent with the presented data in Figure 1 of the manuscript. The distance histogram for the QE data shows a very prominent enrichment of false positives exceeding the distance threshold. While in both cases the desired FDR is not met, we hypothesize that the Velos results are suffering from the low number of identified links. Therefore, reliable FDR estimation is hindered.

In general, Fig. S2a-b shows that MeroX is also able to identify the non-covalent peptide associations using a cleavable cross-linker search. However, it is difficult to judge how many of the identified cross-links below the 25Å cut-off are true cross-links (assuming cleavage of the cross-linker) and how many are non-covalent associations. Since the search itself is not aware of any distance constraint, an obvious assumption is that non-covalent associations should be distributed without preference below and above the distance cut-off. In contrast, true cross-links will be enriched below the distance cut-off. Visually projecting the number of long-distance links to the area below the distance cut-off indicates that a large portion of the within-distance links are in fact non-covalent associations.

In addition, we also analysed the retention times (RT) from linear peptides with SDA modifications (e.g. loop-linked or hydrolyzed cross-linker, see [4] for visualizations of the modifications). Interestingly, the RT of linear peptides is approximately increased by 24 minutes with a single sda-loop modification (Fig. S2c). Subsequently, the RT is almost doubled (42 minutes) when two loop-links were found in a peptide compared to the unmodified version. This information can be used to compare the RT of the linear peptides that were identified in a cross-link / non-covalent peptide association. We used the simplified assumption that identifications are true cross-links when the distance constraints were met and non-covalent association otherwise. In Fig. S2d, the RT difference between the two peptides in a cross-link / non-covalent association is shown. Initially, we tried to map the individual peptide sequences identified by MeroX to the linear (modified) peptide identifications from MaxQuant. For this one of the two peptides identified in a cross-link by MeroX was assumed to carry a loop-link modification. Under these assumptions only a small number of cross-linked peptides yield a RT for both peptides. The reason is that the individual peptides identified by MeroX were not identified in their loop-linked form in MaxQuant. For peptides that are cross-linked, the RT difference from the individual peptides should be randomly distributed. For peptides that are noncovalently associated, the RT difference from

the individual peptides should be closely distributed zero. Because MeroX does not search for loop-link modifications in the search for noncovalently associated peptides the RT difference that is introduced through this modification needs to be accounted for. Therefore, the expected RT difference for the individual peptides from non-covalent associations is on average 24 minutes. Indeed, the two RT difference distributions from cross-links and non-covalent associations look different and match the above described expectation (Fig. 3c). However, the large enrichment of within-distance links with a very small RT difference hints on these identifications being non-covalent peptide associations.

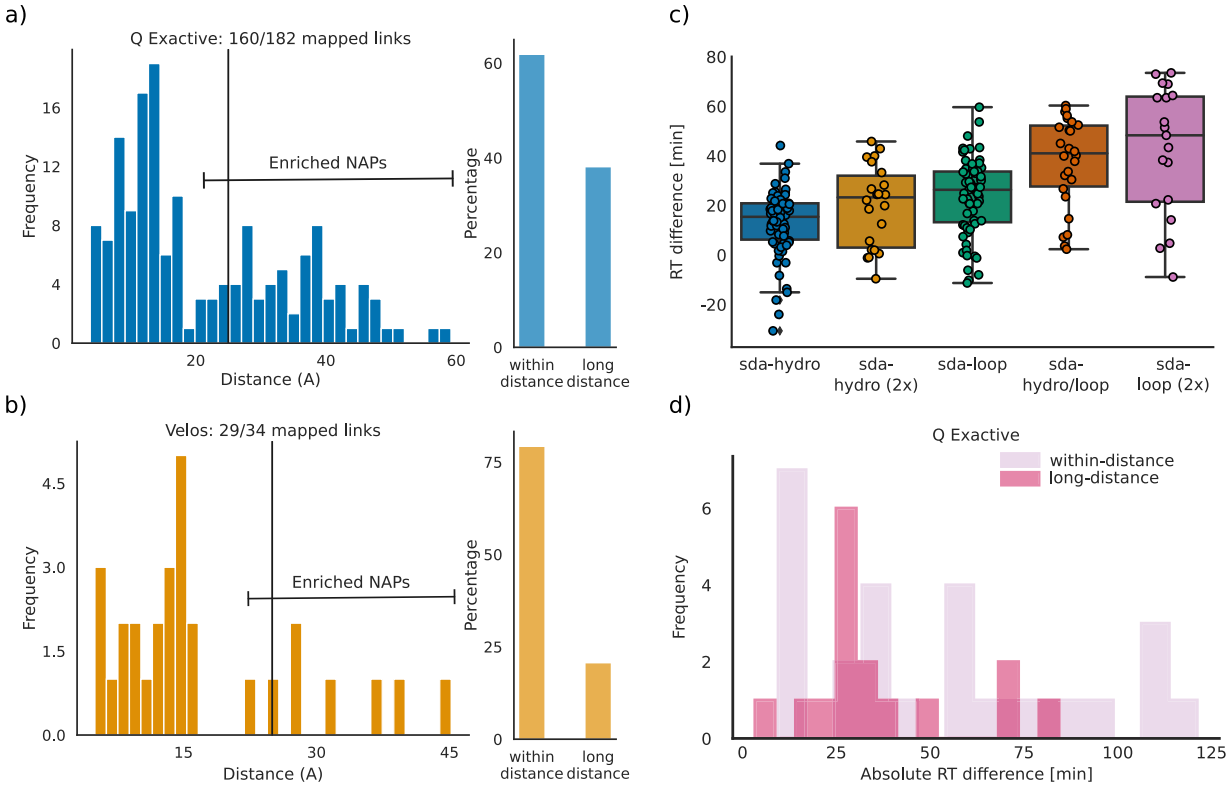


Figure S2: MeroX results and MaxQuant results. (a) Distance-histogram of unique links identified with MeroX for Q Exactive data. (b) Distance-histogram of unique links identified with MeroX for Velos data. (c) Retention time difference of unmodified and cross-linker modified linear peptides identified with MaxQuant. (d) RT mapping of PSMs from a) to linear identifications (MaxQuant search) without RT adjustment for modifications. *Note*: RT - retention time, NAP - non-covalent association, sda-hydro and sda-loop refer to modified cross-linkers [4].

S3 Flow Rate Analysis on Q Exactive High-field

To further investigate the effect of different flow rates on the formation of non-covalent associations we acquired the protein mix (non-cross-linked sample) on the Q Exactive High-field with three flow rates (in triplicates): 0.2 $\frac{\mu L}{min}$, 0.25 $\frac{\mu L}{min}$ and 0.3 $\frac{\mu L}{min}$. The differences in the number of identifications were only small (Fig. S3a) and comparable to the results from the main text (24 PSMs with IS-CID 0). To achieve the desired FDR cut-off of 5% the results were cut after the first decoy hit (S3b).

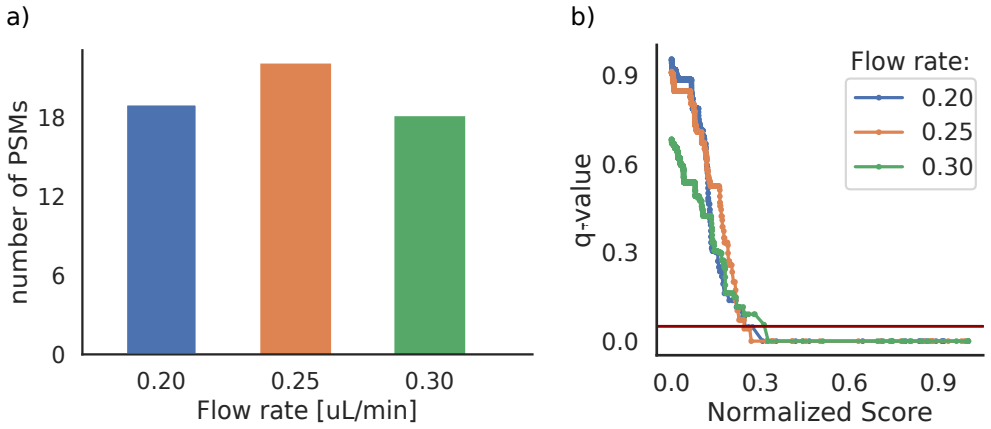


Figure S3: Flow rate analysis. a) shows the number of PSMS passing a 5% FDR cut-off (or until the first decoy hit if the cut-off is exceeded). b) shows the q-value trend with decreasing scores. The horizontal red line marks the 5% cut-off. The search engine score was normalized by division through the respective maxima.

S4 Falsely identified cross-link suggesting homo-dimerization

The spectrum in Fig. S4 shows an example of a cross-link that can falsely lead to the assumption of homo-dimerization.

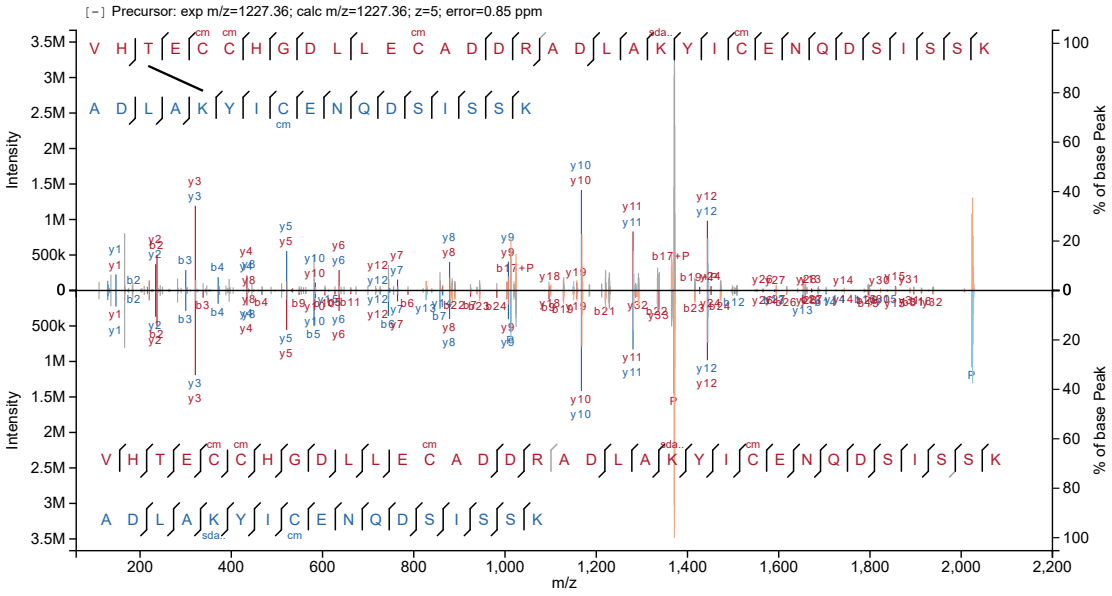


Figure S4: Alternative explanation for a cross-link that would suggest homodimerization. Upper panel, annotation from non-covalent search. Lower panel, annotation from cross-link search. Raw file: V127_J; scan: 34926

References

[1] J. Rappsilber, "The beginning of a beautiful friendship: Cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes," *J. Struct. Biol.*, vol. 173, no. 3, pp. 530–540, mar 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21029779

[2] M. Götze, J. Pettelkau, R. Fritzsche, C. H. Ihling, M. Schäfer, and A. Sinz, "Automated assignment of MS/MS cleavable cross-links in protein 3D-structure analysis." *J. Am. Soc. Mass Spectrom.*, vol. 26, no. 1, pp. 83–97, jan 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25261217

[3] C. Iacobucci, M. Götze, C. Piotrowski, C. Arlt, A. Rehkamp, C. Ihling, C. Hage, and A. Sinz, "Carboxyl-Photo-Reactive MS-Cleavable Cross-Linkers: Unveiling a Hidden Aspect of Diazirine-Based Reagents," *Anal. Chem.*, vol. 90, no. 4, pp. 2805–2809, feb 2018. [Online]. Available: http://pubs.acs.org/doi/10.1021/acs.analchem.7b04915

[4] S. H. Giese, A. Belsom, and J. Rappsilber, "Optimized fragmentation regime for diazirine photo-cross-linked peptides," *Anal. Chem.*, vol. 88, no. 16, pp. 8239–8247, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27454319

Supporting Information: Peptide Retention in Hydrophilic Strong Anion Exchange Chromatography is Driven by Charged and Aromatic Residues

Sven H. Giese[1], Yasushi Ishihama[2], and Juri Rappsilber*[1,2,3]

[1]Bioanalytics, Institute of Biotechnology, Technische Universität Berlin, 13355 Berlin, Germany
[2]Graduate School of Pharmaceutical Sciences, Kyoto University, Kyoto 606-8501, Japan
[3]Wellcome Centre for Cell Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3BF, United Kingdom

Contents

List of Figures

List of Tables

*juri.rappsilber@tu-berlin.de

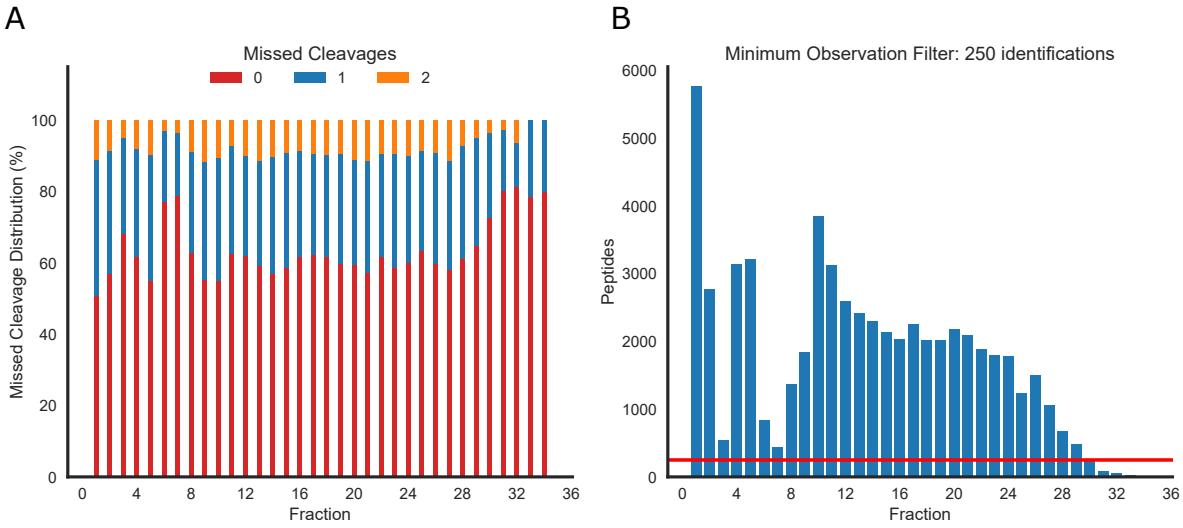


Figure S1: Data overview. (A) Missed cleavage distribution. Stacked bar charts show the number of peptides with 0, 1 or 2 missed cleavages per fraction. (B) Number of peptide identifications per fraction. The horizontal line was set as cut-off - all fractions with fewer than 300 identifications were disregarded from the analysis. A total number of 59,723 non-redundant peptides were analysed before using the cut-off.

S1 Effect Size and Retention Time Influence Differences of the Charged Amino Acids

As established in the main text the effect size of the charged amino acids is very similar. However, the distributions of peptides with 0-5 D or E residues are clearly shifted as shown in Fig. 1 of the manuscript. Table S1 shows the average mean increase of the fraction number per peptide population with 0-5 D/E counts. On average, a single D/E residue in the peptide sequence will shift the peptide 3 fractions. Since the effect size of D/E is very similar (Fig. S2 B) we assume that the estimate holds for either D or E residues. On the other hand the difference between K and R residues is more pronounced (Fig. S2 A).

Table S1: Effect of D and E residues to the retention time shift.

DE count	Mean Fraction	Difference to last Fraction
0	2.61	0
1	5.89	3.28
2	10.73	4.84
3	14.89	4.16
4	17.89	3
5	20.31	2.42

Note: The mean fraction was computed by first filtering all peptide identifications to sequences with 0-5 D or E residues. For each of the five classes the mean fraction was then computed.

S2 Non-charged Amino Acid Contributions to the Retention Time

In the main text we classified the remaining amino acids as 'retaining', 'eluting' and 'other'. This classification is mainly based on investigating an isolated subset of peptides with D/E residue count of 2 and K/R residue count of 1. This subset is then used to visually and statistically infer the influence of the remaining amino acids. As shown in Fig. S3 the number of observations of DE2, KR1 peptides is still very high and distributed over 12 fractions. Based on these peptides we computed the average amino acid composition in each fraction and performed linear regression analysis with the composition as dependent variable and the fraction as target variable. Effectively, modeling the increase or decrease in the sequence composition for all 16 remaining amino acids. The magnitude of the slope can be considered as correlation between the occurrences of amino acids and

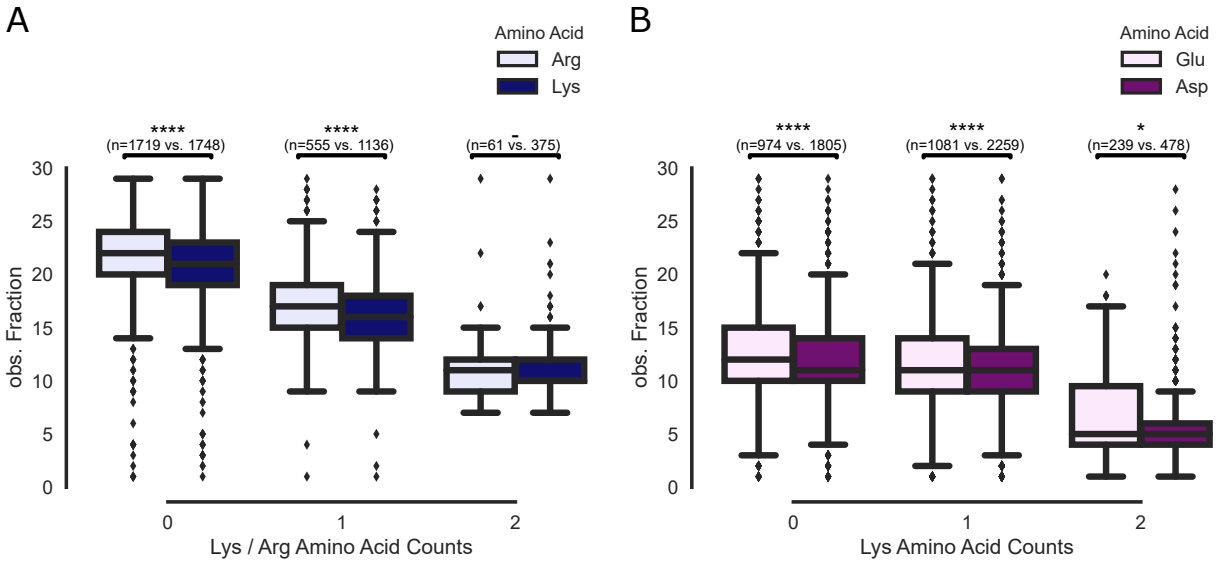


Figure S2: Observed fractions based on a peptide sequence filters. (A) Effect size of Lys and Arg residues. To compare the elution strength of Arg and Lys first all peptide identifications were filtered to only include peptides with and summed D/E residue count of 4. Then the fractions of peptides with 1, 2 and 3 K/R residues were extracted and compared. (B) Effect size of Glu and Asp residues. To compare the retaining strength of Glu and Asp first all peptide identifications were filtered to only include peptides with exactly two D or two E residues. For these two sub-populations then the peptides with 1, 2 and 3 K residues were compared. Significance tests were performed using the Mann-Whitney-U-Test.

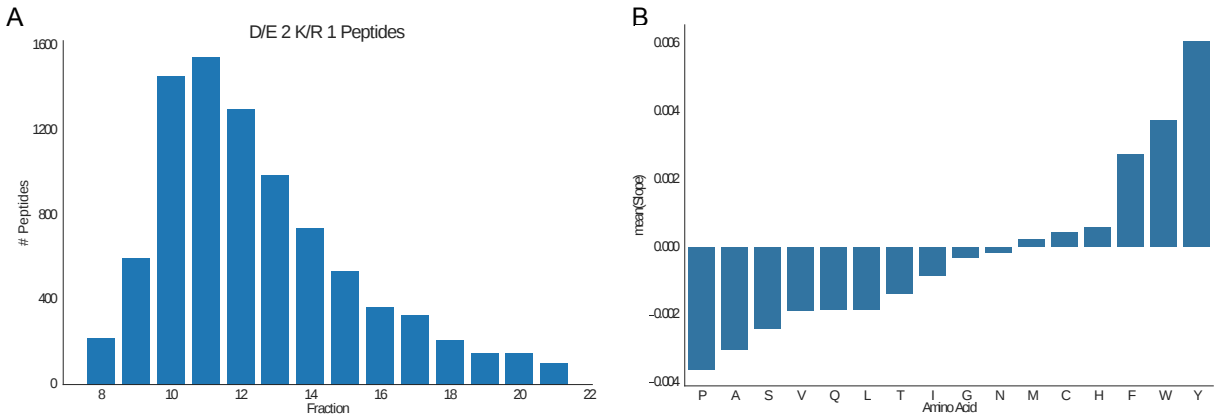


Figure S3: Sub-population of peptides with an D/E 2 and K/R 1 count. (A) The distribution of peptide occurrences is shown depending on the observed fraction. Fractions below 8 and higher than 21 all contained less than 1% (92) of the total number of observations (9,199) and were removed for further analysis. (B) Based on the average composition of the peptides from (A) 20 linear regression models (for each amino acid one) were fitted on the target variable (fraction number) and the dependent variable (average amino acid sequence composition). The slopes of the regression model are shown as bars. Amino acids that have an retaining effect are expected to have a positive slope, amino acids that have an eluting effect are expected to have a negative slope.

a shifted retention time. For large positive slopes, we expected the amino acids to have an retaining effect on the retention time. For large negative slopes, we expected the amino acids to have an eluting effect, see Fig. S3B. The linear regression model was also used to perform a significant test on the slope of the fitted model: with H_0 assuming that the slope is equal to zero and H_1 assuming that the slope is not equal to zero. The test results and fitted lines are visualized in Fig. S4. Based on this we broadly classified the remaining amino acids into retaining contributions (F, W, Y), eluting contributions (P, A, S, V, Q, T) and non-clear or other contributions (L, I, G, N, M, C, H).

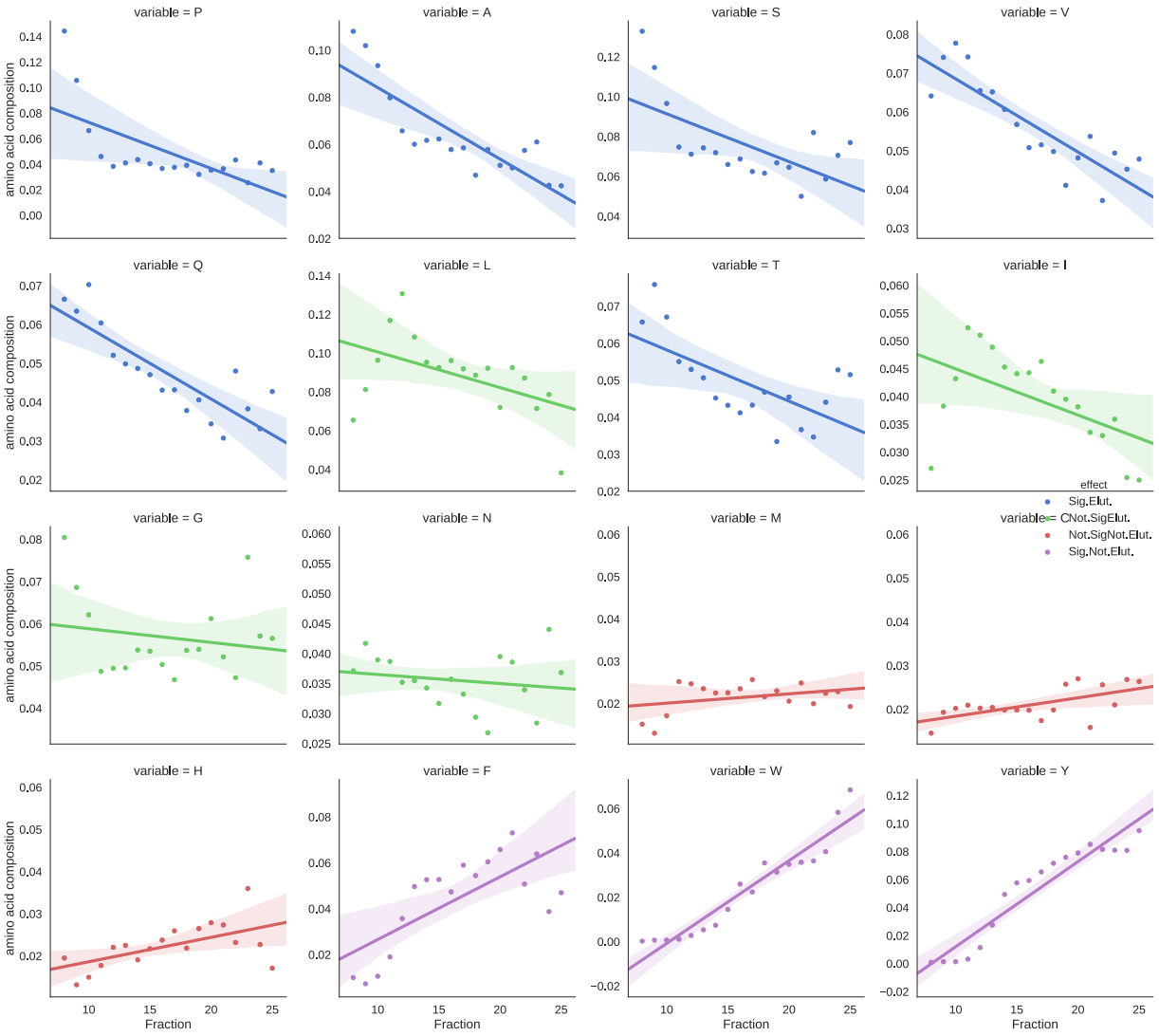


Figure S4: Classification of non-charged amino acid effects on the retention time based on linear regression. As described in Supplementary Fig. S3 linear regression models were fit to the sequence composition data from peptides with an D/E 2 and K/R 1 count. In addition, a simple test with H_0: the slope of the regression model is equal to zero was performed using SciPy. The aromatics F, W, and Y yield significant results and have a potentially large retaining effect. The amino acids P, A, S, V, Q and T also show a significant test results after Benjamini-Hochberg correction [1], but for having an eluting effect. For the amino acids L, I, G,M, N and H the slope of the regression model was not significantly different from zero and were thus classified as ambiguous ('other').

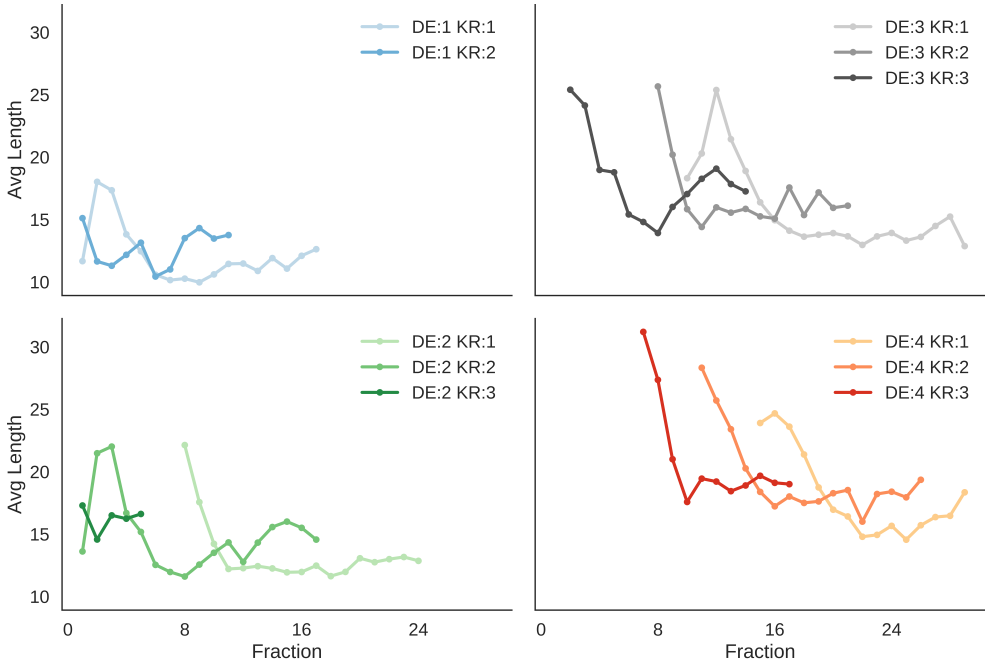


Figure S5: Peptide length influence on the retention time. In each panel a subset of peptides is first extracted (e.g. DE:1 KR:1 first filters all peptides with exactly one D or E residue and exactly one K or R residue). A minimum number of observations of 300 peptides was required per category. The Avg Length represents the mean peptide length of all peptides in a selected fraction.

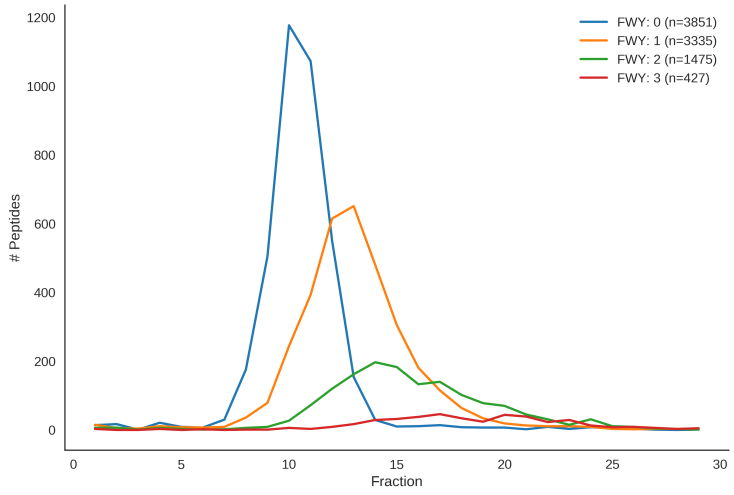


Figure S6: Peptide counts across all fractions with 0,1,2 or 3 WYF residues. In addition to the applied FYW filter all peptides have 2 D/E residues and 1 K/R residue.

S3 Machine Learning - Training, Prediction and Evaluation

Overview

We are interested in learning and predicting the interaction of peptides with the hSAX column - based on the peptide sequence we want to be able to predict when the peptide will elute. Initially, we used a set of classification algorithms in our pre-experiments. The selection of regression methods includes: simple linear regression including the length correction parameter (lcp) with only the 20 amino acids as features ('Pyteomics') [3], a linear regression model with all designed features (Supplementary Table S2), ridge regression, lasso regression, support vector machine regression (SVR) and random forest regression. The selection of classification algorithms includes: feedforward neural network (FNN, Keras implementation with the Theano backend), logistic

regression, random forest, gradient boosting (python package XGBoost[1] [2]) , a support vector machine (SVM) and ordinal logistic regression (python package MORD[2] [7]). Except for Pyteomics, the FNN, MORD and XGBoost the scikit-learn[3] [6] implementations were used.

Input Features

An essential part of classical machine learning algorithms is the engineering of features. Based on initial observations and by investigating the literature we came up with 218 features to summarize the properties of a peptide that might govern retention. These features are summarized in Table S2. Similar to ELUDE and SSRCalc we used hydrophobicity features, consecutive occurrences of amino acids [5] and position specific features for the 20 amino acids [4].

Hyper-parameter Optimization

The above mentioned machine learning algorithms all require a fine tuning of their parameters to achieve the best possible performance (hyper-parameter optimization). Our workflow for optimization, testing and validation the best parameters was as follows: (1) grid search for optimal hyper-parameters with 5-fold cross-validation (CV), (2) selection of the best set of parameters for each classifier based on the achieved accuracy on the test data and (3) validating the best performing classifier on a hold-out validation set that was never used for training. Table S3 gives and overview of the grid search for hyper-parameter optimization. Table S4 summarizes the results for each classifier with the best set of parameters. The best performing classifier was a a feedforward neural network implementation with an CV accuracy of $70 \pm 0.81\%$ (mean $\pm$ standard error of the mean). The linear regression models achieved the lowest accuracy on the test sets with $19\% \pm 0.002$. Pyteomics and the corresponding linear model (lcp) with a minimal set of features were not included in the grid search.

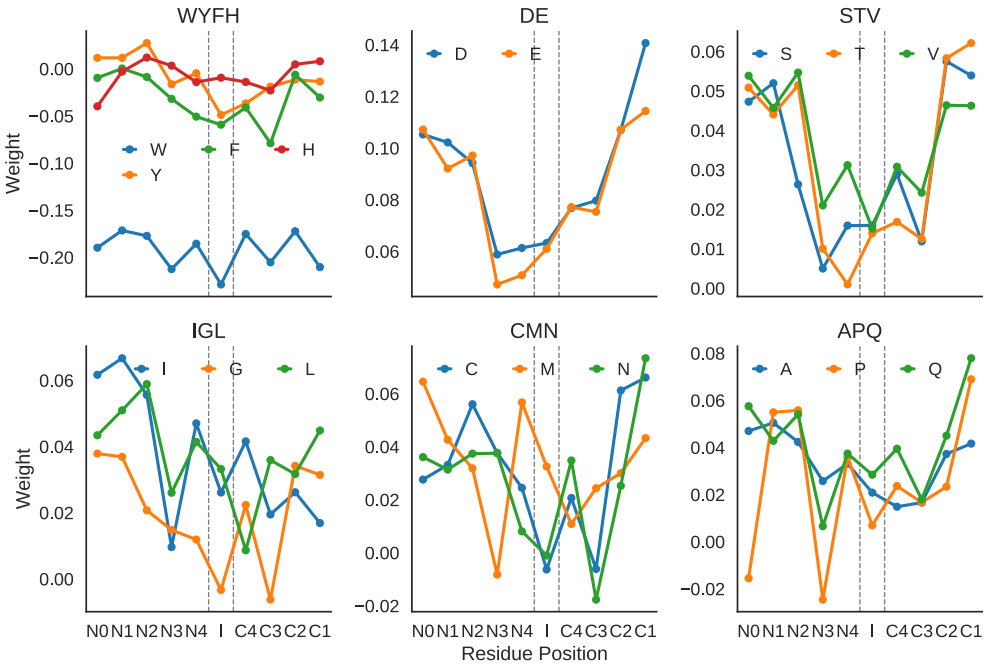


Figure S7: Mean weights from the input layer to the first hidden layer in the FNN. The X-axes indicates the position of an residue in the peptide sequence. The weights are derived from training the FNN on the complete training data (Accuracy: 0.74, Correlation: 0.95). Abbreviations: N - peptide N-terminal, C - peptide C-terminal, I - internal. The numbers indicate the distance to the respective termini.

[1] https://github.com/dmlc/xgboost
[2] https://pythonhosted.org/mord/
[3] http://scikit-learn.org/stable/

Table S2: Extracted Features and their description.

Feature	Description	Total Features
AAcount	Amino acid counts of all 20 amino acids	20
N[AA]1-5	Amino acid counts encoding the n-terminal positions from 1-5.	100
C[AA]2-5	Amino acid counts encoding the c-terminal positions from 2-5. N-term was excluded as mostly R/K was observed.	80
CtermK/R	Indicator if Lysine is the C-Terminal Residue	2
Patterns	Counts the number of coherent amino acid patterns in the peptide sequence of different classes: acidic, basic, aromatics, mixed (acidic+basic) patterns. For example DD, KR, WW, DK.	4
Structural Features	Percentage of amino acids from the sequence that are preferably in the following secondary structure elements: Helix: V, I, Y, F, W, L. Turn: N, P, G, S. Sheet: E, M, A, L.	3
Gravy	Gravy according to Kyte and Doolittle.	1
pI	Isoelectric point of the peptide sequence.	1
loglength	Natural logarithm of the peptide length.	1
Netcharge	Defined as sum of the acidic residues (-1 each), basic residues (+1) and the aromatics F (0.3), W (0.8) and Y (0.6) in a peptide sequence.	1
N-/C-Term distance	Shortest distance of E/D to the C-term and shortest distance of K/R to the N-term.	2
TurnIndicator	Average distance between Proline residues in the sequence.	1
Sandwich	Aromatic patterns that are separated in sequence by one amino acid, e.g. WXY.	1
Aromaticity	Percentage of amino acids belonging to WFY.	1
	Total number of features	218

Note: The count features were scaled with a length correction parameter (lcp).

Table S3: Initial parameter grid for hyper-parameter optimization

Classifier	Parameter Grid
ORL IT	'alpha': [0.1, 0.3, 0.5, 0.7, 0.9]
Lasso	'fit_intercept': [True, False], 'alpha': [0.1, 0.3, 0.5, 0.7, 1], 'normalize': [True, False]
LinearRegression	'fit_intercept': [True, False], 'normalize': [True, False]
Ridge	'fit_intercept': [True, False], 'alpha': [0.1, 0.3, 0.5, 0.7, 0.9], 'normalize': [True, False]
SVM	['C': [0.1, 1, 10], 'kernel': ['linear'], 'class_weight': [None, 'balanced'], 'C': [0.1, 1, 10], 'gamma': [0.001, 0.0001], 'kernel': ['rbf'], 'class_weight': [None, 'balanced']]
OLR AT	'alpha': [0.1, 0.3, 0.5, 0.7, 0.9]
RandomForestClassifier	'n_jobs': [20], 'n_estimators': [100, 500], 'max_features': ('log2', 'auto'), 'max_depth': (None, 4, 7), 'min_samples_split': (2, 15)
XGB	'reg_alpha': [0.01, 0.5, 1], 'n_estimators': [300, 500], 'gamma': [0, 0.1, 1], 'max_depth': [3, 5, 9], 'reg_lambda': [0.01, 0.5, 1], 'nthread': [20], 'learning_rate': [0.1, 0.05]
RandomForestRegressor	'n_jobs': [20], 'n_estimators': [100, 500], 'max_features': ('log2', 'auto'), 'max_depth': (None, 5, 15), 'min_samples_split': (2, 15)
LogisticRegression	'C': [0.01, 0.1, 1, 10], 'multi_class': ['ovr', 'multinomial'], 'n_jobs': [20], 'solver': ['newton-cg'], 'class_weight': [None, 'balanced']

Note: The parameter grid was searched exhaustively with all combinations. The definition of each parameter is available via the documentations of scikit-learn, MORD and XGBoost. The neural network architecture was optimized manually.

Table S4: Best Results after hyper-parameter optimization with 5-fold cross-validation.

Classifier	Best Parameters	Train Accuracy (%)	Test Accuracy (%)
FNN	'layer': 4, 'neurons': [50, 40, 35, 29], 'activation':['relu', 'tanh', 'relu', 'softmax'], 'batch_size':512, 'epochs': 100	79 ± 1.27	70 ± 0.81
SVC	'class_weight': None, 'C': 10, 'kernel': 'linear'	64 ± 0.17	53 ± 0.19
SVR	'C': 10, 'kernel': 'rbf', 'gamma': 'auto', 'epsilon': 0.1	52 ± 0.06	50 ± 0.26
XGBClassifier	'n_estimators': 300, 'learning_rate': 0.1, 'reg_lambda': 0.01, 'reg_alpha': 1, 'max_depth': 9, 'nthread': 25, 'gamma': 0.1	100 ± 0.0	47 ± 0.32
XGBRegressor	'n_estimators': 300, 'learning_rate': 0.1, 'reg_lambda': 0.01, 'reg_alpha': 0.01, 'max_depth': 9, 'nthread': 25, 'gamma': 0.1	67 ± 0.17	46 ± 0.31
RF-Classifier	'max_features': 'auto', 'n_jobs': 20, 'n_estimators': 500, 'min_samples_split': 2, 'max_depth': None	100 ± 0.0	43 ± 0.33
LogisticAT	'alpha': 0.5	43 ± 0.14	43 ± 0.18
RF-Regressor	'max_features': 'auto', 'n_jobs': 20, 'n_estimators': 500, 'min_samples_split': 2, 'max_depth': None	77 ± 0.04	42 ± 0.19
LogisticRegression	'solver': 'newton-cg', 'multi_class': 'multinomial', 'C': 10, 'class_weight': None, 'n_jobs': 20	48 ± 0.07	40 ± 0.2
LinearRegression	'fit_intercept': True, 'normalize': False	19 ± 0.15	19 ± 0.36
Ridge	'alpha': 0.1, 'normalize': False, 'fit_intercept': True	19 ± 0.15	19 ± 0.34
Lasso	'alpha': 0.1, 'normalize': False, 'fit_intercept': True	14 ± 0.07	14 ± 0.06

Note: The grid search results are based on 5-fold cross-validation and sorted after the test accuracy in descending order. Values in the accuracy column represent the mean and standard error of the mean from the CV. A full explanation of the parameters is available through the scikit-learn documentation. Abbreviations: SVC - Support Vector machine Classification, OLR - Ordinal Logistic Regression, AT - All-Threshold, IT - Immediate-Threshold, RF - Random Forest, FNN - Feedforward Neural Network.

References

[1] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing, 1995.

[2] Tianqi Chen and Carlos Guestrin. XGBoost : Reliable Large-scale Tree Boosting System. *arXiv*, pages 1–6, 2016.

[3] Anton A. Goloborodko, Lev I. Levitsky, Mark V. Ivanov, and Mikhail V. Gorshkov. Pyteomics - a Python Framework for Exploratory Data Analysis and Rapid Software Prototyping in Proteomics. *J. Am. Soc. Mass Spectrom.*, 24(2):301–304, feb 2013.

[4] Oleg V. Krokhin. Sequence-specific retention calculator. Algorithm for peptide retention prediction in ion-pair RP-HPLC: application to 300- and 100-A pore size C18 sorbents. *Anal. Chem.*, 78(22):7785–95, nov 2006.

[5] Luminita Moruz, An Staes, Joseph M. Foster, Maria Hatzou, Evy Timmerman, Lennart Martens, and Lukas Käll. Chromatographic retention time prediction for posttranslationally modified peptides. *Proteomics*, 12(8):1151–9, apr 2012.

[6] Fabian Pedregosa and G Varoquaux. *Scikit-learn: Machine learning in Python*, volume 12. 2011.

[7] Fabian Pedregosa-Izquierdo. *Feature extraction and supervised learning on fMRI : from practice to theory.* Theses, Universit{é} Pierre et Marie Curie - Paris VI, 2015.

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

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

[1]Bioanalytics, Institute of Biotechnology, Technische Universität Berlin, 13355 Berlin, Germany
[2]Data Analytics and Computational Statistics, Hasso Plattner Institute for Digital Engineering
[3]Digital Engineering Faculty, University of Potsdam
[4]Wellcome Centre for Cell Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3BF, United Kingdom

Contents

List of Figures

List of Tables

*authors contributed equally
†corresponding author: juri.rappsilber@tu-berlin.de

Supplementary Note 1: xiRT: Multi-task Retention Time Prediction using Neural Networks

Overview

The schematic architecture of the xiRT was presented in Figure 1 of the manuscript, while Supplementary Figure 1 shows a more detailed view (exemplary configuration). Here, we want to give more details about the individual layers. The input layer dimension is dynamically defined by the longest peptide that was identified in the set of PSMs/CSMs. In the example in Supplementary Figure 1 this was set to 59. Subsequently, the input is passed to a predefined embedding layer in TensorFlow. The embedding layer finds a continuous vector representation from a list of positive integers. A hyper-parameter for the network is the length of the embedding vector, here set to 50.

Siamese Architecture

The heart of the xiRT network is a recurrent layer where we either used a Gated Recurrent Unit (GRU-) [1] or a Long short-term memory (LSTM)-layer [2]. These layers are especially suited for the handling sequential data, e.g. language data or peptide sequences. They are available as GPU and CPU implementations in TensorFlow and can thus be used interchangeably within xiRT. The central assumption for recurrent layers is that the order of the input (here amino acids) plays a pivotal role in the prediction process [3]. By optionally applying a bidirectional GRU/LSTM layer, the input sequence is handled forward and backward. To speed up the training process, the activations are further batch-normalized to $\mu = 0, \sigma = 1$. The above-described parts of the network are designed in a Siamese fashion. That means that two input sequences (i.e. the individual peptides in a crosslink) are passed to their custom inputs. However, these layers process the input in the same manner since they share the same weights. The combination of the outputs from the Siamese network can be handled in multiple ways. In Supplementary Figure 1 an *Add-layer* was used, which simply adds the two inputs element-wise. For the retention time prediction of linear peptides there is only a single input and thus no Siamese or additive layers are necessary.

Task Specific Layers

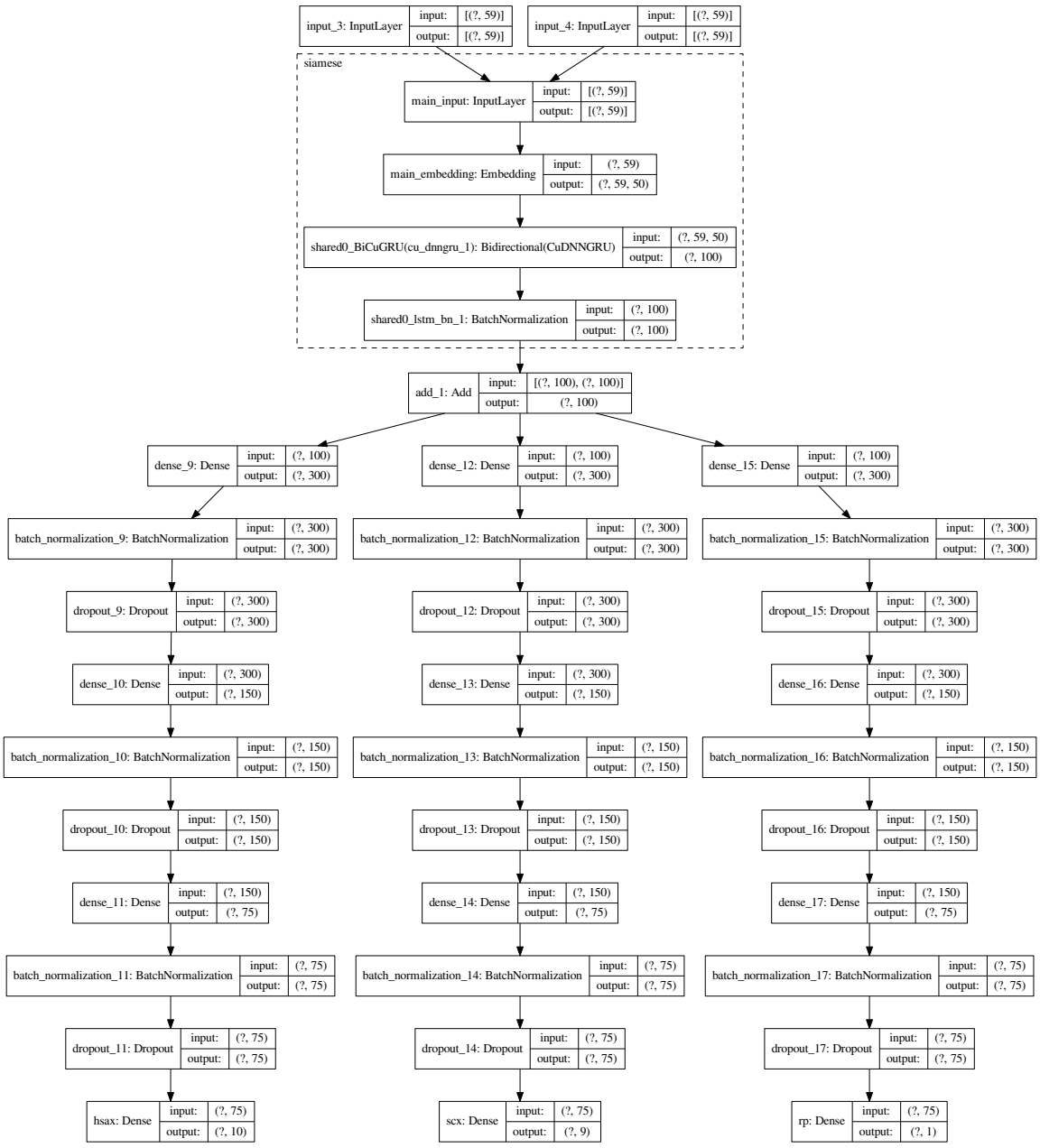
The architecture described above is also shared between the different prediction tasks. In this manuscript, we developed a multi-task network that predicts peptide retention behaviour during SCX, hSAX and RP chromatography. For this, the individual task-networks were designed in a symmetric fashion. They are defined by a sequence of layers with: $layer_i = Dropout(BatchNormalization(Dense(x)))$. Per default we used $i = 3$ and a pyramid-like structure for the dense layers with $n_{neurons} = [300, 150, 75]$. The default dropout-rate was set to 0.1 for all dense layers. Moderate kernel regularization (l2, $\lambda = 0.001$) was also used.

For each task, a custom prediction layer and a loss function are defined. The two employed fractionation techniques SCX and hSAX are handled as ordinal regression problems in which sigmoid activations were used and binary cross-entropy as loss. For the RP, we used a linear activation function and the mean squared error as loss function. Note that the handling of data from fractionation also allows to treat the problem as classification or as regression task and thus the use of softmax or linear activation functions are possible (also configurable in xiRT). The total loss is computed as weighted sum of the three individual losses, e.g. $loss_{total} = w_{fractionation} * (loss_{SCX} + loss_{hSAX}) + loss_{RP}$. Using *Adam* (Adaptive Moment Estimation) as optimizer, the learning rate was fixed to 0.001 during development and optimization on linear data. After optimization for crosslink data a higher learning rate (0.01) achieved faster convergence with similar accuracy together with a batch-size of 256 and was therefore chosen as default value in xiRT.

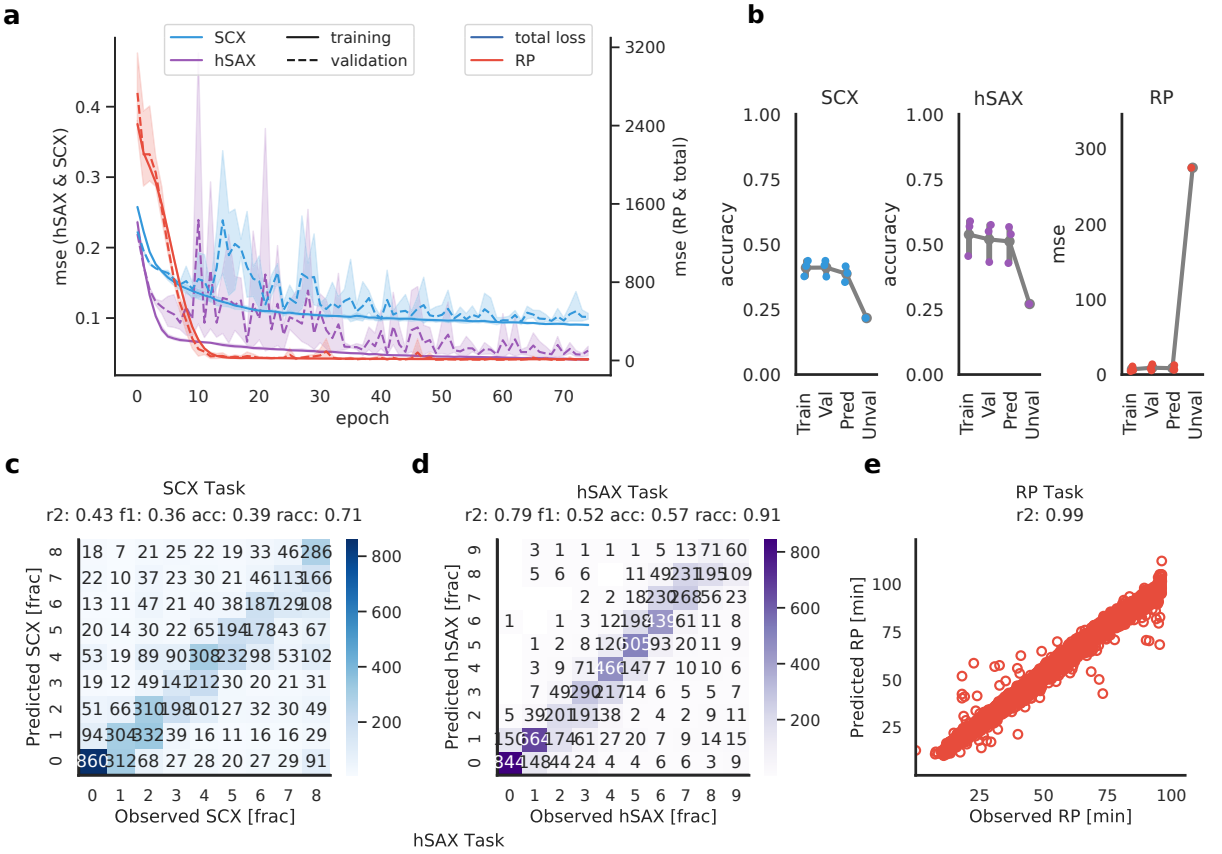
Implementation Details

xiRT is implemented in the popular deep learning framework TensorFlow 2.0 [4]. All training and prediction scripts were run on a TITAN X (Pascal) with 12.8 GB of memory. The usage of a dedicated GPU allows to use optimized recurrent layers in TensorFlow. These layers have a "CuDNN"-prefix, e.g. CuDNNGRU. CuDNNLSTM. Our implementation can also be used on systems without GPUs, at the cost of higher run time. TensorFlow also allows the usage of so-called callbacks. The most important callbacks in our implementation are 1) *ReduceLRonPlateau*, 2) *EarlyStopping* and 3) *ModelCheckpoint*. The 1) callback is used to reduce the learning rate by a factor of 0.5 when the performance has not improved in 15 epochs by a minimum delta of 1e-4.

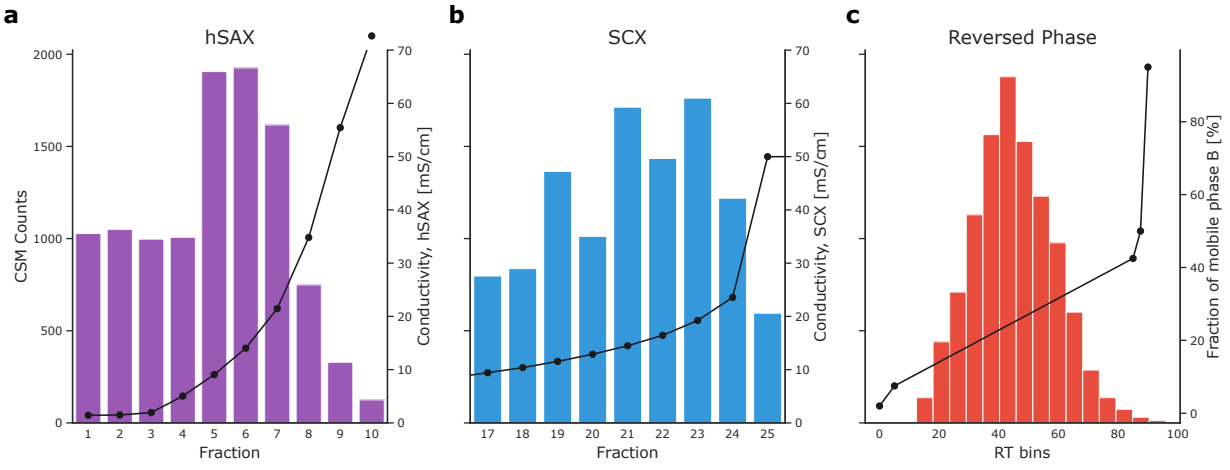
Similarly, 2) is used to speed up the training by stopping the process if no improvements were achieved in a configurable number of epochs. In addition, for the final model the best weights over all epochs are chosen based on the performance on the validation split. Finally, 3) is used to store the weights and model architecture on disk. This allows applying the best model for the respective cross-validation folds and the candidate rescoring. For transfer learning applications these trained models can also be used for new data sets. Most parameters of the network such as learning rate, optimizer, batch size, epochs, callback settings, number of layers / neurons can be adapted through a dedicated YAML file. The online documentation for xiRT on GitHub contains examples for various training and RT dimension scenarios.



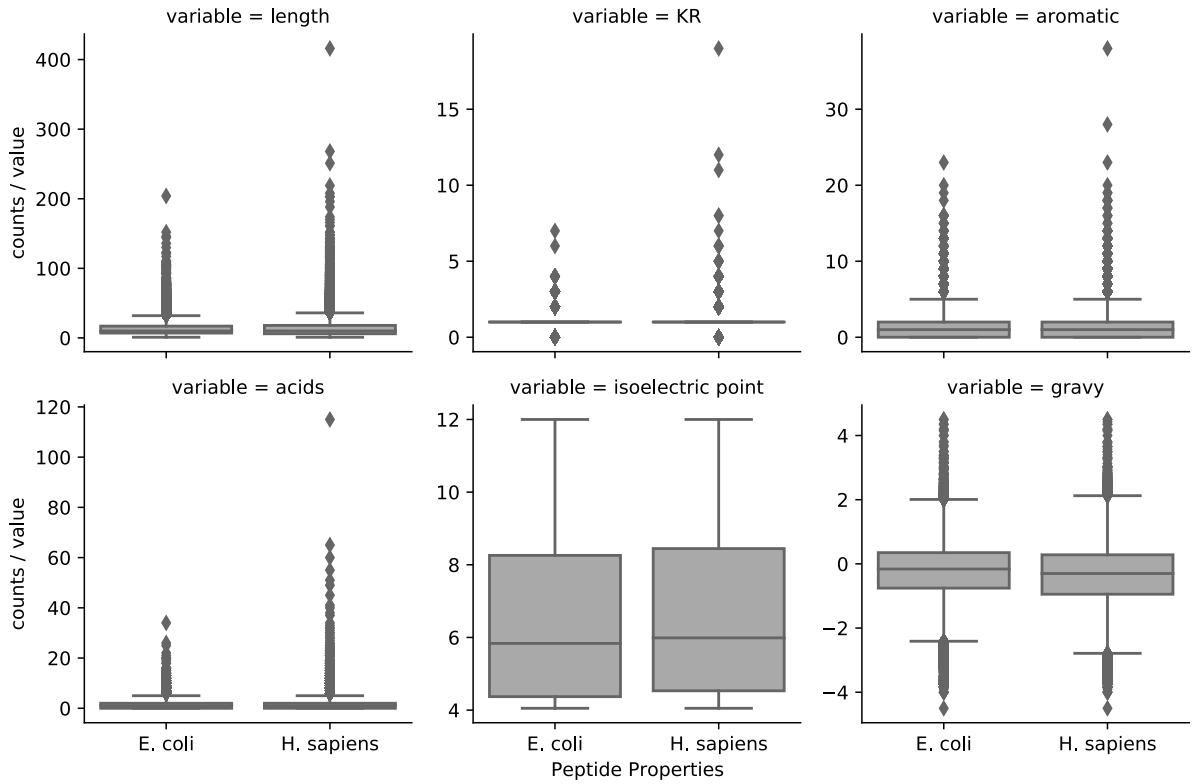
Supplementary Figure 1: Example parameterization of xiRT. Dashed box represents the Siamese network part. Boxes represent individual layers with their names, input and output dimensions. Question marks represent the unknown batch-size at compilation time of the network.



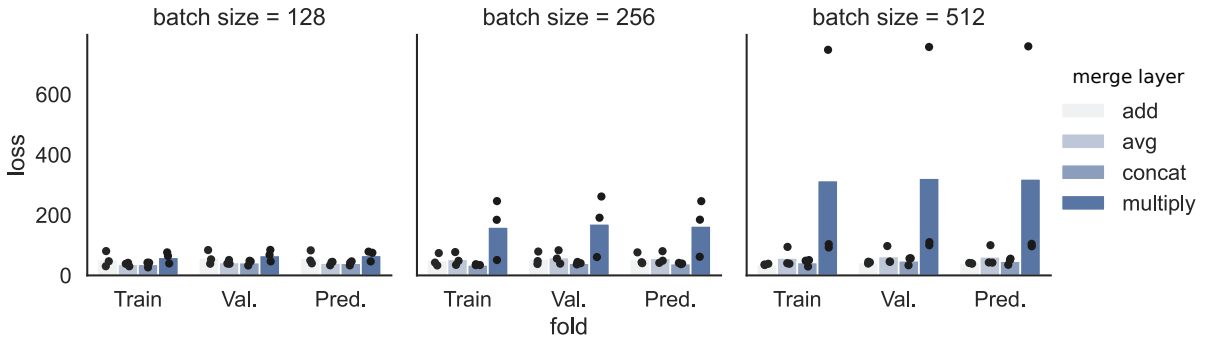
Supplementary Figure 2: Cross-validation results from applying xiRT on linear peptide input. a) Average training performance on all tasks (SCX - blue; hsAX - purple; RP - red) over 75 epochs from k=3 CV-folds. Confidence intervals show standard deviation from a 3-fold CV with the dashed/solid line representing the mean for the validation/training data, respectively.. b) Evaluation metrics for all tasks on the different CV folds. c-e) Representative results for a random prediction fold. Abbreviations: val, validation; pred, prediction, unval, unvalidated; mse, mean squared error; acc, accuracy.; racc, relaxed accuracy ($|error| \leq 1$).



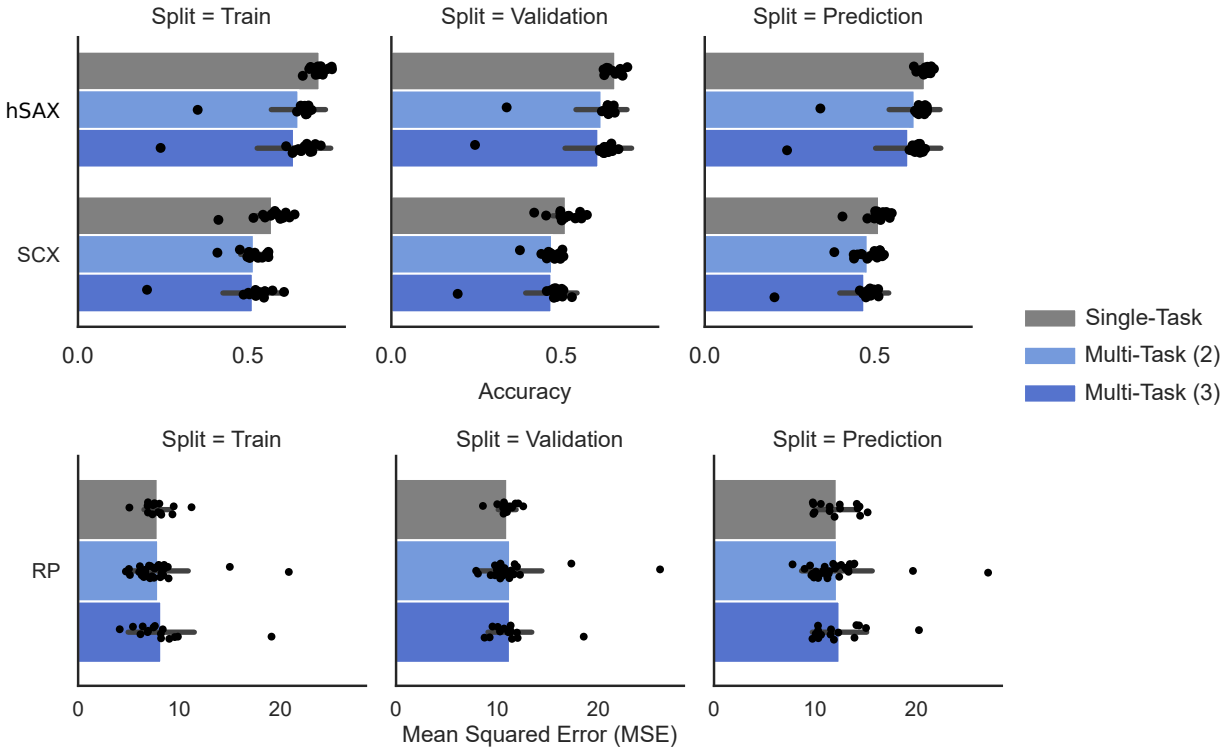
Supplementary Figure 3: Crosslink identifications over fractions / time. (a-b) Distribution of CSMs across native fraction numbers from the off-line fractionation based on strong cation exchange (SCX) and hydrophilic strong anion exchange (hSAX) chromatography. Black lines indicate the eluent concentration (represented by conductivity) at the beginning of the fraction. (c) Distribution of CSMs across reversed-phase retention time bins. Black lines indicate the eluent concentration (fraction of eluting mobile phase B) at the beginning of the fraction. Data corresponds to 11072 CSMs at 1% FDR, all target-target hits excluding matches involving human proteins.



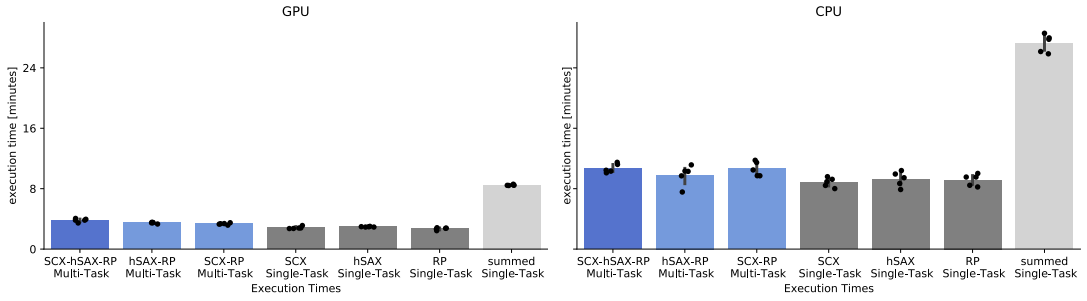
Supplementary Figure 4: Comparison of peptide properties from the E. coli target database and the H. sapiens entrapment database. Variables: KR, K/R count in peptide; aromatic, F/Y/W counts; acids, D/E counts; isoelectric point and GRAVY were computed using Biopython [5]. Boxplots show the median as line in the IQR-box and the whiskers show the 1.5x interquartile range (min & max), points represent the outliers. E. coli box represents n=69175 peptides, H. sapiens box represents n=66083 peptides, respectively.)



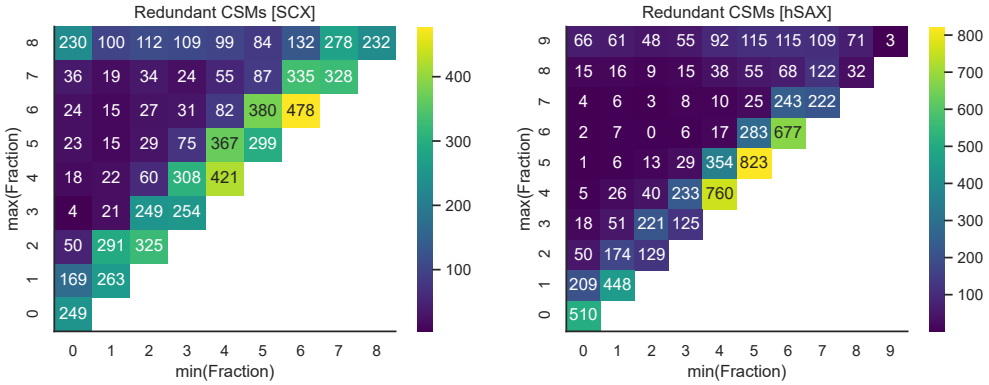
Supplementary Figure 5: Hyper-parameter optimization for xiRT. Appropriate hyper-parameters were assessed following cross-validation (k=3) on crosslinked peptides. The different merge functions (add, average, concat, multiply - from light blue to dark blue) represent tensorflow implementations for the combination of the two input vectors from the Siamese network outputs. Bars indicate the mean.



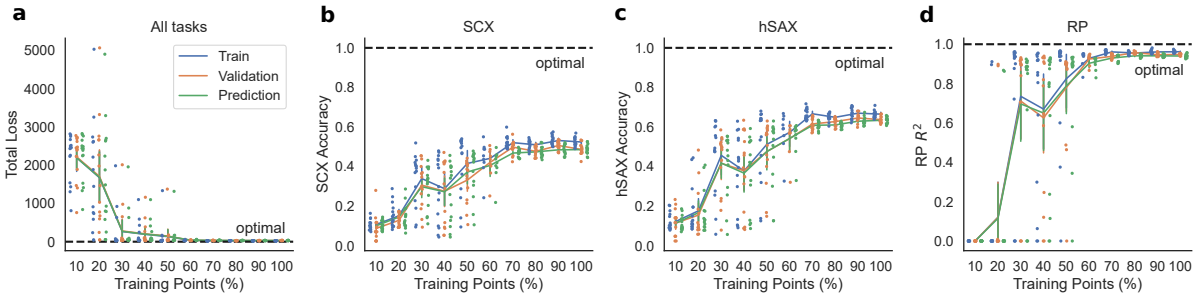
Supplementary Figure 6: xiRT performance for single- (grey) and multi-task parameterizations (with either (2) or (3) tasks in lightblue and blue, respectively). Bars show the mean value from five replica of 3-fold CV (every dot represents a single CV result). One-way ANOVA (type 2) results fail to reject the null hypothesis of an equal mean in all groups at $\alpha = 0.05$ (prediction split only). Results were: SCX ($F = 2.7$, $p - value = 0.08$, $n = 45$), hSAX ($F = 1.69$, $p - value = 0.20$, $n = 45$), RP ($F = 0.04$, $p - value = 0.96$, $n = 60$), with 2 degrees of freedom and 42 total df for the SCX/hSAX test and 57 total df for the RP test. Error bars show the standard deviation. For all bars n=15, except for the Multi-Task (2) in the RP analysis, where RP results are derived from SCX-RP and hSAX-RP, leading to n=30 observations for the RP performance.



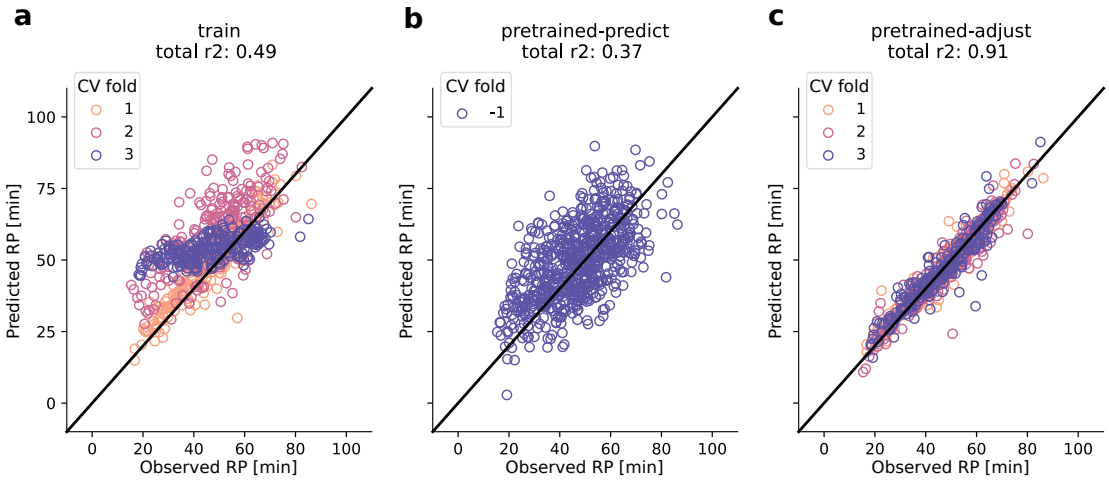
Supplementary Figure 7: xiRT benchmark for single- (grey) and multi-task (blue) parameterizations on different hardware. Every parameter was tested in n=5 replicates. For the 'summed' single-task estimate random replicates were paired. GPU analysis was performed on Intel(R) Xeon(R) CPU E5-1620 v4 @ 3.50 GHz equipped with an TITAN X (Pascal) with 12GB memory. CPU analysis was performed on Intel(R) Core i7 6700K CPU @ 4.00 GHz, with 32GB DDR4 memory. Error bars shown the standard deviation.



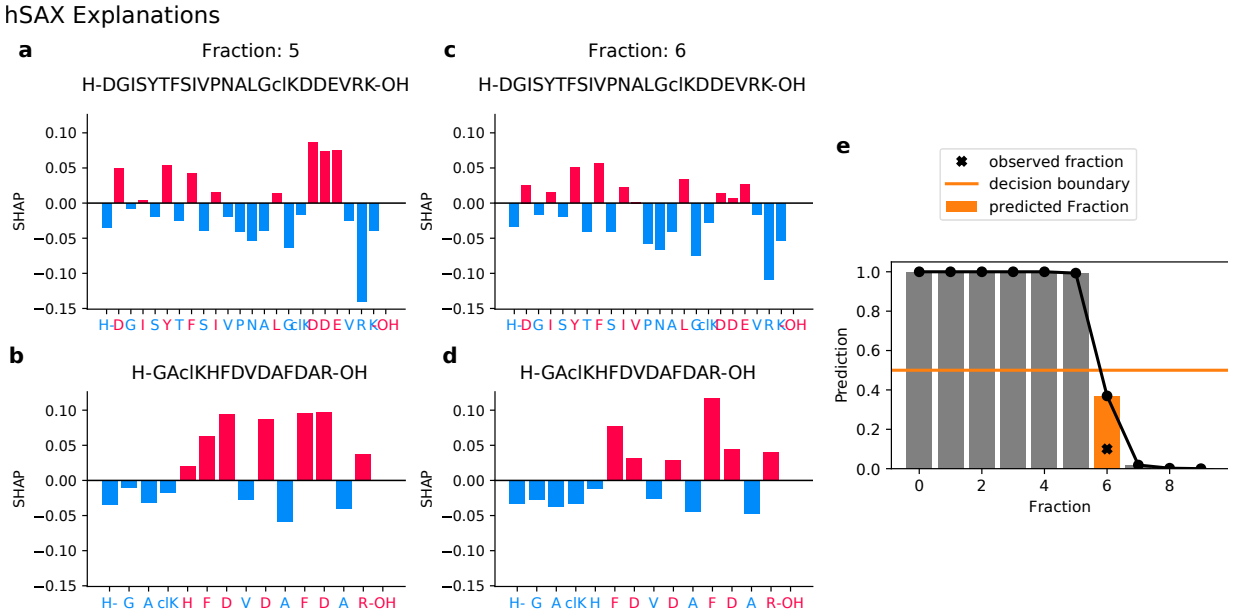
Supplementary Figure 8: Redundancy of CSMs across SCX / hSAX fractions (6843 non-unique CSMs at at 1% FDR). If a CSM was identified in multiple fractions, the span (min and max of the fraction) was calculated and visualized. For example, if a CSM was identified in the fractions 2,3,4, the span would be (2, 4), leading to an increase in the plot at x=2 and y=4. In other words, 60 CSMs were indeed observed with a span of (2, 4). CSM redundancy was high in the last fractions (8 for SCX, 9 for hSAX) emphasizing ambiguous retention behavior. Fraction numbers were transformed to start at zero.



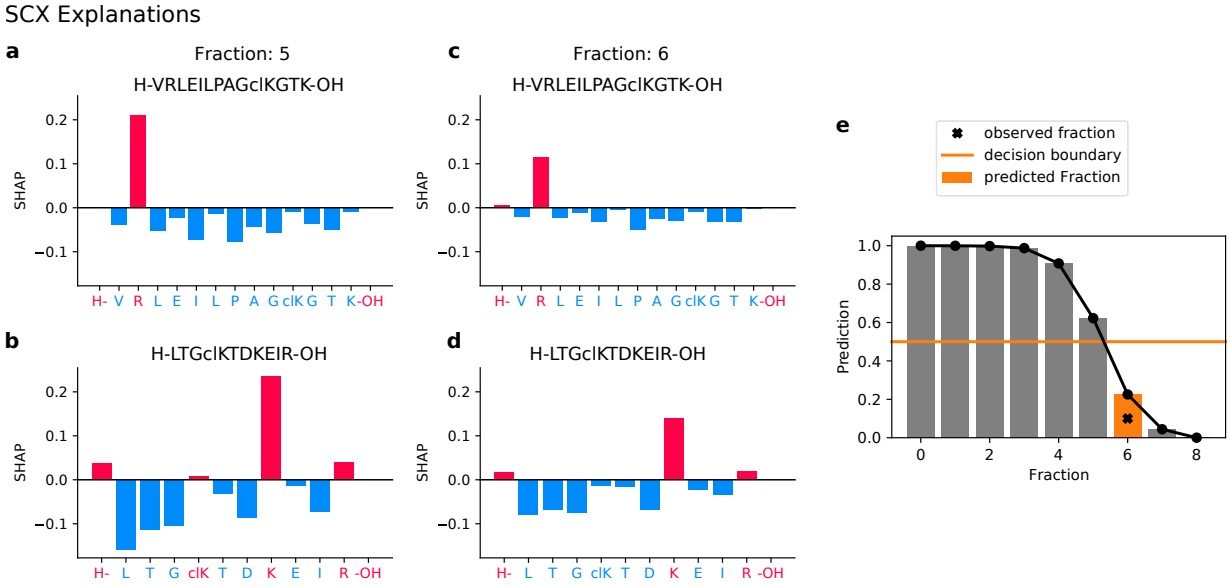
Supplementary Figure 9: Learning results on crosslink data from 3-fold cross-validation (five replicates). a) total unweighted loss is shown, b and c) classification accuracy and d) R^2 for the RP prediction is shown. Data used (train, validation, prediction): 10% (387, 43, 215); 20% (774, 87, 430); 30% (1161, 130, 645); 40% (1548, 173, 860); 50% (1935, 216, 1075); 60% (2323, 259, 1290); 70% (2710, 302, 1505); 80% (3097, 345, 1720); 90% (3484, 388, 1936); 100% (3871, 431, 2151). Vertical bars show the standard deviation with the mean as center. Training, validation and prediction performance is shown in blue, orange and green, respectively.



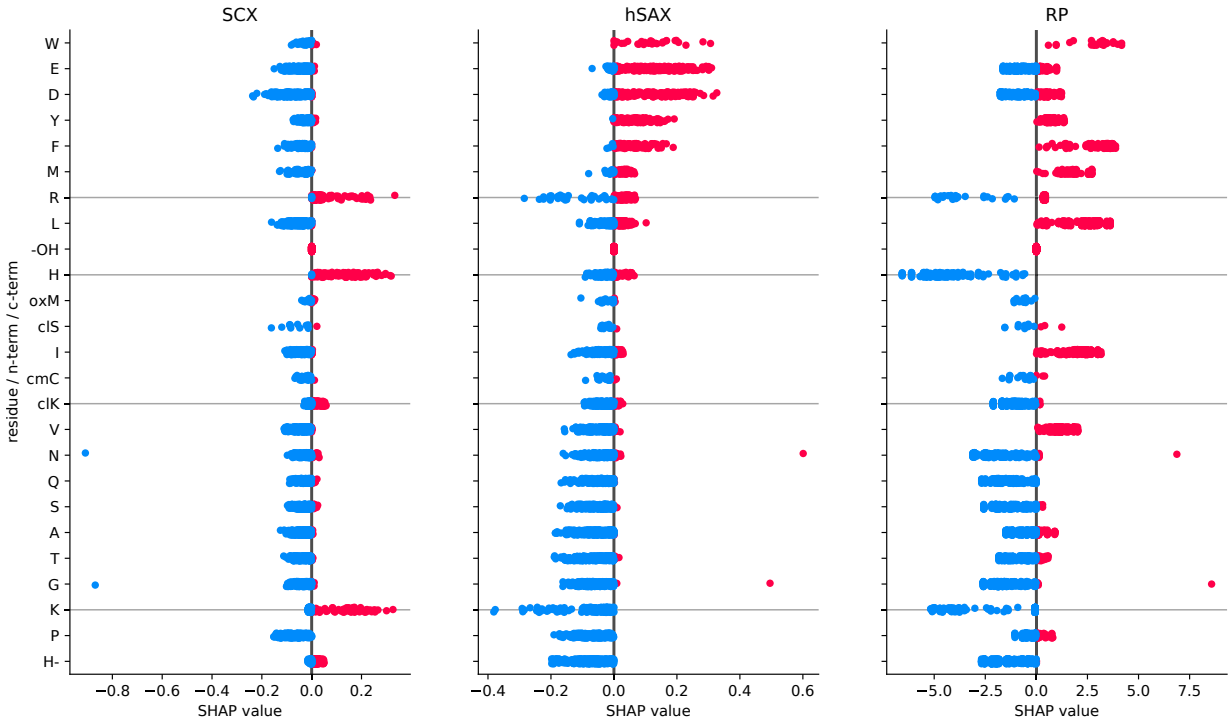
Supplementary Figure 10: Reversed-phase RT prediction for the Fanconi anaemia monoubiquitin ligase complex data set[6] (FA-complex).The panels show the individual cross-validation predictions for the three training set-ups. The modes (a) *train* (solely train on FA-complex data), (b) *pretrained-predict* (apply pretrained model, using the subset of E. coli DSS-crosslink data at 1% CSM-FDR without fine-tuning), (c) *pretrained-adjust* (load the previously described model, include fine-tuning the network during cross-validation). For the pretrained-predict model, no cross-validation was performed. While the data contained 1400 CSMs at 1% CSM-FDR, only about 700 CSMs were used by xiRT due to high peptide sequence redundancy. Cross-validation folds are shown in orange, blue and red for the folds 1 to 3.



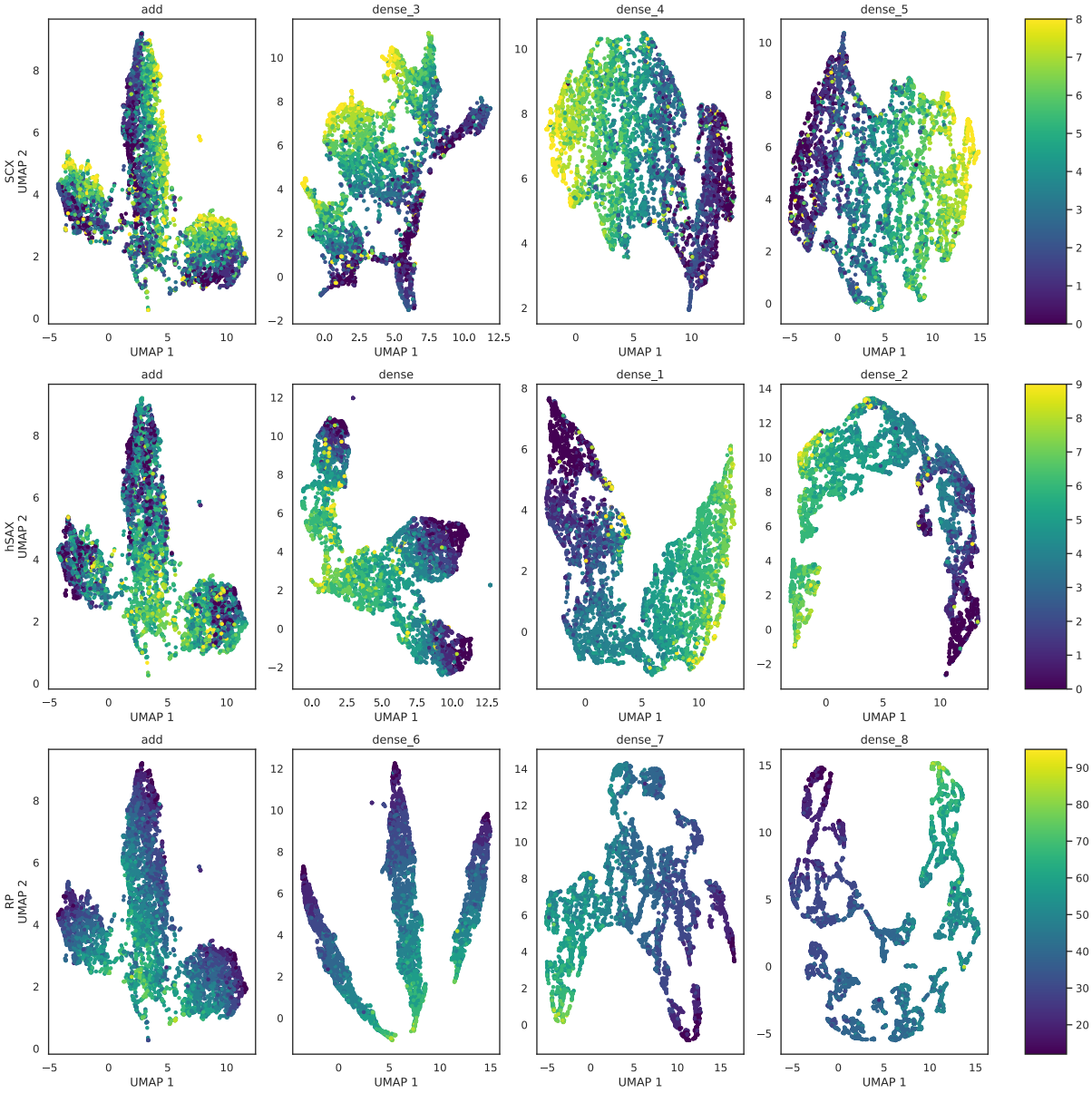
Supplementary Figure 11: SHAP explanations for the hSAX elution behavior of a crosslinked peptide that eluted in hSAX fraction 6 (0-based). a-b) SHAP values for the crosslinked peptide DGISYTFSIVPNALGkDDEVRK-GAkHFDVDAFDAR (H- = N-terminus, clK = crosslinked lysine residue, -OH = C-terminus) for the prediction to elute in fraction 5. c-d) SHAP values for the peptide to elute in fraction 6. e) Predicted output of the network (ordinal fraction prediction) that is translated into the predicted fractions. The fraction is determined by the first prediction that yields a probability lower than 0.5 (orange line). Negative (blue) SHAP-values contribute towards an earlier elution, while positive (red) SHAP-values contribute towards later elution.



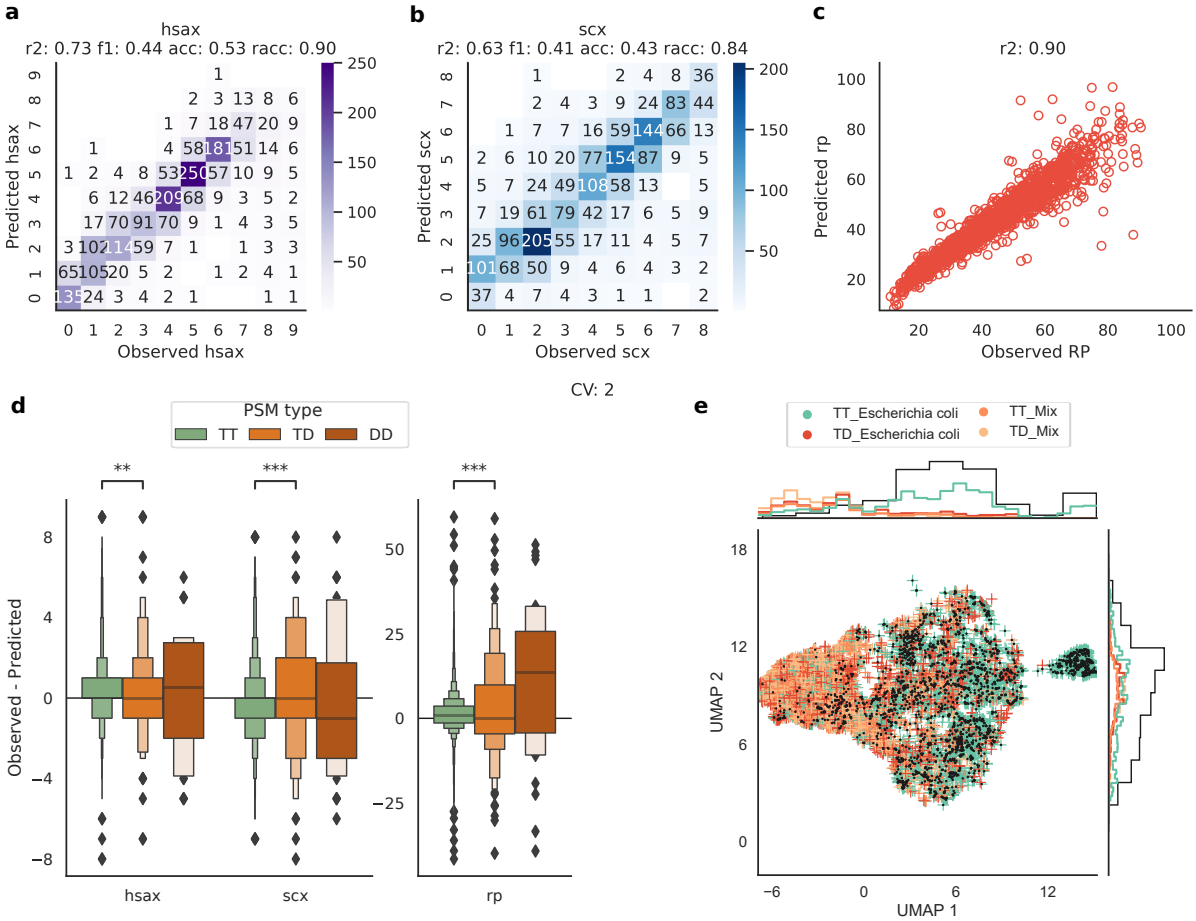
Supplementary Figure 12: SHAP explanations for the SCX elution time of a crosslinked peptide that eluted in SCX fraction 6 (0-based). See Supplementary Figure 11 for a detailed description.



Supplementary Figure 13: SHAP values for individual amino acids of crosslinked peptides contributing towards early or late elution in SCX/hSAX/RP separations. Positive (red) SHAP values contribute towards later elution, negative (blue) SHAP values contribute towards earlier elution. Horizontal grey lines highlight R, K, clK, H residues. SHAP values were computed for 500 randomly drawn CSMs passing 1% CSM-FDR.



Supplementary Figure 14: Visualization of the embedding space throughout the network for each task. UMAP parameters: metric, Euclidean; min_dist, 0.0; n_neighbors, 15. Individual plots correspond to the add (shared) and dense (task-specific) layers in Supplementary Figure 1. UMAP was applied to a cross-validation model from the E. coli DSS data set. Note that the given parameterization of UMAP might be suboptimal for all the selected embedding spaces. Color bar represents the retention time in each dimension either in minutes (RP) or discrete fractions (hSAX/SCX).



Supplementary Figure 15: Evaluation of combining xiRT with pLink2 search results. a-c) Cross-validation results on the second prediction fold using all three RT dimensions (hSAX, SCX, RP). pLink2 data was filtered to a Q-value of 0.01 for the training process in xiRT. d) Error characteristics for TT (green, n=6436), TD (orange, n=214) and DD (brown, n=34) CSMs. P-values are derived from a two-sided, independent t-test with Bonferroni correction between TT and TD observations. P-values: 1.545e-03 (hsax), 5.667e-04 (scx), 3.381e-04 (rp), test-statistics: -3.364e+00 (hsax), -3.632e+00 (scx), -3.586e+00 (rp). e) Dimensionality reduced feature space using UMAP with default parameters. Black dots represent CSMs that passed the 0.01 Q-value cutoff. Only heteromeric crosslink spectrum matches are shown. TT and TD CSMs from E. coli are shown in green and red, while TT and TD CSMs from human peptides crosslinked to E. coli are shown in orange (TT) and peach (TD).

Supplementary Note 2: Hyper-Parameter Optimization on Linear Data

Neural networks are subject to many parameters that need tuning to achieve the best possible performance. Based on our initial work on the prediction of hSAX RTs for linear peptides [7], we came up with an initial architecture and then optimized it manually. Further, we choose a step-wise approach to find suitable hyper-parameters during a 3-fold cross-validation search. The first grid of hyper-parameters is shown in Table 1. For the CV, we split the data into a *training* fold, a *validation* fold (10% of the training fold data) and a testing fold per CV iteration. We also use the term *prediction fold* synonymous to the *testing fold* since we only use the testing fold predictions for CSM rescoring later on. All decoy-PSMs and identifications with a FDR higher than the selected training FDR are assigned to an *unvalidation* fold. After the first round of CV on a set of 576 parameters, another grid-search (320 parameters) was performed with adjusted hyper-parameters (Table 2). This second grid was based on the best performing parameters in the first iteration with slight variations. Note, the linear peptide identifications at 1% PSM-FDR were used for this procedure (n=20802 unique sequences, ignoring identifications to the entrapment database). For the execution of the hyper-parameter search, we again designed a snakemake [8] workflow that can run an arbitrary number of configuration files. The best final parameters were then chosen based on the means of the loss, r^2_{rp}, $accuracy_{hSAX}$ and $accuracy_{SCX}$ in the testing sets during CV. Note that the 2-step optimization offers a reasonable trade-off between finding optimal parameters and decreasing the necessary run time.

The best parameters from optimization (Tab. 1, Tab. 2), showed an average ($\pm$ standard deviation) R2 of 0.99 $\pm$ 0.003 for the RP task, average accuracy 64% $\pm$ 0.9 for hSAX task and 46% $\pm$ 0.7 for the SCX task (Supplementary Figure 2). By using a relaxed accuracy metric (absolute prediction error $\leq$ 1 fraction), hSAX RT prediction achieved 92% $\pm$ 0.3 and SCX RT prediction 74% $\pm$ 0.7.

The network performance across the individual CV-folds of the best parameter was very comparable in terms of training time and performance (Supplementary Figure 2a). The CV was performed on 20802 unique CSMs (train: 12481, validation: 1387, prediction: 6934 observations). The learning trajectory of the number of epochs follows a very smooth learning curve and shows a constant improvement in the training and validation fold with a small gap between the training and validation performance. We also observed that the prediction accuracy for hSAX is better than for SCX in both, training and validation data. This trend is also observable in the prediction folds (Supplementary Figure 2b). In addition, the performance drop from the validation fold to the prediction fold is rather small which is desirable and shows good generalization ability of the network. A lower prediction performance on the unvalidation split can be expected and hence serves as another quality check. The predictions were made with the best classifier from the CV split. The individual predictions for a single CV-fold are more accurate for the RP than for SCX or hSAX (Supplementary Figure 2c-e). While the RP predictions achieve an r^2 of 0.99, the accuracy in SCX and hSAX is limited to 0.64 and 0.45, respectively. The different behaviour of hSAX and SCX might be explained through deviating peptide separation behaviour with the applied gradients (Supplementary Figure 3). While the shape of the gradients is similar, the hSAX gradient led to a more uniform distribution of crosslinked peptides across the elution window in contrast to the more confined elution of crosslinked peptides in later SCX fractions. Therefore, adjacent SCX fractions are expected to show a higher overlap in their identifications than fractions from hSAX.

Supplementary Table 1: First parameter grid for the optimization on linear peptide data.

Parameter	Parameter Grid	Selected Parameter
recurrent type	CuDNNGRU	CuDNNGRU
recurrent units	25, 75, 125	75
recurrent activity l2 lambda	0, 0.001	0.001
recurrent kernel l2 lambda	0	0
dense layers	3	3
dense neurons	(300, 150, 75), (150, 100, 50)	(150, 100, 50)
dense kernel l2 lambda	(0.001, 0.001, 0.001)	(0.001, 0.001, 0.001)
dense dropout	(0.3, 0.3, 0.3), (0.1, 0.1, 0.1)	(0.1, 0.1, 0.1)
dense activation	(relu, relu, relu), (swish, swish, swish)	(swish, swish, swish)
embedding length	50, 100	50
batch size	256, 512	256
class weight	1, 250, 500	250

Note: Parameters with a prefix "recurrent" were used for a single recurrent layer. Parameters with the "dense" prefix were used for the task specific layers. A total set of 576 parameter combinations were used during the grid search. Remaining settings were left at defaults. The best parameter was determined based on the testing folds in a 3-fold CV experiment. Training time was limited to 75 epochs and early stopping patience was set to 15.

Supplementary Table 2: Second parameter grid for the optimization on linear peptide data.

Parameter	Parameter Grid	Selected Parameter
recurrent type	CuDNNGRU, CuDNNLSTM	CuDNNGRU
recurrent units	75	75
recurrent activity l2 lambda	0, 0.001	0.001
recurrent kernel l2 lambda	0, 0.001	0.001
dense layers	3	3
dense neurons	(300, 150, 75)	(300, 150, 75)
dense kernel l2 lambda	(0.001, 0.001, 0.001)	(0.001, 0.001, 0.001)
dense dropout	(0.2, 0.2, 0.2), (0.1, 0.1, 0.1)	(0.1, 0.1, 0.1)
dense activation	(relu, relu, relu), (swish, swish, swish)	(relu, relu, relu)
embedding length	50, 75	50
batch size	256	256
class weight	1, 50, 100, 200, 250	50

Supplementary Note 3: xiRT Explainability Analysis

In this section we describe the analysis of the SHAP values from the learned multi-task model. For this, the used tensorflow-version needed to be downgraded to 1.15 together with SHAP (v. 0.36.0). As background data 100 randomly chosen CSMs were provided. To use the DeepExplainer, the trained network had to be dissected into the single tasks. Furthermore, the ordinal regressions setup for hSAX and SCX complicates the analysis since each sigmoid activation of the output vector can be explained via SHAP (padded positions were ignored). Therefore, we only focused on the SHAP values for the relevant prediction decision, i.e. the sigmoid activiation that was $<= 0.5$. With this special model architecture, the returned SHAP-values failed the 'check_additivity' flag in the SHAP package and the check was thus disabled. However, the magnitude and overall explanations from the DeepExplainer show realistic feature importance values for the RT contributions on residue level. Since the SHAP values only represent an approximation of the contributions we further explored their magnitude. In Supplementary Figure 11 we demonstrate the explainability via SHAP of a crosslinked peptide's predicted retention time (hSAX fraction). The residues D, E, R and K behave mostly as expected. In addition, aromatics (Y, F) also contribute to stronger retention and hence later elution times, while A contributes towards earlier elution times. These observations are in line with an earlier study on the hSAX RT behavior[7]. Similarly, an explanation for a SCX prediction is shown in Supplementary Figure 12. The global SHAP values based on the raw sequence inputs to xiRT are shown in Supplementary Figure 13. For hSAX again, D, E, F, Y, W belong to the major contributors towards extended retention times. For SCX, the positive contribution is

mainly attributed towards R, K and H. Note that crosslinked K residues, contribute much less towards later elution times than non-crosslinked K residues.

Supplementary Table 3: RT features used for prediction on *E. coli* data set.

#	Feature Name	Description
1	hsax-error	crosslinked - raw error between observed and predicted (hSAX)
2	scx-error	crosslinked - raw error between observed and predicted (SCX)
3	rp-error	crosslinked - raw error between observed and predicted (RP)
4	hsax-error-peptide1	raw peptide 1 error (hSAX)
5	scx-error-peptide1	raw peptide 1 error (SCX)
6	rp-error-peptide1	raw peptide 1 error (RP)
7	hsax-error-peptide2	raw peptide 2 error (hSAX)
8	scx-error-peptide2	raw peptide 2 error (SCX)
9	rp-error-peptide2	raw peptide 2 error (RP)
10	peptide1_mean	median of all peptide1 error (absolute values)
11	peptide1_sum	sum of all peptide1 error (absolute values)
12	peptide1_max	maximum of all crosslinked errors
13	peptide1_min	minimum of all peptide1 errors
14	peptide2_mean	median of all peptide2 errors (absolute values)
15	peptide2_sum	sum of all peptide2 errors (absolute values)
16	peptide2_max	maximum of all peptide2 errors
17	peptide2_min	minimum of all peptide2 errors
18	cl_mean	median of all crosslinked errors (absolute values)
19	cl_sum	sum of all crosslinked errors (absolute values)
20	cl_max	maximum of all crosslinked errors
21	cl_min	minimum of all crosslinked errors
22	initial_prod	log2 product (absolute values + 0.1) of all initial errors (#1-9)
23	initial_sum	sum (absolute values) of all initial errors (#1-9)
24	initial_min	minimum (absolute values) of all initial errors (#1-9)
25	initial_max	maximum (absolute values) of all initial errors (#1-9)
26	hsax-error_square	squared hsax-error for crosslinked errors
27	hsax-error_abs	absolute hsax-error for crosslinked errors
28	scx-error_square	squared scx-error for crosslinked errors
29	scx-error_abs	absolute scx-error for crosslinked errors
30	rp-error_square	squared rp-error for crosslinked errors
31	rp-error_abs	absolute rp-error for crosslinked errors
32	hsax-error-peptide1_square	squared hsax-error for peptide1 errors
33	hsax-error-peptide1_abs	absolute hsax-error for peptide1 errors
34	scx-error-peptide1_square	squared scx-error for peptide1 errors
35	scx-error-peptide1_abs	absolute scx-error for peptide1 errors
36	rp-error-peptide1_square	squared rp-error for peptide1 errors
37	rp-error-peptide1_abs	absolute rp-error for peptide1 errors
38	hsax-error-peptide2_square	squared hsax-error for peptide2 errors
39	hsax-error-peptide2_abs	absolute hsax-error for peptide2 errors
40	scx-error-peptide2_square	squared scx-error for peptide2 errors
41	scx-error-peptide2_abs	absolute scx-error for peptide2 errors
42	rp-error-peptide2_square	squared rp-error for peptide2 errors
43	rp-error-peptide2_abs	absolute rp-error for peptide1 errors

Note: Features computed from xiRT predictions. All errors or predictions are derived from the same xiRT model for crosslinked peptides. In the case of individual peptide predictions (peptide1/peptide2), the second sequence in the input is set to all-zeroes. This feature set was used in the *E. Coli* analysis with all three RT dimensions.

Supplementary Table 4: Unique and redundant CSMs across hSAX and SCX fractions.

	Total	Unique	Redundant	hSAX (red/same)	hSAX (red/diff)	SCX (red/same)	SCX (red/diff)
Counts	39226	4500	6843	3729	3114	2849	3994
%	100	40	60	33	27	25	35

Note: Unique and redundant CSM identifications at 1% CSM-FDR (separate for heteromeric and self-links). Unique CSMs are combinations of peptide 1, peptide 2, link site and charge state that were only identified once. Redundant (red) CSMs were identified more than once and thus can either have different RT times ("diff") or the same RT times ("same"). Percentages show the observations divided by the sum of unique and redundant CSMs (rounded). The theoretical accuracy limit for hSAX and SCX was derived by summing the percentages of unique and 'red/same' CSMs (hSAX: 73%, SCX: 65%).

Supplementary Table 5: Rescoring gains with different number of chromatographic dimensions.

	level	reference	RP	SCX-RP	hSAX-RP	SCX-hSAX-RP
heteromeric	CSM	724	902 (+1.25x)	977 (+1.35x)	1092 (+1.51x)	**1199 (+1.66x)**
heteromeric	Peptide	507	619 (+1.22x)	664 (+1.31x)	737 (+1.45x)	**801 (+1.58x)**
heteromeric	Residues	414	508 (+1.23x)	546 (+1.32x)	603 (+1.46x)	**654 (+1.58x)**
heteromeric	PPI	109	135 (+1.24x)	131 (+1.2x)	**157 (+1.44x)**	152 (+1.39x)
self	CSM	10357	10404 (+1.0x)	10428 (+1.01x)	10439 (+1.01x)	**10443 (+1.01x)**
self	Peptide	6521	6565 (+1.01x)	6586 (+1.01x)	6598 (+1.01x)	**6601 (+1.01x)**
self	Residues	4810	4853 (+1.01x)	4873 (+1.01x)	4886 (+1.02x)	**4888 (+1.02x)**
self	PPI	478	514 (+1.08x)	531 (+1.11x)	540 (+1.13x)	**543 (+1.14x)**

Note: The data corresponds to all *E. coli* target-target identifications at 5% CSM-, Peptide-, Residue-level FDR and 1% PPI-FDR. Rescoring was performed using a linear SVM. Highest values are marked in bold. The hyper-parameters for the rescoring were chosen dynamically via cross-validation for each run ('class_weight': 'None', all conditions; 'C': 100 (RP, SCX-RP, SCX-hSAX-RP); 'C': 10 (hSAX-RP)), according to the sklearn API. Values are rounded to two digits.

Supplementary Table 6: CSMs / PPIs involving a human protein (rescored results).

PSMID	Protein 1	Protein 2 (E. coli) initial	Protein 2 (human) corrected	Peptide 1 (E. coli)	Peptide 2 (human) initial	Peptide 2 (E. coli) corrected
2262348	P0AFG6	P50552	P0AFG6	SEEKclASTPAQR	KELQKclVK	KIKclELVAK
3165715	P0AFG6	P50552	P0AFG6	EDVEKclHLAK	KELQKclVK	KIKclELVAK
2545576	P0AFG6	P50552	P0AFG6	LLAEHNLDASAI-KclGTGVGGR	KELQKclVK	KIKclELVAK

Note: The displayed CSMs correspond to the rescored identifications involving a human peptide as shown in the manuscript (Figure 4). Three human target CSMs are shown that result in a single PPI between a human protein and the E. coli protein SucB at 1% PPI-FDR (up to 5% for lower FDR levels). Manual inspection revealed the SucB peptide KIKELVAK as a better match, i.e. a peptide of the same E. coli protein that the peptide 1 is from. It had not been matched as it carries a rare modification that was not included in our original search.

Supplementary Note 4: pLink2 Processing

The recalibrated MGF files were searched with pLink 2 (2.3.9) with the following search parameters: Flow Type, HCD; Cross-Linker, DSS with AlphaSites = BetaSites = [KSTY; Enzyme, Trypsin; Missed Cleavages, 2; Peptide Mass [600, 6000]; Precursor Tolerance, 5ppm; Fragment Tolerance, 3ppm; Fixed Modifications, Carbamidomethylation[C]; Variable Modifications, Oxidation[M]. The filter parameters were as follows: Filter Tolerance, 10 ppm; FDR, separate FDR 1% at CSM level; Compute E-value, False.

We further processed the unfiltered results table from pLink2 in order to get all CSMs (including decoys) and their associated error estimates for usage in xiRT. In short, we added the information about peptide origin, target-decoy origin, species, peptide positions and RT in SCX/hSAX/RP. These steps were only performed with the peptides that pass the 0.5 Q-value threshold (similar to the xiSEARCH processing, as only 50% CSM-FDR data was used). These additional annotation steps were necessary since the filtered pLink2 results (*.filtered_cross-linked-spectra) do not provide the necessary information (e.g. decoy hits and error estimates are not provided).

The generated file was then used as input for xiRT. For xiRT, the same settings as for xiSEARCH were used. In total 35822 peptides were used as input data. During the CV 3866 peptides were used for training, 430 for validation and 2147 for prediction (prediction-fold is visualized in Supplementary Figure 15, together with a 2D-feature space representation using UMAP[9]. Using crosslinks identified from pLink2 lead to a comparable prediction performance as using xiSEARCH (3871 training peptides, 431 validation peptides, 2151 prediction peptides).

Supplementary References

[1] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," in *Proc. 2014 Conf. Empir. Methods Nat. Lang. Process.* Stroudsburg, PA, USA: Association for Computational Linguistics, jun 2014, pp. 1724–1734. [Online]. Available: http://aclweb.org/anthology/D14-1179

[2] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, nov 1997. [Online]. Available: https://www.mitpressjournals.org/doi/abs/10.1162/neco.1997.9.8.1735

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, may 2015. [Online]. Available: http://www.nature.com/articles/nature14539

[4] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Oper. Syst. Des. Implementation, OSDI 2016*, 2016.

[5] P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. de Hoon, "Biopython: freely available Python tools for computational molecular biology and bioinformatics," *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, jun 2009. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp163

[6] S. Shakeel, E. Rajendra, P. Alcón, F. O'Reilly, D. S. Chorev, S. Maslen, G. Degliesposti, C. J. Russo, S. He, C. H. Hill, J. M. Skehel, S. H. W. Scheres, K. J. Patel, J. Rappsilber, C. V. Robinson, and L. A. Passmore, "Structure of the Fanconi anaemia monoubiquitin ligase complex," *Nature*, vol. 575, no. 7781, pp. 234–237, nov 2019. [Online]. Available: http://www.nature.com/articles/s41586-019-1703-4

[7] S. H. Giese, Y. Ishihama, and J. Rappsilber, "Peptide Retention in Hydrophilic Strong Anion Exchange Chromatography Is Driven by Charged and Aromatic Residues," *Anal. Chem.*, p. acs.analchem.7b05157, mar 2018. [Online]. Available: http://pubs.acs.org/doi/10.1021/acs.analchem.7b05157

[8] J. Koster and S. Rahmann, "Snakemake–a scalable bioinformatics workflow engine," *Bioinformatics*, vol. 28, no. 19, pp. 2520–2522, oct 2012. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts480

[9] L. McInnes, J. Healy, N. Saul, and L. Großberger, "UMAP: Uniform Manifold Approximation and Projection," *J. Open Source Softw.*, vol. 3, no. 29, p. 861, sep 2018. [Online]. Available: http://joss.theoj.org/papers/10.21105/joss.00861