Investigating chemical crosslinking of protein-RNA complexes by mass spectrometry

Dissertation

for the award of the degree

"Doctor rerum naturalium"

of the Georg-August University School of Science (GAUSS)

submitted by

Alexander Wulf

from Cuxhaven, Germany

Göttingen, 2021

Investigating chemical crosslinking of protein-RNA complexes by mass spectrometry

Dissertation

for the award of the degree

"Doctor rerum naturalium"

of the Georg-August University School of Science (GAUSS)

submitted by

Alexander Wulf

from Cuxhaven, Germany

Göttingen, 2021

Thesis Committee

Prof. Dr. Henning Urlaub, Bioanalytical Spectrometry Group,
Max Planck Institute for Biophysical Chemistry,
Clinical Chemistry, University Medical Center Göttingen

Prof. Dr. Markus Bohnsack, Department of Molecular Biology,
University Medical Center Göttingen

Prof. Dr. Patrick Cramer, Department of Molecular Biology,
Max Planck Institute for Biophysical Chemistry

Members of the Examination Board

Primary reviewer: Prof. Dr. Henning Urlaub, Bioanalytical Spectrometry Group,
Max Planck Institute for Biophysical Chemistry,
Clinical Chemistristry, University Medical Center Göttingen

2nd reviewer: Prof. Dr. Markus Bohnsack, Department of Molecular Biology,
University Medical Center Göttingen

Further members of the Examination Board

Prof. Dr. Ralf Ficner, Department of Molecular Structural Biology,
Georg-August-Universität Göttingen

Prof. Dr. Jörg Stülke, Department of General Microbiology,
Georg-August-Universität Göttingen

Dr. Alexander Stein, Department of Membrane Protein Biochemistry,
Max Planck Institute for Biophysical Chemistry

Date of oral examination: November, 22nd, 2021, at the Max Planck Institute for Biophysical Chemsitry

1 Summary

RNA plays a quintessential role in cellular processes such as gene expression, transcription, RNA processing, and translation. Proteins carrying out those diverse functions are inherently flexible to accommodate dynamic RNA structures, leading to a structurally adjustable RNA-binding protein (RBP) complex that often evades crystallographic or cryo-EM analyses. To elucidate protein-RNA interfaces of such RBPs, crosslinking coupled with mass spectrometry (XL-MS) can provide direct evidence of interacting amino acid residues and nucleotides. XL-MS captures transient interactions that are stabilized through the crosslinking event, and enable peptide-centric analyses that also shed light on the crosslinked nucleotide.

Here, a chemical crosslinking workflow for the chemical crosslinkers 1,2,3,4-diepoxybutane (DEB), nitrogen mustard (NM), and 2-iminothiolane (2-IT) is presented that allows for specific, reliable, and unambiguous identification of protein-RNA interfaces at amino acid resolution. The workflow consists of 6 steps: chemical crosslinking, RNA and protein digestion, C18 chromatography, TiO2 enrichment, ESI-LC-MS/MS, and data analysis using the RNP_{xl}/Nuxl software. Acquired data from chemical crosslinking coupled with mass spectrometry (cXL-MS) of model protein Hsh49 with various synthetic RNAs features detectable nucleic acid fragments to all four nucleotides. Moreover, insights from this cXL-MS data demonstrate that certain crosslinkers display a nucleotide preference and that cXL-MS generally identifies complementary cross-link sites to canonical UV-crosslinking. Here, lists of mass shifts that are specific to each nucleotide and crosslinker combination are reported, allowing spectral annotation to identify crosslinked spectrum matches (CSMs) by the RNPl/Nuxl software.

The Hsh49 model system is expanded by its cognate protein Cus1 and U2 RNA, forming a part of the yeast spliceosomal complex, and analyzed by cXL-MS, revealing RNA-binding activity in both RNA recognition motifs. Additionally, methyltransferase Dnmt2 and the negative elongation factor complex (NELF) are analyzed by the developed method. This revealed previously unknown alternative confirmations of Dnmt2 and substantial RNA-binding events within the tentacle regions of the NELF complex, adding valuable information on the protein-RNA interplay in these complexes. When expanding cXL-MS of protein-RNA complexes to live cells, multiple RBPs from *E. coli*, *B. subtilis*, and human HeLa cells were identified in states bound to their cognate RNAs *in vivo*, which illustrates the powerful application of this method to complex samples.

Identified RBPs are mostly of the ribosomal protein family, fitting insights generated in previous studies by others well. Lastly, high-field asymmetric waveform ion mobility mass spectrometry (FAIMS) is used as a molecular filter to enhance identification of low-abundant crosslinked peptide-RNA moieties entering the mass spectrometer, resulting in higher numbers of CSMs and consequently, a more comprehensively painted picture of the RNA-interactome in *E. coli* compared to single shot analyses. Overall, these data show that the method presented here provides a robust, reliable, and unambiguous way to identify protein-RNA crosslinking sites both *in vitro* and *in vivo* that complements canonical UV-crosslinking, leading to a more detailed depiction of protein-RNA interfaces in RBPs.

2 List of abbreviations

Table 1: List of alphabetically sorted abbreviations and acronyms.

Abbreviation	Full name/meaning
A	Adenine
ACN	Acetonitrile
Adobe	Adobe Systems Incorporated
ADP	Adenosine diphosphate
AMP	Adenosine monophosphate
API	atmospheric pressure ionization
ATP	Adenosine triphosphate
APS	Ammonium persulfate
B.	*Bacillus*
BCA	Bicinchoninic acid
C	Cytosine
CID	collision-induced dissociation
CLIP	Cross-linking immunoprecipitation
cRNA	coding RNA
CSD	cold-shock domain
CSM	crosslink spectrum matches
cXL-MS	chemical crosslinking mass spectrometry
dA	deoxy-adenosine
dC	deoxy-citidine
dG	deoxy-guanosine
DNA	deoxyribonucleic acid
ds	double-straned
dT	deoxy-thymidine
DDA	data-dependent acquisition
DIA	data-independent acquisiton
DTT	Dithiothreitol
eCLIP	enhanced CLIP
E.	*Escherichia*
EtOH	Ethanol
FAIMS	High field asymmetric waveform ion mobility spectrometry
G	Guanine
Gal	Galactose
GDP	Guanosine diphosphate
Glc	Glucose

GMP	Guanosine monophosphate
GTP	Guanosine triphosphate
HCD	Higher-energy collisional dissociation
HeLa	Henrietta Lacks
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
hnRNP	heterogeneous nuclear ribonucleoprotein
HPLC	High pressure/performance liquid chromatography
IAA	Iodoacetamide
iPrOH	Isopropyl alcohol (2-propanol)
IPTG	Isopropylβ-D-1-thiogalactopyranoside
KH domain	K-Homology domain
k-NN	k-nearest neighbor
Lac	Lactose
LiCro	LiChrosolv
LB	Lysogeny broth
LC	Liquid chromatography
lncRNA	long non-coding RNA
MI	Marker ion
MS	Mass spectrometer/mass spectrometric
mRNA	messanger RNA
miRNA	micro RNA
MT	Methyltransferase
mwco	molecular weight cut-off
ncRNA	non-coding RNA
NiNTA	Nickel nitrilotriacetic acid
NT	nucleotide
OT	Orbitrap$^{\text{TM}}$
PCI	Phenol:chloroform:isoamyl alcohol
PIPES	piperazine-N,N'-bis(2-ethanesulfonic acid)
PolII	RNA-Polymerase II
RBP	RNA-binding protein
RNP	Ribonucleopotein
RP	reversedd-phase
RNA	Ribonucleic acid
RNAP	RNA-polymerase
RRM	RNA recognition motif
rRNA	ribosomal RNA
S.	*Saccharomyces*
SAH	S-Adenosyl-L-homocysteine

SDS-PAGE	Sodium dodecyl sulphate–polyacrylamide gel electrophoresis
siRNA	small interfering RNA
snRNA	small nuclear RNA
snoRNA	small nucleolar RNA
SOB	Super optimal broth
SOC	SOB with catabolite repression
ss	single-stranded
T	Thymine
T4	T4-phage DNA ligase
TN	true negative
TP	true positive
TEA	Triethylamine
TEV	Tobacco Etch Virus nuclear-inclusion-a endopeptidase
TRIS	Trisaminomethane $(HOCH_2)_3CNH_2$
tRNA	transfer RNA
U	Uracil
UDP	Uridine diphosphate
UMP	Uridine monophosphate
UTP	Uridine triphosphate
XL-MS	crosslinking mass spectrometry
Znf	Zinc finger

List of Figures

List of Tables

Contents

3 Introduction

3.1 Mass spectrometric identification of peptide biomolecules in structural biology

Mass spectrometry (MS) has long been an analytical instrument for various analyses that require absolute proof if a certain compound is present in a sample or not. While compounds can be any chemical, synthetic or naturally occurring, MS proved to be immensely valuable to the fields of proteomics and metabolomics by providing a highly sensitive method that can be used in a high-throughput fashion. The field of proteomics in general has immensely benefited from advances in MS technology, producing faster and more sensitive instruments that can identify hundreds of proteins in a matter of minutes. Apart from classical identification of proteins, MS can also aid in answering questions in the realm of structural biology through native MS, top-down protein sequencing, cross-linking-MS, and hydrogen–deuterium exchange-MS, enabling characterization of biomolecular structures, functions, and interactions [1]. Examples of questions in structural biology that were answered using MS techniques are prodigious: membrane interactions of the endosomal Vps34 complex, elucidation of the RNA-interface in type IV CRISPR-Cas complexes, conformational dynamics of the ABC transporter P-glycoprotein, generating tRNA binding models, organization of the nuclear pore complex scaffold, elucidation of the structural components of the ribosome, nucleotide binding to the p97 ATPase, study of nucleosome formation, the interfaces of the H2A-H2B-Nap1 histone-chaperone complex, interactions of mitochondrial proteins under different treatments, the norovirus assembly pathway, protein interactions upon stimulation of synaptic vesicles, the exosome complex topology, Hfq-dependent translational regulation, the architecture of the dynein cofactor dynactin, and many more. [2–17].

Mass spectrometric identification of proteins is carried out in three fundamental steps: i) sample preparation and ionization of analytes, ii) recording of mass spectra of analytes, and iii) computer-based analysis of mass spectrometric data. This general process is depicted in figure 1. First, purified samples, cells, or tissues are treated and lysed to generate a mixture of different proteins. In classical MS-based protein identification, the sample is then subjected to SDS-PAGE analysis to reduce complexity, but other chromatographic approaches such as size-exclusion chromatography (SEC), reversed-phase chromatography (RP-chromatography) of affinity-based chromatography may be used instead. Once the protein(s) of interest are sufficiently separated from other co-purified proteins, enzymatic digestion of the protein (mixture) generates peptides with specific N-terminal amino acids or C-terminal amino acids. Trypsin is most commonly used as it cleaves the peptide backbone C-terminally from arginine or lysine (if not preceded by proline), ensuring a positive charge on the peptide from the side chain that facilitates ionization. Other enzymes such as chymotrypsin (cleaves after aromatic amino acids), LysC (cleaves C-terminal from lysine), LysN (cleaves N-terminal from lysine), AspN (cleaves N-terminal from aspartate), GluC (cleaves C-terminal from glutatmate) and ArgC (cleaves C-termianl from arginine) are also used either alone or in combination with other enzymes, which are reviewed by Giansanti *et al.* [18]. If complexity is still very high and/or specific enrichment steps on the peptide level are required, a second chromatographic round of separation may follow digestion.

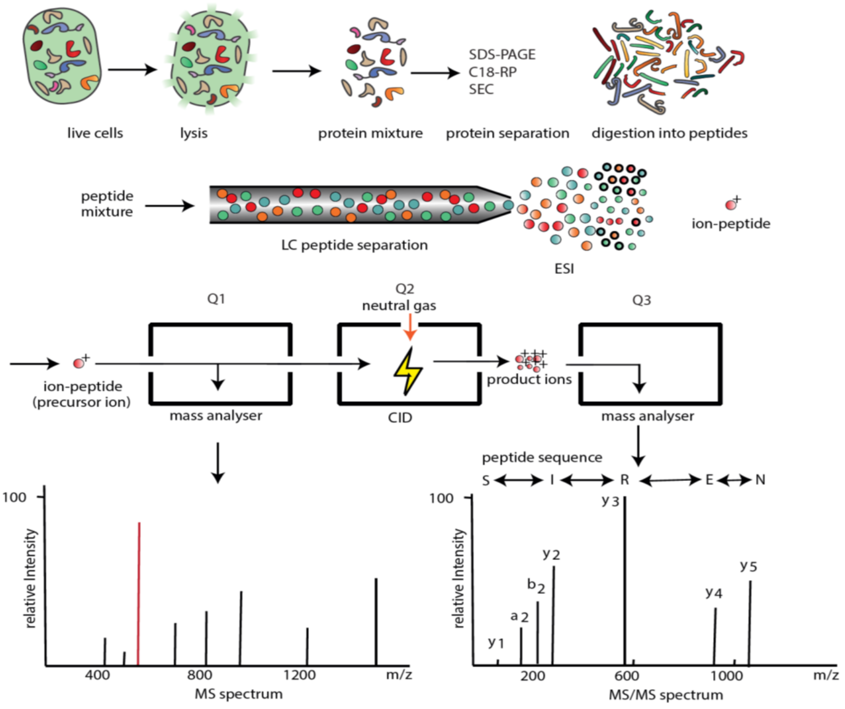


Figure 1: General workflow of DDA mass spectrometric identification of biomolecules. Identification of biomolecules by mass spectrometry generally follows a sequential process. First, cells or tissues of interest are homogenized and lysed to generate a complete cellular extract. Proteins are then separated by SDS-PAGE or chromatographic separation to reduce sample complexity. Separated samples are then digested into peptides, usually by trypsin and the resulting peptide digest is subjected to LC-chromatographic separation before ionized (using ESI for example). Ionized sample is then subjected to mass spectrometric analyses, usually a triple quadropole-type mass spectrometer. In quadropole 1 (Q1) precursor ions are analyzed by the mass analyzer and their respective m/z values are reported. Precursors can be selected in subsequent scans and enter Q2, where they collide with a neutral gas, resulting in fragments ions that are directed into Q3, where they are recorded. The resulting MS2 spectrum is then used to infer peptide sequence information based on characteristic shifts between the peak that correspond to specific amino acids. This approach is termed data-dependent acquisition (DDA) as fragmentation is only performed on selected precursor ions identified in a previous precursor scan. Spectral annotation is typically done by software as mass spectrometers can easily generate millions of spectra in one run. Figure adapted from Coon *et al.* [19]

Ultimately, peptides are found in aqueous solution, but need to be vaporized to enter the mass spectrometer as ions. Different ionization techniques such as electron impact (EI) ionization, Atmospheric Pressure Chemical Ionization (APCI), and Matrix Assisted Laser Desorption Ionization (MALDI) exist [20, 21]. Because the peptide mixtures are routine separated via C18-chromatography prior to being subjected into the mass spectrometer, an ionization technique from aqueous solutions needs to be applied [22]. Chapter 3.1.1 delineates the principle of electrospray ionization used in ionizing analytes from aqueous solutions in more detail. Vaporized ions then enter the mass spectrometer and are most commonly analysed in data-dependent acquisition (DDA) mode by tandem mass spectrometers. For DDA, the mass spectrometer consists of at least three quadropoles or ion traps to first record a survey scan (MS1) of all incoming ions in quadropole 1 (Q1) by an ion detector, recording the mass to charge ratio (m/z) of the detected ions. Then, the mass spectrometer can select a specific m/z ratio by adjusting the electromagnetic field of the quadropole after the survey scan has been recorded. The selected precursor ion is routed into Q2 (or an ion trap) for fragmentation by colliding with an inert gas such as N_2 in a collision chamber. This collision-induced dissociation (CID), or high-energy collisional-induced dissociation (HCD) in Orbitrap instruments with a C-trap, fragments the precursor ion into fragment ions. Fragment ions are then routed into Q3 or an Orbitrap where their specific m/z values are recorded (MS2). Since every MS2 scan is dependent on precursor ions recorded in MS1 scans, DDA provides unambiguous identifications of analytes if they are well time-resolved. This means that the mass spectrometer must be either fast enough to record many MS2 scans after one survey scan to generate MS2 scans for all analytes entering the mass spectrometer, or that incoming analytes display a reduced complexity over time, granting the mass spectrometer enough time to record MS2 spectra for all analytes without missing out on analytes detected on MS1. Oftentimes, sample complexity is too high, and the mass spectrometer is set to a fixed number of MS2 scan following an MS1 scan (commonly set to n = 20, Top20) that excludes any precursor mass from MS1 past the most intense top 20 peaks. This undoubtedly leads to loss of information, especially from less-intense precursor peaks that stem from an ion package of fewer ions, warranting the development of other orthogonal separation techniques to reduce complexity and increase sensitivity such as high-field asymmetric wavelength ion mobility mass spectrometry (FAIMS) discussed in chapter 3.2. This approach to study a protein by cleaving it into smaller peptides first, and to study those in detail to deduce the original complex protein, is called "bottom-up" proteomics in contrast to "top-down proteomics", where entire proteins are subjected to the mass spectrometric analyses, skipping the protein digest [23, 24]. Bottom-up proteomics works exceptionally well for highly modified proteins, carrying multiple different post-translational modifications (PTMs) that add an immense level of complexity to analyzing the protein by mass spectrometry [25]. Not only does one need to distinguish between different PTMs, but their positions is may also be variable across proteins, a feat that is currently best solved by bottom-up proteomics, where individual peptides with their respective PTM modifications can be reliably analyzed. All studies performed in this thesis are of a "bottom-up" nature because of the interest in localizing the modification site unambiguously, requiring enzymatic digestion of the protein into smaller peptides.

3.1.1 Electrospray ionization

Electron spray ionization (ESI) was invented in 1988 as a soft ionization technique by John Fenn and colleagues, transferring entire proteins into the gas phase by applying ESI [22]. Molecules are transferred from the liquid aqueous phase into the gas phase in an endoergic desolvation process, thereby ionizing the analyte. Peptides generated by tryptic cleavages provide excellent physicochemical properties to facilitate desolvation and ionization because they are naturally positively charged (tryptic cleavage restriction) in the acidified solution and are typically 8 amino acids long on average [22, 26, 27]. Analytes pass the transfer tube and enter the capillary where high voltages are applied to polarize charges in solution, forming a cone at the tip due to electrostatic Coulomb forces between charges and the surface tension of the liquid [22, 27]. Applying the proper Taylor Cone Voltage to the needle results in a threshold of charge density to be reached (Releigh limit), resulting in small droplets are emitted from the Taylor cone, which is also termed Rayleigh fission [22]. Emitted droplets evaporate quickly as the electrode is heated, leaving soluble charged ions to ultimately enter the gas phase [22]. There are different theories as to how those droplets are formed, namely the single ion droplet theory (SIDT) and the ion evaporation model proposed by Dole *et al.* and Iribarne *et al.*, respectively, that can be reviewed elsewhere [28, 29].

3.1.2 Orbitrap mass spectrometers

All mass spectrometers used in this study used a combination of mass analyzers, namely quadropole mass analyzers, linear ion traps (LIT), and orbitraps (OT), all used for different aspects of the MS-analysis of analytes. The Q consists of four rod-shaped electrodes to which a radio-frequency (RF) and a direct current (DC) is applied, creating an electric field that can be altered by manipulating RF and DC [30]. Selection of precursor masses post MS1 survey scan is achieved by altering RF and DC in such a way that only the selected m/z precursor ratio is directed through Q1 with a stable trajectory, while all other precursors collide with the electrodes, discharging [30]. A schematic of one of the most recently developed OT mass spectrometers, the OT Exploris 480, is depicted in figure 2. Here, vaporized ions enter the mass spectrometer through the transfer tube and are focused into a single beam in the electrodynamic funnel. An Advanced Active Beam Guide (AABG) is essentially a routing Q that is slightly bend to filter out uncharged particles that exit the ion funnel by letting them collide with the electrodes, whereas ions are routed by the electric field through the bend AABG. The classical Q1 is called advanced quadropole technology (AQT) and serves as the selecting Q post survey scan, where all ions are detected by the detector. Selected ions are transfered through a special LIT, the C-trap, before fragmented in the ion-routing multipole (collision chamber, Q3) that also redirects fragmented ions back into the C-trap. The C-trap can store ions for some time before ejecting them into the OT, allowing parallelization of precursor selection in Q1, fragmentation in the Q2, and MS2 recording in the OT [30, 31]. The orbitrap mass analyzer is comprised of two electrodes: one external electrode setting boundaries to ion trajectories in the periphery, and an internal electrode that resembles a spindle [30, 31]. Injected fragment ions enter the OT as an ion package from the C-trap and adopt a spiral trajectory around the internal electrode. Ion trajectories induce currents within the internal electrode that can be measured and Fourier-transformed into the appropriate m/z ratio [30, 31].

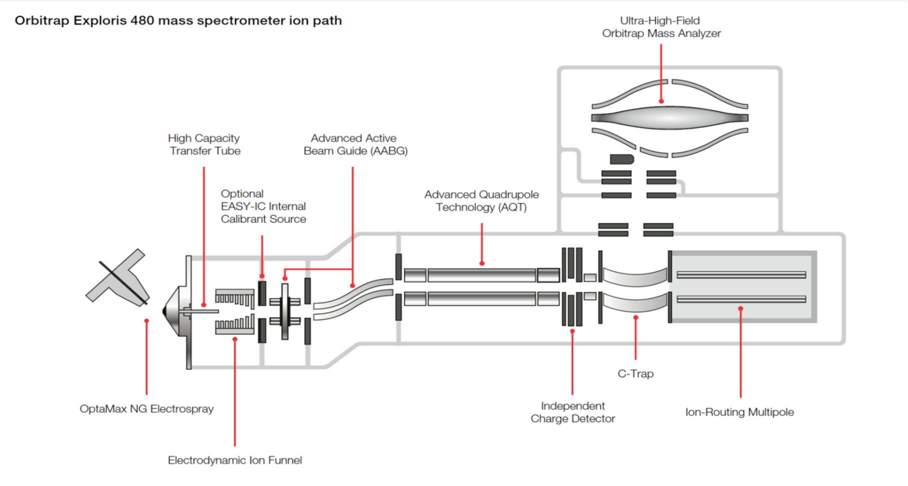


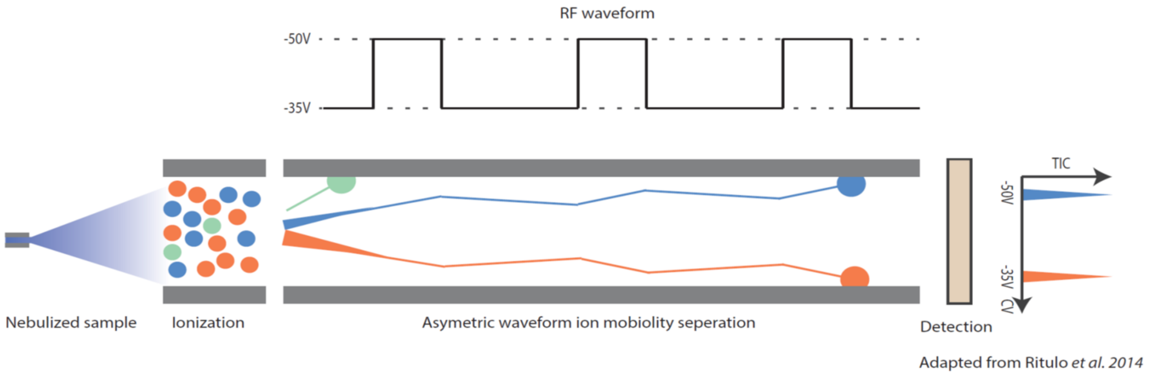
Figure 2: Schematic of the OT Exploris 480 mass spectrometer. Thhermo Fischer's OT Exploris 480 mass spectrometer contains an orbitrap (OT) mass analyzer that was further developed into an untra-high field OT mass analyzer that can achieve resolutions of up to 480,000 at m/z = 200. The general setup of the Exploris is similar to the OT Lumos instruments, Fusion instruments, and HF-X instruments. Ionized sample is routed through an ion funnel, focusing incoming precursor ions before being routed through the beam guide into Q1. Selected precursors from survey scans (DDA) can be routed into Q2 via the C-trap. Q2 is an ion-routing multiple used for higher-energy collisional dissociation (HCD) of precursors that results in fragments ions that can be stored in the C-trap prior to determining their m/z values in the OT. The C-trap allows parallel fragmentation in the ion-routing multipole while fragmented ions are being analyzed in the OT and a bundle of additional ions is waiting in the C-trap to be routed into the OT. The C-trap and ion-routing multipole could also be used in MS^n experiments, where fragment ions are sequentially selected in the C-trap and re-fragmented in the ion-routing multipole before ultimately subjected to the OT. Image taken from Thermo Fischer's OT Exploris 480 website [32]

OT mass analyzers are "high resolving" mass analyzers, meaning that small differences in m/z ratios can be distinguished [31]. This is important as tryptic peptides are 8 amino acids long on average, but can easily surpass 50 amino acids [27, 31]. A precursor may then have charge states ranging from 2-8 charges [27]. Since naturally occurring isotopes of carbon and nitrogen, 13C and 15N, respectively, are incorporated into peptides at low frequencies, the resulting change in mass, the delta mass, is +1 Da. A corresponding m/z ratio varies from charge state +2 being 0.5, to charge state +3 being 0.33. and so forth. As stable isotopes usually occur at a frequency of 1%, the resulting peaks are small in comparison to the monoisotopic peak, comprised of all 12C and 14N [27]. As charge states increase, additional peaks corresponding to the diisotopic species decrease in distance on the m/z scale and thus require high resolving power.

3.2 (high) field-asymmetric waveform ion mobility spectrometry (FAIMS)

High-field asymmetric waveform ion mobility spectrometry (FAIMS) is an ion-mobility technique that separates gas-phase ions at atmospheric pressure (760 Torr) and room temperature, first described in 1993 in combination with ESI [21]. Ion-mobility mass spectrometry (IMS) is an developing field with a wide range of applications. IMS enables researchers to not only gain information on the mass of the analyte, but it can separate isomers, isobars, and conformers in addition to measuring ion size while increasing the signal-to-noise ratio (S/N ratio) [33]. IMS features an additional dimension of data points along the gradient by separating structurally similar ions and ions of the same charge state into families of ions that share the same mass-mobility correlation line when plotted [33]. Nowadays, there are four different types of IMS that can be used: i) drift-time ion mobility spectrometry (DTIMS), ii) aspiration ion mobility spectrometry (AIMS), iii) differential-mobility spectrometry (DMS), also called (high) field-asymmetric waveform ion mobility spectrometry (FAIMS), and iv) traveling-wave ion mobility spectrometry (TWIMS) [33]. Of the four IMS methods, DTIMS exclusively measures collisional cross-sections correlating to the shape of transferred analytes, and features exceptionally high resolving power.[33]. DMS and FAIMS, on the other hand, offer continuous-ion monitoring capability as well as orthogonal ion mobility separation, offering exceptionally high-separation selectivity mainly used in preparative IMS [33]. TWIMS is the newest method, featuring low resolving power, good sensitivity, and easiest integration into commercially available mass spectrometer.

A FAIMS instrument is placed in front of the ion source of the mass spectrometer and acts as an ion filter that can be set to continuously transmit one type of ion [21]. Figure 3 illustrates the principle of FAIMS for three different nebulized ion species that enter the FAIMS device. Gas-phase ions enter space between the two electrodes of the FAIMS device, being consequently subjected to the alternating electric field [21]. A rapidly alternating radio frequency wave (RF wave) is applied to the electrodes, altering ion trajectories dramatically. [21]. The green ion in this case is immediately discharged as its trajectory is changed too much towards the electrode. A second compensation voltage (CV) is applied and counteracts the strong first alternating voltages in some degree, pulling ions back towards the center of the drift tube [21]. By alternating the two CVs, ions are pushed and pulled depending on the voltages applied and only ions whose physicochemical properties allow for minimal deviations from a perfect trajectory at those two CVs can pass the drift tube without being discharged and enter the mass spectrometer to be analyzed. [33]. FAIMS thus greatly diminished the heterogeneity of incoming ions eluting from the LC-system by removing contaminating solvent ion clusters, and grouping ions into ion families that transverse the drift tube given set CVs [21].



Adapted from Ritulo *et al. 2014*

Figure 3: Schematic of FAIMS ion separation. High field asymmetric waveform ion mobility spectrometry (FAIMS) is an additional molecular filtering technique prior to MS-analysis. This atmospheric pressure ion mobility technique separates gas-phase precursor ions by their behavior in strong and weak electric fields applied by a circular electrode in front of the ESI inlet into the mass spectrometer. The applied radio frequency waves alter the trajectory of the trans-versing ions through the circular electrode. Only ions with adequately altered trajectories based on applied counter voltages (CVs) can pass through unimpededly. The two compensation voltages illustrated here, are -50 V and -30 V and allow ions of the orange type and the blue type, but not the green type to pass through the electrode to enter the mass spectrometer. FAIMS can increase peptide identification by providing orthogonal selectivity that enhances signal-to-noise ratios for selected ion species that would otherwise be missed.

The molecular ion filtering application of FAIMS is ideal to remove unwanted ions such as singly charged ions that do not stem from peptide precursor ions that must be doubly charged when cleaved by trypsin. Removing low-mass solvent cluster ions results in an improved S/N ratio for mass spectra of peptides compared to conventional ESI-LC-MS [21]. Additionally, FAIMS allows to identify differing distinguishable charge states of an entire protein, namely cytochrome c, alluding to the fact that the the high electric field strength induces ion mobility of these species that is sensitive to the structure of the protein ion [21]. In addition to more classical IMS, FAIMS has also been used extensively to enhance proteome coverage by identifying about 10% more peptides from a sample, as well as boosting identifications of crosslinking sites from protein-protein crosslinked samples [34–36]. There is one caveat to enhance selectivity that somewhat limits the wide applicability of FAIMS: it requires a constant flow of clean, dry nitrogen gas, usually 5 L per minute.

3.3 Fragmentation of Biomolecules: peptides and RNA

Selected precursor masses from MS1 survey scans can be isolated and routed into the collision chamber for fragmentation via CID HCD in OT mass spectrometers. Fragmentation of the precursor ion into smaller fragment ions that are subsequently analyzed in the OT and recorded provide specific clues as the structure of the precursor ion based on specific fragmentation patterns for various biomolecules. Peptides break at very specific sites within the peptide backbone during HCD, generating specific fragments that correspond to primary amino acids that constitute the peptide, allowing computer algorithms to quickly deduce the peptide sequence from specific fragment ions.

Figure 4 A) illustrated how peptides are fragmented according to Roepstorff *et al.*[37]. Based on this nomenclature, breaking the peptide backbone in three different places results in three different types of ion pairs, denoted as a/x-ions, b/y-ions, and c/z-ions [37]. The fragments containing the amino-terminus are called a, b, and c-ions, while fragments containing the carboxy-terminus are called x, y, and z-ions [37, 38]. In HCD fragmentation of peptides, b/y-ions are most typically observed, albeit a-ions also appear frequently, but in fewer numbers. Those a-ions exhibit a prominent relationship from b-type ions that marks them as a/b-ion pairs by the loss of a C=O group, corresponding to a specific shift of 27.9949 Da.[37, 39]. Additionally, the loss of water or ammonia is most commonly observed in peptide fragmentation, resulting in shifts of 18.01056 Da and 17.02655 Da, respectively [39]. Since all fragment ions carry the amino acid side chain and only differ in their termini, specific mass shifts for all 20 canonical amino acids enable deduction of the peptide sequence from lattering ion serious that only differ by the specific side chain shift [38]. By utilizing trypsin as the proteolytic enzyme, the C-terminus is fixed to either be an arginine or a lysine residue, setting the corresponding y1-ion to either 175.1190 m/z or 147.1128 m/z, respectively. Figure 4 B) illustrates how RNA is can be fragmented at the phosphodiester bond, leading to four specific fragment ions. Using the nomenclature proposed by Domon *et al.*, generated ion pairs are denoted a/w-ions, b/x-ions, c/y-ions, and d/z-ions [39]. Fragment ions containing the 3'-hydroxyl end of the RNA are termed w, x, y, and z-ions, whereas fragments containing the 5'-hydroxyl end are termed a, b, c, and d-ions [39]. In HCD fragmentation of peptide-RNA heteroconjugates under peptide-centric conditions, MS2 spectra most often contain a diagnostic marker ion of the nucleobase alone, or its nucleosidic/nucleotidic equivalent, resulting in y or w-ions.

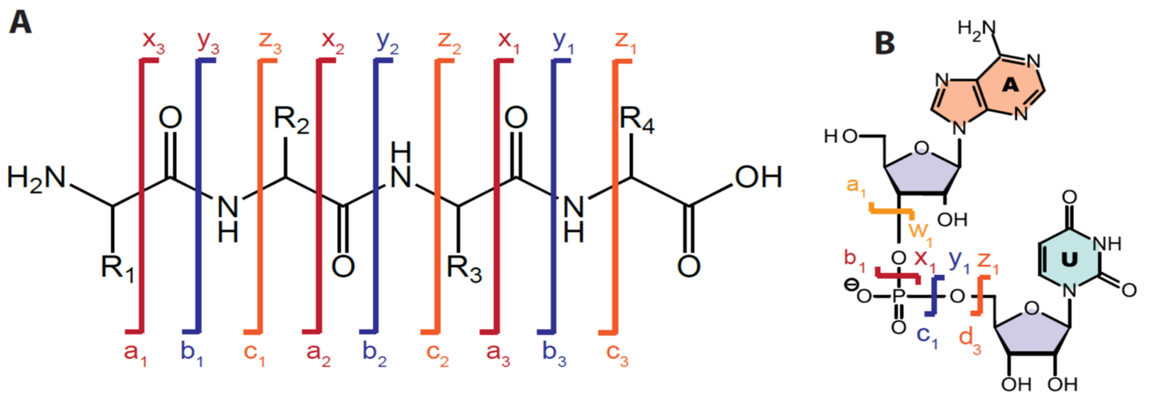


Figure 4: Schematic of peptide and RNA fragmentation. A) The peptide bond between two adjacent amino acids can be broken in three different ways: i) breaking it between the C_α and the carbonyl-carbon results in a/x ions. ii) breaking the peptide bond between the carbonyl-carbon and the amide nitrogen results in b/y ions. iii) breaking the peptide bond between the amide nitrogen and the next C_α-atom results in c/z ions. Ions containing the N-terminal part of the peptide are called a, b, or c ions, whereas ions containing the C-terminal part are called x, y, and z ions. In HCD, the peptide bond usually breaks between the carbonyl-carbon and the amide nitrogen, resulting in b and y ions. **B)** The RNA phosphate-sugar backbone also fragments during HCD in four different ways. i) A nucleoside loss resulted from breaking on either side of the phosphoester link generates d/z ions when the oxygen of the phosphoester is also lost, and ii) c/y ions when an intact nucleoside is lost. iii) If the second phosphoester-oxygen is lost, b/x ions result in an incomplete monophosphote attached to a nucleoside. Complete loss of a monophosphate nucleotide results in a/w ions. Ions incorporating the 3'-end of the RNA are called w, x, y, and z ions, whereas ions incorporating the 5'-end are called a, b, c, and d ions. MS2 spectra of crosslinked peptide-RNA heteroconjugates often contain nucleobase marker ions and sometimes marker ions corresponding to entire nucelosides and nucleotides, the y, and w-ions, respectively.

Besides CID and HCD, there are numerous other fragmentation techniques available that are reviewed elsewhere and were not used in this study [39, 40]. Since CID usually leads to fragmentation at the most labile chemical bond, PTM analyses with very labile modification such as histidine phosphorylation are usually best carried out using HCD because the higher amount of energy applied to the colliding molecules leads to peptide backbone fragmentation and not the loss of the PTM [39]. HCD fragmentation predominantly resulting in a, b, and y-ions in additional to the aforementioned restriction in y and b-ions for tryptic peptides to be either Lys or Arg, allows for rapid identification of peptide sequence from generated fragment ions by computer software.

3.4 Computer-based data evaluation of mass spectrometric data

Spectral data of peptide sequencing information from mass spectrometric analyses are nowadays analyzed by computer software, dealing with millions of MS1 and MS2 spectra. As such over 30 different search algorithms, differing in their bioinformatic approach to index spectral data, parallelization of computational processes, quantification of features, times to complete spectral annotation, handling of various PTMs, and most importantly, statistical meta-analyses to reduce the number of false positive identifications [41]. To identify proteins via mass spectrometry, recorded MS spectra are searched against a protein database, containing all possible proteins of interest by performing an *in silico* digest of the proteins in the database. By specifying what enzyme was used to digest the actual sample, the computer generates a list of all possible peptide fragments that fit the given enzymatic restrictions. Common modification of the peptide such as alkylation of cysteines or oxidation of methionine can also be specified, allowing greater coverage of highly modified peptides. Theoretical spectra are then calculated for all entries on the list of possible fragments and recorded spectral data is compared to the theoretical spectra. Differing statistical analyses determines if an actual spectrum matches with a theoretical spectrum and a peptide spectrum match (PSM) is reported. Multiple PSMs are aggregated into protein groups to which the peptide fragments match, allowing to determinedly conclude the presence of a certain protein in the sample.

One of the most common search engines was developed in 2011 by Cox *et al.* and has since been enjoying immense popularity and constant improvement: the search engine Andromeda, implemented in MaxQuant [42, 43]. MaxQuant's Andromeda search engine also features label-free quantification based on extracted-ion chromatograms that can add reliable quantification results of identified peptides to every run analyzed [42]. Similarly, Proteome Discoverer from provides a commercial alternative to MaxQuant with the robust search engine Sequest and numerous additional nodes for various analyses [44]. Since data presented in this thesis is mainly of qualitative nature, aspects of quantification have largely been left out, even though label-free quantification has been used to estimate protein abundances in some cases as indicated in the chapters below. Additionally, more complex analyses such as determining a crosslinking site in either protein-protein or protein-nucleic acid crosslinking requires specialized software such as pLink2 or RNP$_{\mathrm{xl}}$ that will be discussed in chapter 3.10 [45, 46].

3.5 Supervised machine learning algorithms for classifying mass spectrometric data

The field of machine learning is vast and has most influenced all aspects of scientific research in one way or another. In mass spectrometry, recent advances in mass spectrometry-based proteomics have enabled tremendous progress in the understanding of cellular mechanisms, disease progression, and the relationship between genotype and phenotype by novel machine learning approaches used to classify meta-data [47]. Meta-analyses cluster identified features in various ways to visualize findings such as clustered abundances, time-course specific profiles of proteins, expanding and integrating identified features into extensive knowledge data bases such as pathways, functional annotation, and structural annotation [47]. Overall, these integrated approaches aid in identifying concepts, mechanisms, and insights on a proteomic level. Other mass spectrometric fields such as metabolomics and imaging mass spectrometry are also increasingly utilizing machine learning to integrate metabolites into pathways and classify images from tissues analyzed by MALDI-imaging [48].

Machine learning approaches can roughly be divided into two sections: i) supervised machine learning, and ii) unsupervised machine learning. While unsupervised machine learning with its powerful neural networks and support vector machines, is rightfully the preferred approach to complex data, it requires substantial knowledge on the topic and excellent coding abilities that would be best applied in a separate thesis for candidates of bioinformatics. Supervised machine learning, on the other hand only requires a solid understanding of statistics and moderate coding abilities. Additionally, supervised machine learning easily formulates insights that can be understood intuitively, and supervised machine learning is bound by input parameters and dictated algorithms that are easier to understand [49]. In any machine learning project, a sufficiently large portion of the data set needs to be allocated as a training set upon which the algorithm learns the data and deduces classification features. The large data set that is usually split in an 80/20 ratio, allocating 80% of the data being to a training data set and 20% of the data to a testing data set, used in evaluating model performance after machine learning [49, 50]. Here, the instance-based algorithm k-nearest neighbor (k-NN), the C5.0 decision tree algorithm, and the rule-based algorithm repeated incremental pruning to produce error reduction (RIPPER) are all commonly used to classify multivariable data [49, 50].

3.5.1 k-nearest neighbor

k-NN is also called a lazy learning algorithms, as the generalization process is skipped entirely and no statistical model is actually built [51]. As a positive feature, it requires less computation time during the training phase compared to eager-learning algorithms (such as decision trees or rule-based algorithms) but more computation time during the classification process of unseen data [51]. k-NN is based on the principle that the instances within a data set will generally exist in close proximity to other instances that have similar properties [51, 52]. k-NN looks at all features, variable associated with a classification outcome, and compares the classifying label of an unclassified data point by observing the classifying labels of its neighbors [51]. k-NN then determines the nearest instances to the unclassified data point and assigns the same classifying label to it [52]. If data point separate well into distinct regions across the different parameters, k-NN provides a very fast and effective algorithm to accurately classify data "of the same kind", having very close neighbors. A more thorough review on k-NN and its applications can be found elsewhere [52]. In classifying spectral data from protein-RNA crosslinking, spectral data must only be classified as a true positive hit, a spectrum displaying enough information to confirm the presence of an RNA chemically attached to the sequenced peptide, or not.

3.5.2 C5.0 decision trees

Decision tree algorithms such as C5.0, however, aim at identifying features that best divides the training data, and the single most effective feature in doing so is called the root node of the tree [49]. It may be somethings such as the number of amino acids within a peptide that best predicts a classification event. Having identified the root node, the remaining partition is consecutively divided into sub-trees by the same method until the training data is fully classified [49]. Decision trees can thus, perfectly classify the training data set with copious amounts of sub-trees that are often too specific and individualized to hold up in unclassified data [49]. As a consequence, overfitting quickly becomes a problem. to overcome this problem, there are two commonly applied measures that can reduce overfitting: i) stopping the training algorithm before it reaches complete classification, limiting the number of sub-trees to generate, and ii) prune the decision tree as necessary once completed [49]. Decision trees take longer computation times, as pruning and evaluation against a testing validation set iteratively removes unnecessary nodes that do not lead to significant information gain [49]. It is quite easy to generate multiple decision trees on large data sets and then to select the tree with the fewest leaves, resulting in a tree that classifies "most of the testing data" accurately.

3.5.3 Rule-based RIPPER algorithm

Lastly, the RIPPER algorithm is one of the best known rule-based algorithms utilizing a repeated process of repeated growing and pruning akin to C5.0 decision trees [49]. It formulates very restrictive rules during the growing phase similar to un-pruned decision trees to maximize fitting the data, but relaxes those rules during the pruning phase repeatedly to reduces instances of overfitting [49]. RIPPER classifies data points into more than two classes by ordering them in based on their prevalence and then treating them in order as a distinct two-class problem, sorting features into a potentially true class, and a potentially false class [49]. RIPPER has been successfully used in genomics to classify hundreds of gene expression profiles of embryonic cells in lead to discovery of novel biomarkers. [53]. A more comprehensive description of RIPPER can be found elsewhere [49]. Those three classification algorithms may also be useful in peptide-centric mass spectrometric approaches to identifying peptide-RNA crosslinking sites.

3.6 Ribonucleoproteins

RNA-protein interactions play an important role in a myriad of cellular processes. Most notably, RNA-protein interactions are the driving force of ribosomal mRNA translation into proteins, transcription and post-transcriptional regulation also depend heavily on protein-RNA interactions. Ribonucleoprotein (RNP) complexes, comprised of RNA and protein, exert tremendous influence on mRNA-splicing, mRNA-polyadenylation, mRNA-transport, mRNA stability and ultimately mRNA translation [54]. Clinically, RNA also plays a central role in many human diseases, particular through non-coding RNAs (ncRNAs). Disruption of homeostatic microRNAs (miRNAs), small nucleolar RNAs (snoRNAs), and large intergenic non-coding RNAs (lincRNAs) have been shown to also play a role in neurological, cardiovascular, and developmental diseases in addition to processes leading to tumorgenesis [55, 56]. Increasing attention is also paid to circular RNAs (circRNAs) that are increasing with age and seem to be related to neurological diseases such as Alzheimer's disease [57]. Consequently, interest in identifying protein-RNA interactions is ample and crosslinking coupled with mass spectrometry provides a powerful tool to study such interactions.

Structurally, it has been known for years now, that RBPs regulate gene expression by binding to different regions of mRNAs[58]. In this fashion, RBPs bind to 5' and 3' untranslated regions (UTRs) and the coding region, masking those RNA sequences if bound and unmasking regions such as the ribosomal binding site (RBS) if unbound [59]. Moreover, it seems that not only does the mRNA exhibits dynamic behavior, but also the RNP itself: the splicing complex consists of multiple RNPs that undergoes massive restructuring events such as factors associating and disassociating from the spliceosome, as well as specialized ATP-dependent RNA helicases such as the DEAD-box helicases that are essential enzymes remodeling RNPs temporally [60, 61]. The spliceosome has been studied well over the years, elucidating how posttranscriptional excision of introns from pre-mRNA generates a fully functional, protein-encoding mRNA transcript that only contains exons. Interestingly, RNPs are shown to engage in liquid-liquid phase separation through their intrinsically disordered regions (IDRs) that leads to the formation of granules in various cells in a stage-dependent manner that remains to be fully understood [62–64].

RNA-binding proteins (RBPs) can be roughly sorted into two groups based on the role the RNA plays in the protein-RNA complex, illustrated in figure 5 In one instance, unstructured RNAs are properly folded into a secondary/tertiary RNA structure by protein chaperones that also help in retaining proper RNA structure [58, 65]. One of the best known examples for RNA chaperones from *E. coli* is the cold-shock protein CSP-A, discussed below or RNA-helicases [66, 67]. Instead of the protein acting upon the RNA and inducing biological function, the RNA itself can direct protein function and thus confer biological activity. As such, the RNA provides a structural framework for the protein, leading the protein to its target [58]. miRNAs for example bind to various target mRNAs and recruit proteins of the Argonaute familiy that bind to the guiding RNA (gRNA) and then degrade the target mRNA, silencing gene expression post transcription by preventing mRNA translation in eukaryotes [58]. Interestingly, prokaryotic Argonaute proteins such as *Tt*Ago targets ssDNA and leads to foreign DNA degradation presumably as a defense mechanisms against bacteriophages by *E. coli*.

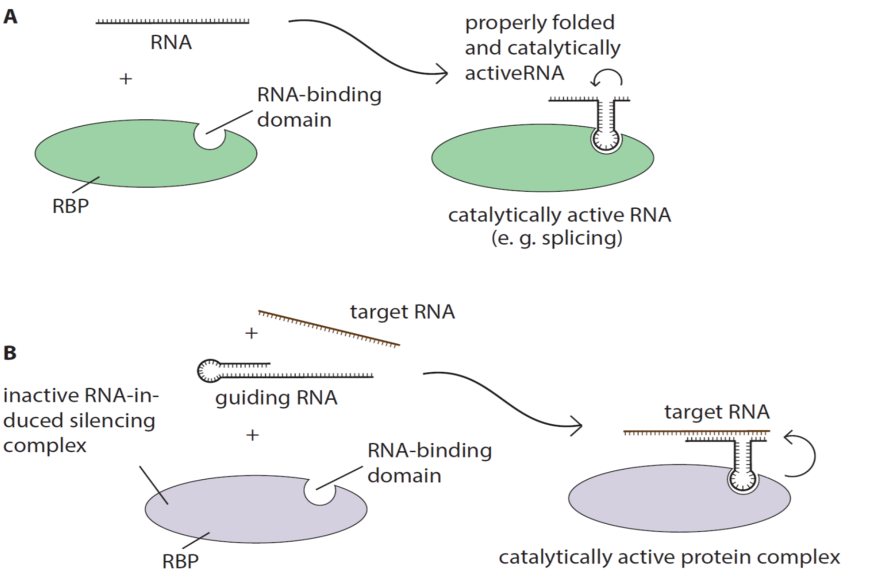

A
RNA
+
RNA-binding domain
RBP
properly folded and catalytically activeRNA
catalytically active RNA (e. g. splicing)
B
target RNA
+
inactive RNA-in-duced silencing complex
guiding RNA
+
RNA-binding domain
RBP
target RNA
catalytically active protein complex

Figure 5: Two types of RBPs. RNA-binding proteins (RBPs) fall within two distinct categories. **A)** an (partially) unstructed RNA molecule is assumes its proper, biologically active seconday/tertiary structure upon binding to an RBP. In this case, the proteins induces proper folding and confers biological activity. RNA chaperones serve as a good example for this kind of RBP, as they induce and maintain proper RNA secondary/tertiary structures. **B)** Alternatively, the RNA may already have assumed its proper secondary/tertiary structure that is recognized by the RBP and binds to it. Here, the RNA confers biological functions by juxtaposing enzymatically acitve parts of the protein to its substrate. Members of the Argonaut protein family for instance bind miRNAs to silence mRNAs and regulate gene expression. Adapted from Hentze *et al.*[58].

3.7 RNA-binding motifs

RBPs are very diverse both functionally and structurally. Structurally, RBPs need to accommodate RNA secondary and tertiary structures and may even be directly involved in superstructure formation. While the DNA-double helix stably assumes its double helical conformation, RNAs can build a plethora of superstructures, comprised of loops, sheets, helices, barrels, and turns that are described in more detail by others [68, 69]. There are over 2,000 known RNA-binding proteins, some of which can be classified as canonical RBPs based on known RNA-binding domains, and a large number of non-canonical RBPs present in many ribosomal proteins binding to rRNA, with still unidentified RNA-interacting regions [62]. Certain protein domains have been identified over the years to be primarily RNA-binding, of which the RNA-recognition motif (RRM) is the best studied example of an RNA-binding domain across all species [70, 71]. In addition to the RRM, other RNA-interacting motifs such as the K-homology domain (KH-domain), the zinc finger proteins, the cold-shock domain (CSP), the ribosomal protein S1-like domain, and the double stranded RNA-binding domain (dsRBP) have been identified as some few canonical structural RNA-binding domains (figure 6. Each of the domains described above is briefly characterized in the following sections.

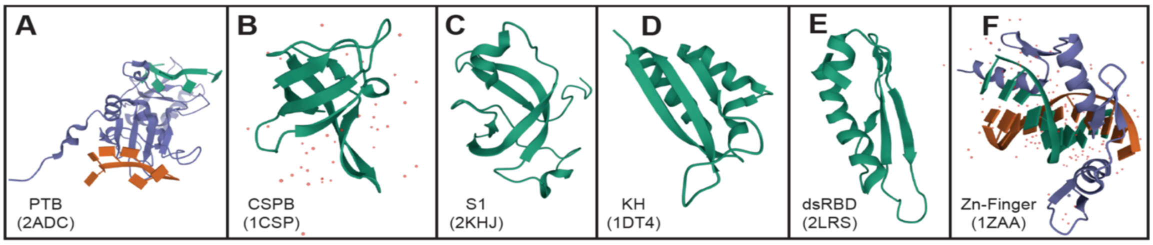


Figure 6: Examples of well-characterized RBPs. A) the polypyrimidine tract binding protein (PTB) contains two RNA-recognition motifs (RRMs) that are involved in binding various pre-mRNAs, as a regulator of alternative splicing and subsequent RNA processing steps. The RRM domain consists of four β-strands and two α-helices arranged in an alpha/beta sandwich. Some RRMs also contain a third α-helix. RNA is depicted in orange and in green [72]. **B)** cold-shock protein CSPB contains the cold-shock protein RNA binding domain which was originally identified in prokaryotic organisms as an RNA chaperone. Three anti-parallel β-sheets confer RNA binding properties and the CSP domain is also found in eukaryotic proteins [73]. **C)** Ribosomal protein S1-like RNA-binding domain, identified first in ribosomal protein S1 constitutes a five-stranded anti-parallel β-barrel, present in many RNA-associated proteins and is similar to the CSP domain [74]. **D)** The K-homology domain consists of two α-helices connected through a GXXG-loop and additional flexible loops and is found in all domains of life [75]. **E)** The double-stranded RNA-binding domain dsRBD features a pentameric secondary structure of αβββα-topology and is involved in post-transcriptional gene regulation [76]. **F)** The Zinc-finger domain binds both DNA and RNA and is quite versatile. It consists of a short β-hairpin and an α-helix, sometimes embedded into a ββα-structure. Two cysteines and two histidines side chains fixate a central zinc-ions into a tetrahedral array, hence its name zinc-finger domain. RNA is depicted in orange and in green [77].

3.7.1 The RNA-recogniation motif

Undoubtedly one of the best described RNA-binding domains, is the RRM domain that is even estimated to be present in about 1% of all eukaryotic proteins, illustrating the diverse role RNA plays in many cellular processes [62]. It comprises of a conserved amino acid sequence of 90 residues that folds into a characteristic β1α1β2β3α2β4 topology [62, 70, 71]. The four-stranded β-sheets are packed against two α-helices, forming a plane that is the prime mediator of RNA-interaction (figure 6 A) [71]. This fold is highly conserved across species and was identified to bind pre-mRNAs and mRNAs in many heterogeneous nuclear ribonucleoproteins (hnRNPs) such as the polypyrimidine tract binding protein (PTBP1) [72]. PTBP1 contains four RRMs, binding primarily to the nucleobases U and C in pyrimidine motifs, where binding is mediated by surface-exposed aromatic residues such as Tyr or Phe that are part of β-strand1 and β-strand3 [71, 72]. Binding of PTBP1 of mRNAs occurrs both in the nucleus and in the cytoplasms, where PTBP1 acts as a regulator of alternative splicing inside the nucleus, as well as regulating mRNA stability, mRNA localization, poly-adenylation, and mRNA translation in the cytoplasm [72]. The RRM motif is fairly rigid in its topology, assuming a stable confirmation that can be crystallized. Both Hsh49 and NELF-E discussed below contain RRMs that could be crystallized, while the flexible IDRs of NELF-E evaded crystallization efforts.

3.7.2 The cold-shock domain

Another evolutionarily diverse nucleic acid-binding protein domain is the cold-shock domain (CSD). It seems to have originated in prokaryotes as an RNA-chaperoning domain, ensuring proper RNA secondary/tertiary structure of different RNAs, but evolved into a regulatory domain of euaryotic translation [66, 78, 79]. Akin to the RRM, CSD domains contain about 70 highly conserved amino acids that have been found to be binding to single stranded RNAs in both prokaryotic and eukaryotic systems [79]. In *E. coli* the cold-shock protein CSP-A is the major cold-shock response protein to adapt to cold temperatures by enhancing transcription of CCAAT-containing promoters found in the genes for gyrase and histone-like nucleoid structuring protein HN-S protein [66, 80, 81]. As a response, gyrase ensures that chromosomal dsDNA is not negatively supercoiled in *E. coli*, leaving chromosomal DNA in an unwound, actively describable state [81]. Another example of CSD is the prokaryotic transcription termination factor Rho that has been extensively analyzed by NMR, revealing that the RNA-binding domain is more similar to a CSD domain than a canonical RRM [82]. Structurally, the CSD assumes as an anti-parallel, five-stranded β-barrel with connecting loops of variable length (figure 6 B)) [82].

3.7.3 The ribosomal protein S1-like domain

As the name suggests, the ribosomal protein S1-like domain was first identified in ribosomal protein S1, but has since been identified in numerous other RNA-binidng proteins, making up a diverse RBP family [83, 84]. A well documented example of a ribosomal protein S1-like domain is polynucleotide phosphorylase (PNP) from *E. coli* which structure has been elucidated by NMR analyses: A five-standed anti-parallel β-barrel displays conserved amino acid residues of various types on the distal side of the barrel, and conserved residues in the connecting loops adjacent to the barrel that mediate protein-RNA interactions on the medial side [85]. Structural similarities to the CSD domain are apparent and may hint at an evolutionary lineage of conserved RNA-binding domains, of which CSD is thought to be older [85]. Another less commonly known S1-like domain carrying RBP is NusA [85]. NusA plays an important role in transcriptional pausing by DNA-dependent RNA-polymerase RNAP [86]. Therefore, NusA plays a critical dual role in stimulating transcriptional pausing and transcription termination [86].

3.7.4 The K-homology domain

Similar in legth, the K-homology domain (KH-domain) consists of 70 highly conserved amino acids found in a diverse number of nucleic acid-binding proteins such as the heterogeneous nuclear ribonucleoproteins (hnRNPs) [87]. The three homologous KH-domains were first identified in the human hnRNP K that functions both as an enhancer of gene transcription and as a repressor of gene transcription [88]. Protein regulation is achieved by phosphorylation, altering protein function from mediating nucleic acid binding activity to providing a protein scaffolding platform for other signaling proteins [88, 89]. Such functional interconversion of proteins is not uncommon, especially not if interconversion is mediated by phosphorylation that is known to acitivate/inactivate many proteins as an easily reversable PTM [90]. hnRNP-K has been identified as an oncogene, abberantly expressed in cancerous tissues with expression levels soaring high above homeostatic levels [87]. Unfortunately, high levels of hnRNP-K are correlated with to poor prognoses in different cancerous diseases, indicating a degree of malignancy and thus serving as a diagnostic marker [87]. Since KH-domains can be found in multiple copies, it has been found that multiple copies of the KH-domain are functioning both cooperatively binding to the same (AU-rich) target sequence and independently of each other, binding to different regions of the target RNA [87]. Crystallographic structures show reveal a βααββα-topology in various proteins, spanning from prokaryotes (PNPases) over archeal exosome subunits to prokaryotic/eukaryotes small ribosomal protein S3 (figure 6 D)) [84, 87, 91, 92].

3.7.5 The double stranded RNA binding domain

In contrast to other RNA-binding domains that almost exclusively bind ssRNA, the double-stranded RNA-binding domain (dsRBD) has been shown to bind specifically dsRNA [93, 94]. The dsRBD domain is also known as DSRM (Double-Stranded RNA-binding Motif) and is comprised of 65-68 conserved amino acid residues [93]. dsRBD proteins are mainly involved in posttranscriptional gene regulation by converting adenine nucleobases in mRNAs to inosine nucleobases, and activating inhibitory proteins during protein translation [93, 95]. This domain is also found in RNA editing proteins. Interaction of the dsRBD with RNA is unlikely to involve the recognition of specific sequences, despite its specificity in deaminating adenosine nucleobases [94]. Nevertheless, multiple dsRBDs may be able to act in combination to recognize the secondary structure of specific RNAs in mRNA localization, such as the Staufen protein recognizing maternal mRNAs during egg development in *Drosophila melanogaster* [93]. [93].

NMR analysis of the dsRBD Staufen revelas an αβββαtopology (figure 6 E)) [93].

3.7.6 The Zinc finger domain

Another well-characterized nucleic acid-binding domain is the zinc finger domain (Znf domain). As its name suggests, some of the these domains bind zinc ions, but iron ions or no ion bound have also been reported, making the small protein motif quite versatile [96]. Also in allusion to their name, the small domains contact the target nucleic acids via finger-like protrusions, stabilized by salt bridges, on multiple occasions [96]. The zinc finger domain has been first described in the transcription factor TFIIIA from *Xenopus laevis*, where it binds DNA [96, 97]. Since then, its bound molecules have expanded to RNA, dsRNA:DNA hybrids, proteins, and lipids, where binding largely depends on the primary protein structure of the ZnF domains, the linker regions between them, as well as the secondary and tertiary structure of Znf proteins sequence [77, 96, 97]. Znf domain are found in in many different proteins across diverse functional and structural roles. They are involved in gene transcription, mRNA-transport, mRNA translation, cytoskeletal organisation, epithelial development, cell adhesion, protein folding, chromatin remodelling and zinc sensing [96]. Fortunately, Znf domains assume stable structures of two major types, rarely undergoing conformational changes upon binding, facilitating crystallization efforts [98]. The classical Znf type is of the C2H2-type, where two Cys and two His bind a single zinc ion tetravalently to hold it in place [96]. Structurally, this type forms a short β-hairpin and an α-helix illustrated in figure 6 F), and three subgroups of C2H2 Znf domains exist that were identified by others [77, 96, 99].

3.8 Identifying the RNA in RNPs (CLIP, PAR-CLIP, CRAC)

Detection of the RNA component in RNPs has been an established part of genomics for years. UV crosslinking and immunoprecipitation (CLIP) purifies short endogeneous RNA fragments that were UV-crosslinked to specific RBPs by seuqencing the RNA fragments. the CLIP technology is described in detail elsewhere and has not been used in this thesis, but it illustrates how the RNA portion of an RNP can also be studied in a high-throughput manner: In a standard CLIP, RBPs are UV-irradiated at 254 nm to covalently link protein with RNA. Next, the RNA is digested by RNAses and RBPs bound to RNA fragments are immunoprecipitated or affinity purified [100]. An adapter sequence is ligated to the RNA fragment, and purified crosslinked complexes are subjected to SDS-PAGE to further purify the complex from other contaminating proteins that were co-purified [101]. Next, proteinase K digestion cleaves the protein non-specifically until only the crosslinked amino acids remain bound to the RNA fragment. This small modification on the RNA fragment results in abbrogated cDNA synthesis using reverse transcriptase (RT) during cDNA libary generation via PCR and high-throughput RNA-sequencing (RNA-seq) [101]. Due to the infrequent read-through sequencing of the crosslinking site, the increase in truncated cDNA fragments can be detected and mapped onto the the corresponding RNA sequence that reveals the crosslinking site on the RNA [101].

There are many variations on the original CLIP method such as high-throughput sequencing of RNA isolated by CLIP (HITS-CLIP), individual-nucleotide resolution CLIP (iCLIP), infrared CLIP (irCLIP), enhanced CLIP (eCLIP), photoactivatable ribonucleoside-enhanced CLIP (PAR-CLIP) and Proximity-CLIP, excellently reviewed by Hafner *et al.*[46]. Both nucleobase analogs are more photoreactive than their natural counterparts and results in a T to C mutation and a G to A transition during PCR amplification, respectively [101]. crosslinking and analysis of cDNAs (CRAC) employes a different purification strategy compared to conventional CLIP by using affinity-based purification such as nickel affinity under denaturing conditions as an alternative to immunoprecipitation [101].

3.9 Identifying the proteins and peptides in RNPs using mass spectrometry

In addition to techniques aimed at identifying the RNA component of an RBP, there are different methods that detect the protein component of the RBP. Since mass spectrometry has been proven to be such a powerful technique in proteins identification in bottom-up approaches, peptide identification is performed by mass spectrometry. The RNA component is covalently crosslinked to the protein by UV-irradiation and also detected by mass spectrometry in various ways. Differences in those approaches arise from differences in isolation the RBPs from the proteomic background using affinity-based techniques, or immunoprecipiation. The following paragraphs describe each technique very briefly to demonstrate the current status on method development.

3.9.1 Identifying RNPs by UV-crosslinking coupled with mass spectrometry at amino acid resolution

UV-crosslinking at 254 nm coupled with mass spectrometry has been the "gold standard" in detecting peptides crosslinked to RNA nucleotides for years [58]. UV-crosslinking is thought to be a "zero-length" crosslinking event because a covalent bind between the interacting RNA nucleobase is directly formed with the amino acid side chain of a protein without mediating factor in between the two [46]. This limits crosslinkable nucleotides to just interacting nucleotides that make direct contact with the protein surface. Kramer *et al.* developed the original experimental workflow of UV-irradiation coupled with mass spectrometry to analyze UV-induced crosslinking sites at the amino acid residue level [46]. Key component to identifying crosslinked peptides and crosslinked RNA moieties is the RNP$_{xl}$ software that searches spectral data from any crosslinked protein-RNA complex or crosslinked proteome for peptide-RNA heteroconjugates [46]. This approach has been successfully demonstrated *in vitro* and *in vivo* on yeast protein-mRNA complexes [46].

Undoubtedly one of the strongest point for conventional UV-crosslinking coupled with mass spectrometry is the ability to identify the crosslinked amino acid based on spectral data. This may sound trivial, but an RNA attached to a peptide imposes multiple difficulties for search engines: i) The RNA is not a single PTM, but rather comprised of four nucleotides. ii) the length of the RNA may vary between mononucleotides or oligonucleotides being crosslinked to the peptide, increase combinational complexity in nucleobase combinations. iii) The RNA does not fragment the same way a PTM may be lost during HCD. Kramer *et al.* reported nucleotide specific mass adducts of uracil nucleobases that are observed during CID fragmentation in addition to the most prevalent neutral losses on the RNA moieties such as water losses, ammonia losses, and phosphate losses [46]. The authors noted that there was an uracil nucleotide bias that limited the detectable RNA nucleotides to U [46]. Since then, the approach has been extensively utilized to investigate many protein-RNA complexes, expanded to include the more reactive uracil analog 4-SU, and extended to DNA crosslinking identifying nucleotides other than uracil [3, 5, 102–105].

3.9.2 Other methods using UV-crosslinking coupled with mass spectrometry to identify RBPs

In classical RNA interaction capture (RIC), RBPs are crosslinked *in vivo* by UV irradiation [106–108]. RBPs binding to polyadenylated RNAs are consequently covalently bound to the proteins and can be "captured" with oligo(dT) magnetic beads [106]. Following stringent washes to remove artificially bound proteins, the mRNA interactome is determined by quantitative mass spectrometry [106, 107]. In theory, only RBPs bound to RNA in a physiological environment should be identified as crosslinking takes place *in vivo* [106]. Similarly, enhanced RNA interactome capture (eRIC) refines the original RIC approach by using a LNA-modified (locked nucleic acid) capture probe that constitutes a class of bicyclic RNA analogs, having an exceptionally high affinity and specificity toward their complementary DNA or RNA target molecules [109]. This method based on the use of an LNA-modified capture probe is claimed to include greater specificity and increased signal-to-noise ratios compared to existing methods such as RIC [107].

In RBDmap, cells are again crosslinked using UV-irradiation, and RBPs are captured using oligo(dT) capture, partially digested from the RNAs using LysC, and the partial digested subjected to another round of oligo(dT) capture [110]. Using MS-based analyses, the first oligo(dT) release should represent all RBPs, whereas identifications from the second oligo(dT) capture are thought to identify the RNA-binding domains of RBPs [110]. In peptide crosslinking and affinity purification (pCLAP), it is not entire RBPs that are oligo(dT) captured, but rather partially digested RBP peptides following UV-crosslinking that are subjected to affinity capture [110]. Consequently, pCLAP RIC, eRIC, and RBDmap all rely on oligo(dT) caputure of poly-adenylated mRNAs, failing to identify RNAs that do not carry a poly-A cap, limiting identifications to the mRNA-interactome. A possible way to circumvent this mRNA bias is the use of 4-SU UV-crosslinking at 350 nm. In a variant approach to RNA-associated protein purification (TRAPP) using 4-SU analogously to PAR-CLIP, cells are grown on 4-SU, incorporating the photo-reactive uracil-analog into all RNAs that can be subsequently UV-crosslinked in this PAR-TRAP approach [111]. Purification of RBPs is not performed by oligo(dT) capture, but silica beads, avoiding the poly-A-containing mRNA bias [112].

Another approach features the use of segmentally isotope-labeled RNA that is crosslinked by UV-irradiation and analyzed in tandem mass spectrometry [113]. This CLIR-MS/MS approach utilizes stable isotopes incorporated into the RNAs in specific regions that can be detected by a shift in expected delta masses of +1 Da, allowing ot localize the crosslinking position on the RNA [113]. Localizing the position on the peptide uses conventional mass shifts identified from UV-crosslinking and this approach was tested successfully on polypyrimidine tract binding protein 1 [114]. Overall, though, crosslinking approaches are severly limited to UV-irradiation, activating primarily uracil nucleobases to crosslinking to amino acid residues of RBPs, or to the photo-activation of its analog (4-SU), reacting with the same nucleotide bias towards amino acid residues [46, 112].

3.10 Chemical protein-protein crosslinking mass spectrometry

Chemical crosslinking mass spectrometry (XL-MS) provides a powerful tool to answer many questions in structural biology as crosslinking can be performed under native conditions, capturing the structural confirmations present. In combination with X-ray crystallography and cryo-EM, XL-MS can add highly complementary data. XL-MS has been traditionally defined for protein-protein crosslinking, investigating how different protein subunits interact with each other or how larger protein complexes are formed and change confirmation by providing information on the interfaces between the proteins. To achieve this, protein-protein crosslinking relies on small chemicals comprised of a spacer arm connected to two function groups that mediated the crosslinking events on each end [12, 115, 116]. The spacer length is variable, and so are the functional groups on either side of the crosslinker. Most commonly used, functional groups contain N-hydroxysuccinimide esters that specifically target primary amino groups on either end. Primary amino groups are found at the N-termini of the protein, as well as lysine side chains [12, 116]. The crosslinker 3-sulfo-N-hydroxysuccinimide ester (BS3) for example crosslinks lysine residues on either end, while adding a spacer reagion of 11.4 Å [12, 116]. Other chemical groups are reactive toward thiol groups found in Cys side chains or more reactive towards carboxyl groups found in Asp and Glu side chains [116]. As such protein-protein crosslinking has been aiding in structural anayses of various complexes, including ribosomes, mitochondria, the negative elongation complex, and many more, demonstrating its core position as a technique in structural biology [60, 117, 118]. However, XL-MS can also add valuable information on nucleic acids bound to a RNP complex, elucidated by UV-crosslinking for example, demonstrating the versatility of XL-MS in structural biology.

3.11 Chemical protein-nucleic acid crosslinking

Chemical crosslinking mass spectrometry cXL-MS for protein-RNA complexes utilized small chemicals similar to the N-hydroxysuccinimide esters used in protein-protein XL-MS that covalently link both RNA nucleotides with the interacting amino acids of the protein. In contrast to canonical UV-irradiation, chemical crosslinking is not of a "zero-length" nature, as the crosslinking reagent is comprised of atoms, occupying space between the two crosslinked moeties. As such, the spacer length of crosslinking reagents is specific to the chemical used, analogous to the crosslinkers used in protein-protein crosslinking. This spacer length, however, may be advantageous because it would allow to identify proximal amino acis that are not in direct contact with the RNA, but still close enough to define the protein-RNA interface. Since RBPs are usually quite flexible in their overall conformations to accomodate the highly dynamic nature of RNA superstructure, chemical crosslinking of RBPs may actually help in defining the protein-RNA interface by covering a greater surface available for crosslinking. The chemical structure of the three crosslinking reagents 1,2,3,4-Diepoxybutane (DEB), nitrogen mustard (NM), and 2-iminothiolane (2IT) used in this this are depicted in figure 7 and are briefly introduced in the following sections.

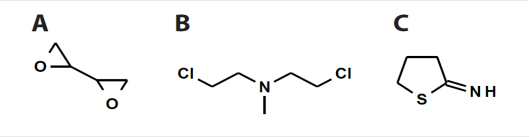


Figure 7: Chemical structure of crosslinkers used. A) The primary chemical crosslinking reagent used in this thesis is the epoxide 1,2,3,4-diepoxybutane that is known to react with both RNA and DNA, creating interstrand crosslinking events and nucleotide adducts that are cancerogenic to live cells. **B)** the nitrogen mustard mechlorethamine is an alkylating reagent, reacting with both DNA and RNA by forming interstrand crosslinks that lead to irreparable DNA damage. Its anti-neoplastic effects are used in chemotherapeutic treatments of Hodgkin's lymphoma. **C)** 2IT or Traut's reagent is a widely used chemical reagent that converts primary amines such as lysine side chains into UV-reactive thiols (sulfhydryl groups) and has been successfully to thiolate various proteins in addition to crosslinking RNA.

3.11.1 1,2,3,4-Diepoxybutane

1,2,3,4-Diepoxybutane (DEB) is a carcinogenic crosslinking reagent with two epoxide functional groups that can react as electrophiles and form covalent bonds with proteins, DNA, and RNA upon ring-opening (figure 7 A)) [119–121]. Early studies of DEB-mediated crosslinking of ribosomal proteins bound to the 16S and 23S RNA in *E. coli* by 2-dimensional gel system analysis identified multiple 30S protein crosslinked to the 23S rRNA: S3, S4, S5, S7, S8, S9, S12, S13, S14 and S18 [120]. Moreover, 50S ribosomal proteins L1, L2, L4, L13, L14-L21, L15, L16, L17, L18-L23, L19-22-24, L27 and L28 were identified to be crosslinked to the 23S RNA. DEB-mediated crosslinking revealed additional crosslinking events between 16S RNA and S1, S2 as well as S6 from the 30S subunit; and between 23S RNA and L10, L11, L7/12 from the 50S subunit [122]. Others have found elongation factor G to be crosslinked to the 23S rRNA upon DEB treatment, that seems to bind to a sequence of 27 NTs between position 1055 and 1081 of the 23S rRNA, forming a RNP [123].

Recent examples of DEB crosslinking include DNA-crosslinking of human fibrosarcoma cells, in which over 150 different proteins were chemically crosslinked to DNA [124]. Crosslinked proteins included transcription factors, tubulins, histones, and splicing factors. DEB-mediated crosslinking identified proteins involved in DNA-binding, transcriptional regulation, cell signaling, DNA repair, and DNA damage response [124]. Mass spectrometric analyses of crosslinked proteins have indicated that 1-(S-cysteinyl)-4-(guan-7-yl)-2,3-butanediol conjugates were formed, confirming that DEB creates DNA-lesions between cysteine side chains and the N-7 guanine positions found in DNA [124]. DEB-mediated crosslinking of DNA-strands of the 5S rRNA gene showed monoalkylation at all deoxyguanosines in addition to interstrand alkylation between strands in both free and nucleosome-bound states of the DNA, illustrating that DEB can seamlessly crosslink nucleosome condensed DNA [121, 125]. Overall, these studies suggest that DEB is an excellent crosslinking reagent for protein-nucleic acid crosslinking, potentially preferring G nucleotides that usually evade capture by UV-crosslinking methods in RNA [125].

3.11.2 The nitrogen mustard mechlorethamine

The nitrogen mustard bis(2-chloroethyl)methylamine or mechlorethamine (NM) is a highly carcinogenic chemical that has been described to induce alkylation reactions of DNA in 1942 and crosslinks in RNAs to ribosomal proteins in 1978[126, 127]. NM contains two chloroethyl and a methyl group attached to a tetrahedral nitrogen [128] (figure 7 B)). As a crosslinker, the two chloroethyl groups act as functional groups, undergoing a one-step SN2 intramolecular cyclization reaction, forming an aziridinium ion and chloride ions as a leaving group [128]. The aziridinium ion can then be attacked by nucleophilic nucleobases present in nucleic acids or nucleophilic amino acid side chains. NM is supposed to add a 12 to 15 Å spacer between the heteroconjugates, but evidence on how that number was computed is missing [126]. NM has been studied extensively in DNA-protein crosslinking events, showing that cysteine thiols within proteins can be crosslinked to N-7 positions of guanine in DNA via NM quite effectively by mass spectrometry similar to DEB-mediated crosslinking [126, 129]. Studies using DNA-analogs comprising of A and G nucleotides have unambiguously confirmed the N-7 position of purine nucleotides to be highly reactive towards NM, leading to G-A crosslinks, as well as A-A crosslinks in DNA [126, 128, 130].

Clinically, nitrogen mustards are used as chemotherapeutic agents as NM leads to formation of both intra-strand and inter-strand crosslinks in the DNA-double helix that cannot be repaired by the cellular machinery, leading to cell cycle arrest and apopotosis of tumorigenic cells [129]. More specifically, it has been found that DNA lesions of the inter-strand type (ICLs) are most cytotoxic by creating blocks between two DNA-strands that halt replication and transcription completely [129]. Intra-strand lesions are also cytotoxic, but to a lesser degree, possibly by the cellular repair machinery being able to repair bulky lesions in one strand by mismatch repair, nucleotide excision repair, or base excision repair[130, 131]. Due to the high level of cytotoxicity to any live cells in patients, nitrogen mustards have only been used sparsely clinically, focusing only on late-stage and drug resistant tumors such as Hodgekin's lymphoma, leukemia, multiple myeloma, and ovarian carcinoma [129]. Interventions with NM chemotherapeutics are coupled with localized treatments in treating ovarian carcinoma to minimize exposure to the patient's healthy cells in surrounding tissues [129].

While RNA crosslinking events inhibiting the transcriptionaly machinery have been noted to in many studies, direct RNA-centric mass spectrometric evidence is scanty. Even so, direct inhibition of transcription by NM could be shown *in vitro*, as well as interactions with polyribonucleotides, ribosomes, and enzymes involved in translation [126, 132, 133]. Overall, though, studied effects *in vivo* have largely been limited to detrimental outcomes due to DNA-damage with RNA-interactions "occurring on the side". Thus, NM has assumed the functional role of an alkylating agent of DNA, RNA, and proteins, making it an excellent candidate for chemical crosslinking of protein-nucleic acids coupled with mass spectrometry [128].

3.11.3 Traut's reagent 2-iminothiolane

2-iminothiolane (2IT), also known as Traut's reagent, was first used as a thiolating reagent in 1973 [134]. Traut *et al.* thioloated the 30S ribosome from *E. coli*, where the imido ester group of 2-IT reacts with the N-terminal amino group of lysine amino acid residues, converting the primary amine into a thiol that can be activated by UV-irradiation at 265 nm [134, 135]. In protein-RNA crosslinking using 2-IT, an activated thiol can react with uridine nucleotides for example, leading to a chemical crosslinking reaction with a spacer length of 8 Å between the two crosslinked moieties [135]. 2-IT has been extensively studied in need of converting amino groups into thiols, but also in the context of protein-RNA crosslinking. Urlaub *et al.* determined that UV-irradiation and 2-IT chemical crosslinking induced crosslinking events in ribosomal proteins from *E. coli* in 1995 [7]. Crosslinks within components of the small ribosomal subunit were identified by 30S ribosomal proteins S3, S4, S7, S14, and S17, whereas the large subunit was identified through crosslinks in 50S ribosomal proteins L2, L4, L6, L14, L27, L28, L29 [7]. Additionally, it has been shown that ribosomal protein S3 contains a KH-domain, being involved rRNA-binding [7]. Also, a crosslink within the 50S ribosomal protein L36 can be found within the zinc finger-like motif that bind rRNA as well [7]. More than 20 years have passed since the initial experiments have been carried out and drastic advances in the technologies used in mass spectrometry warrant re-visiting 2-IT crosslinking coupled with OT mass spectrometers.

3.12 Enrichment strategies for protein-nucleic acid heteroconjugates

Both UV-crosslinking and chemical crosslinking are grossly inefficient in their crosslinking yields, and are usually masked by highly abundant non-crosslinked proteins in highly complex samples during LC-MS[136]. Even though peptide mixtures are chromatographically separated by C18-RP chromatography, highly abundant proteins yield a large amount of peptides that elute at the same time, resulting in high MS1 signals that always outcompete low abundant peaks that are consequently never triggered for MS2 fragmentation. In low-complexity samples, fewer peptide elute at potentially the same time, allowing for potentially RNA-crosslinked peptides to generate a sufficient MS1 signal that is triggered for MS2 fragmentation. Enrichment strategies may be applied at the protein level, or at the peptide level post digestion, and aim at reducing non-specific, non-crosslinked peptide background noise. A reduction in background noise from linear peptides results in an increase of signal intensities for low-abundant RNA-crosslinked peptides that are now consequently selected, fragmented, and sequenced in MS2. A very simple, yet powerful approach of enriching peptide-RNA heteroconjugates is the use of TiO_2, originally designed to enrich for phosphopeptides [10, 46]. Similar to phospho-enrichment, the negatively charge phosphate in nucleic acid backbones binds to either TiO_2 or IMAC based resin well, allowing for specific enrichment [137, 138]. This method has been widely used in UV-based crosslinking in various samples. Other methods include biotinylated oligo(dT) probes, silica-based enrichment of nucleic acids, simple size-exclusion chromatography (SEC), basic reversed-phase chromatography, cleavable IMAC, and cleavable biotin-tagging to name only a few [3, 138–143].

3.13 The yeast spliceosomal Hsh49:Cus1 complex

The yeast spliceosomal protein Hsh49 constitutes a good example of an RBP because it is inherently small, only contains three proteins domains -two of which are structurally identical RRM domains-, and only contains a small portion of its sequence that cannot be crystallized because of the flexibility in the loop region (figure 8) [144, 145]. The human orthologue of Hsh49 is SF3b49 which is part of a heptameric protein complex (splicing factor 3b (SF3b)). Hsh49-binding protein Cus1 is also part of SF3b, and so are SF3a, the Sm proteins, Msl1p/Lea1p, and U2 snRNA. Together, these proteins form the U2 snRNP, which plays an important role in pre-mRNA splicing.

The sequence of Hsh49 and its RRMA domain architecture is highly conserved among different eukaryotic species [145]. Structurally, Hsh49 contains two RRMs, displaying the canonical RRM fold that binds to the 5' stem-loop I of U2 RNAin addition to the branchpoint recognition site [145]. Recent studies conducted in yeast, show that the N-terminal RRM1 binds to RNA, whereas the C-terminal RRM2 has not been shown to exhibit RNA-binding abilities [145]. A structural model based on crystallographic data has recently been published by Roon *et al.* (figure 8 [144]. In tis model, Hsh49 is bound to Cus1, forming a heterodimer [146]. Interestingly, the model shows that Cus1 binds to the α-helical structure of Hsh49 RRM1, opposite the four-stranded β-sheet, leaving the canonical RNA-binding surface of RRM1 available to bind mRNA [144, 146]. Binding affinity of RRM1 is increased when complexed with Cus1, probably due to the formation of an extended RNA-binding surface, spanning the two proteins [144].

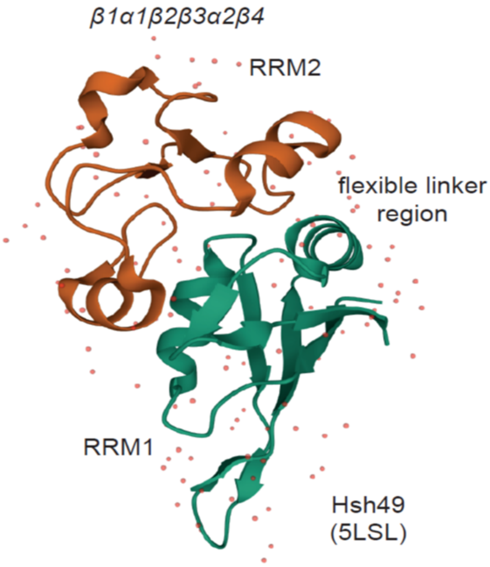


Figure 8: Crystal structure of Hsh49. Hsh49 is a yeast RNA-binding protein comprised of two RRM domains and a flexible linker region that past an α-helix that could not be crystalized by Van Roon *et al.* [146]. The minimal protein structured is naturally reduced to its simplest functional domains of which RRM1 is described to bind RNA in yeast with no detectable RNA-binding in RRM2, but orthologues of Hsh49 in *Caenorhabditis elegans* for example reverse the RNA-binding preferences of the two domains. Overall, its simplistic structural components and well-described RRMs served as the model protein upon which chemical crosslinking coupled with mass spectrometry was based upon in this thesis.

3.14 Methyl transferase Dnmt2 from *Saccharomyces pombe*

There are various enzymes used in tRNA-charging and modifying tRNAs. One class of tRNA-modifying enzymes are methyltransferases (MTs) that catalyze the transfer of a methyl group from the cofactor S-adenosyl-methionine (SAM) to the carbon-5 of an acceptor nucleotide [147]. Those MTs have originally been considered to be DNA MTases, but studies have shown that tRNAs are also recognized as substrates of this enzyme, now also playing a role in tRNA modifications [147]. In *S. pombe*, the methyltransferase Dnmt2 modifies the tRNA^{Asp} by methylating a codon-recognizing nucleotide (cytosine-38) at the 5'-carbon (figure 9) [5]. Functionally, methylation of tRNAs by Dnmt2 has been shown to protect tRNAs from cleavage, promotes the aminoacylation of tRNA^{Asp} *in vitro*, and contributes to translational accuracy [5]. Dnmt2 is structurally composed of two distinct domains: The SAH cofactor-binding domain, as well as the nucleotide-binding domain used in binding cytosine-38 [5]. Unfortunately, the tRNA^{Asp} could not be resolved by crystallization due to the flexibility of the RNA, but crosslinking sites derived from UV-XL-MS of the complex has provided valuable spatial restrictions for complex modeling *in silico* [5].

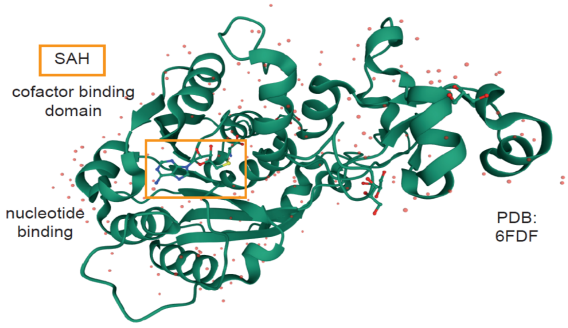


Figure 9: Crystal structure of Dnmt2 from *Saccharomyces pombe*. Methyl transferase Dnmt2 serves as an important tRNA-modifying enzyme in *S. pombe* by methylating cytosine-38 of various tRNAs such as tRNA^{Asp}, tRNA^{Glu}, and tRNA^{Val}. Here, Dnmt2 was crystalized by Johansson *et al.* in the presence of tRNA^{Asp} that was omited to better show the structural subunits of the protein [5]. The methyl group is donated by the cofactor S-adenosyl-L-homocysteine (SAH) and catalysis occurs at a flexible "active site loop", containing cysteine-80 that is crucial to catalysis. The SAH molecule is boxed in yellow and the flexible active site loop is depicted as the lateral unstructured region at the lower bottom.

3.15 The negative elongation factor NELF

There are many different ways to regulate gene expression. Most commonly, transcription factors activate or repress certain genes by opening or blocking binding sites of the DNA for DNA-dependent RNA-plolymerase (RNAP) to initiate transcription of the target gene. Gene regulation can also occurr by a mechanism known as promoter-proximal pausing of RNA polymerase II (Pol II) [148]. The protein complexes DRB sensitivity-inducing factor (DSIF) and negative elongation factor (NELF) both act to stabilize a paused Pol II, where a tilted DNA-RNA hybrid structure is formed that blocks binding of incoming nucleotides triphosphates (NTPs) to be incorporated into the growing pre-mRNA transcript [148].

Structurally, the NELF complex is comprised of the four NELF subunits NELF-A, NELF-B, NELF-C/D, and NELF-E (figure 10 that restrain Pol II mobility and prevents anti-pausing factor TFIIS from binding the supercomplex [148]. Restraint is achieved by binding to the polymerase funnel, two mobile polymerase modules, and the trigger loop as previously shown [148]. Additionally, NELF possesses two flexible tentacles regions within NELF-A and NELF-E that can contact DSIF and exiting RNA and their involvement in RNA-binding has been alluded to previously [148]. The crystal structure presented in figure 10 lacks the two tentacle regions from both NELF-A and NELF-E because they are comprised of IDRs and evade crystallization processes. Interestingly, Rawat *et al.* reported that both tentical regions within NELF-A and NELF-E are cooperatively involved in liquid-liquid phase-separation *in vitro*, positing an explanation how NELF condensates within the nucleus in stressed human cells, leading to transcritpional downregulation that supports cellular survival [149].

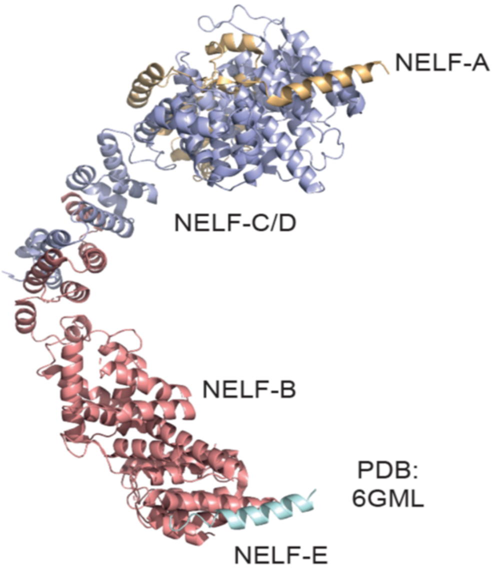


Figure 10: Crystal structure of the NELF complex. The negative elongation factor (NELF) consists of four subunits, NELF-A, NELF-B, NELF-C/D, and NELF-E. It is important in promoter proximal pausing of DNA-dependent RNA-polymerase II (RNAPII) and has been well characterized by Vos *et al.* [150, 151]. The NELF complex is known to be RNA-binding via the NELF-AC highly conserved core, as well as the NELF-E RRM (not shown here) and NELF-B in association with NELF-AC. Additionally, NELF-A and NELF-E contain flexible tentacle regions that are so disordered, attempts at crystallizing them failed so far. Those regions are thought to be involved in RNA-binding, yet mass spectrometric evidence at amino acid resolution has been lacking so far.

3.16 *Escherichia coli* as a model organism

The Gram negative bacterium *Escherichia coli* has been studied as model organism and used as a protein-producing factory for years [152]. The species *E. coli* contains a diverse family of pathogenic and apathogenic members that may cause no disease at all, serving as a commensal organism, mild, moderate or severe diseases such as the pathologies associated with enterohaemorrhagic *E. coli* (EHEC), strain O157:H7 that can cause bloody diarrhea, hemolytic uremic syndrome, kidney failure, and even death [153]. By studying apathogenic types of *E. coli*, reasearchers hope to gain insights into bacterial processes that also occur in pathogenic types and may be targeted pharmacologically. Different strains have been isolated and artificially created to accommodate laboratory work safely by removing most virulence factors such as pathogenic components of the outer membrane and toxins [154]. Additionally, apathogenic *E. coli* are routinely used in the laboratory for the production of proteins because it is easy and cost-effective to grow *E. coli* in standard media.

E. coli BL21 DE3 is routinely used for protein expression from plasmids introduced into competent bacterial cells [152]. This strain contains the lambda DE3 pro-phage, carrying the T7-plolymerase that is under control of the *lac* promoter [152, 154]. Protein expression can thus be easily induced by IPTG, activating the *lac* promoter [152, 154]. Additionally, *E. coli* BL21 is lacking the lon-protease that could degrade produced proteins upon IPTG-induction [152]. Moreover, this strain also lacks the OmpT protein, another protease that might reduce the quality and quantity of expressed proteins if present [152]. In some cases, however, *E. coli* may not be a good system for some selected protein expressions, as prokaryotes cannot glycosylate eukaryotic glykoproteins, resulting in artificial eukaryotic proteins at best [152]. These reasons allow for easy production of various proteins in high numbers and of high quality, making this strain ideal for protein expression.

While *E. coli* BL21 DE3 is nowadays almost exclusively used for protein expression, it can still be used as a model organism for prokaryotic cells without being exposed to an actual pathogen. *E. coli* BL21 DE3 has been genetically altered to only produce a truncated version of the core saccharide that is part of the outer cell wall in addition to a deletion of the region containing genes that encode functional flagella, rending it non-motil and not invasive [154]. *E. coli* BL21 DE3 is still a living bacterial cell, though, undergoing biochemical processes common to most prokaryotes that can be studied with novel techniques in proof of concepts studies such as this thesis.

3.17 *Bacillus subtilis* as a model organism

Bacillus subtilis is a Gram-positive, endospore forming rod shaped apathogenic bacterium that has been used extensively in the laboratory [155]. It does contain a thick cell wall, consisting of a thick layer of peptidoglycan that may impede uptake of certain compounds by the bacterium [156]. *B. subtilis* is easy to grow and secretes target proteins efficiently into the medium, making it the Gram-positive equivalent to *E. coli* when it comes to ease of protein production [155]. Moreover, it is used in the study of biofilms and quorum sensing, as well as the study of its antimicrobial peptides; the ribosomal peptides (RPs) (bacteriocins), the polyketides (PKs), the non-ribosomal peptides (NRPs) and the volatiles used in competition with neighboring bacteria for the same food sources [156–158]. Overall, this bacterium is an excellent candidate for studying systemic effects of a treatment or method on live cells across fundamentally different prokaryotic cells when coupled with apathogenic *E. coli* cells due to their robustness, low maintenance-cost, and safety in handling.

3.18 Human HeLa cells as a model organism

HeLa cells stem from a biopsy obtained from a woman, called Henrietta Lacks who died in 1951 in the wake of an aggressive adenocarcinoma of the cervix that was sent to the Laboratory of Dr. Geroge O. Gey at Johns Hopkins [159]. At Johns Hopkins it proliferated rapidly and researchers became interested in this human cell line, now called HeLa and has been an important cell line in human molecular biology [159]. While it is generally accepted that HeLa cells are cancerous, it is still widely believed that these cells somewhat reflect human physiological conditions. Analyses of the genomic material of different HeLa cells revealed extensive genomic rearrangements [160, 161]. First, Henrietta Lack's genome consisted of 46 normal chromosomes, but HeLa cells contain 70 - 90 chromosomes [161]. Second, these rearrangements are alluding to catastrophic chromosome shattering that does not reflect normal conditions [161]. Moreover, gene expression profiling drew attention to several pathways, including cell cycle and DNA repair, that exhibited extremely different expression patterns from those found in normal human tissues [159, 161]. Thus, it is important to account for the strikingly aberrant characteristics of HeLa cells and not to assume physiological homeostasis when designing and interpreting experiments with HeLa cells.

4 Aims of this thesis

Based on previous knowledge regarding chemical crosslinking using DEB, NM, and 2-IT, chemical protein-RNA crosslinking is known to occur, a reliable workflow for the LC-MS-based identification of chemical protein-RNA crosslinking sites on amino acid resolution, is missing. Thus, seven aims were derived to develop such cXL-MS method for protein-RNA crosslinking: i) Validating that the chemical crosslinkers DEB, NM, and 2-IT can crosslink protein to RNA. ii) Identification of nucleotide specific adduct masses that unambiguously correspond to the chemically crosslinked nucleotide to the model protein Hsh49. This enables crosslinking site determination at amino acid resolution for all three crosslinkers. iii) Applying a reliable and robust chemical workflow for crosslinking coupled with mass spectrometry to different protein-RNA complexes in a proof of concept study *in vitro*. Hsh49:Cus1, Dnmt2, and the NELF complex serve as excellent model proteins/complexes because they can be analyzed by cXL-MS with their native RNAs. iv) Comparing cXL-MS to canonical UV-crosslinking to identify commonalities and differences on both amino acid residue level, as well as detectable nucleotides. v) Expanding the method of chemical crosslinking coupled with mass spectrometry to entire live cells *in vivo*. *E. coli*, *B. subtilis*, and HeLa cells are easy to grow, handle, and readily accessible to test the cXL-approach in both prokaryotic and eukaryotic cells. vi) Testing different FAIMS settings to identify protein-RNA crosslinking sites after FAIMS-separation. And lastly, vii) Identifying major RBPs from chemical crosslinking of *E. coli* cells from various approaches.

5 Materials and Methods

5.1 Contributions of other researchers to this work

1. Luisa M. Welp, former Master's student, now colleague in the Bioanalytical mass spectrometry group, Urlaub Lab (MPIbpc) (Hsh49 protein purification under supervision and as instructed by established protocol, Dnmt2 crosslinking and analysis of spectra from 2IT-mediated and NM-mediated crosslinking as part of her supervised Master's Thesis.)

2. Monika Raabe, technician in the bioanalytical research group, Urlaub lab (MPIbpc) (extensive Sample preparation of *E. coli* cell lysates for FAIMS analyses as instructed, following established protocol)

3. Sven Johansson, former PhD student from the Ficner lab (University of Göttingen)(Expression and purification of the Dnmt2 tRNA$^{\mathrm{Asp}}$ complex)

4. Dr. Piotr Neumann, postdoctoral researcher at the Ficner lab (University of Göttingen) (modeling of Dnmt2 bound to tRNA$^{\mathrm{Asp}}$)

5. Dr. Jana Schmitzova, former postdoctoral researcher in the Lührman lab (MPIbpc) (purification of the Hsh49:Cus1 complex as instructed by established protocol)

6. Dr. Seychelle Vos, former postdoctoral researcher from the Cramer lab (MPIbpc) (Provided the purified NELF:TAR complex)

5.2 Materials

Table 2: List of Chemicals used. Some chemical names have been abbreviated.

Product	Company	City/Country
2IT (2-iminothiolane)	Thermo Fisher	Waltham, MA, USA
Acetic acid (AcOH)	Merck KGaA	Darmstadt, GER
Acetone ($(CH_3)_2CO$)	Merck KGaA	Darmstadt, GER
Acetonitrile (ACN), LiChrosolv	Merck KGaA	Darmstadt, GER
Ampicillin (Amp)	Sigma Aldrich	St. Louis, MO, USA
Beta-mercapto Ethanol (β-ME)	Carl Roth GmbH	Karlsruhe, GER
BS3, no-Weigh ™	Thermo Fisher	Waltham, MA, USA
Calcium chloride ($CaCl_2$)	Merck KGaA	Darmstadt, GER
cOmplete ™ Protease Inhibitor	Merck KGaA	Darmstadt, GER
Coomassie Brilliant Blue G 250™	SERVA	Heidelberg, GER
D-(+)-glucose	Merck KGaA	Darmstadt, GER
DEB (1,2,3,4-diepoxybutane)	Sigma Aldrich	St. Louis, MO, USA
Dithiothreitol (DTT)	Alexis Biochemicals	San Diego, CA, USA
Ethanol (EtOH)	Merck KGaA	Darmstadt, GER
Formaldehyde (CH_2O)	Sigma Aldrich	St. Louis, MO, USA
Formic acid (FA)	Sigma Aldrich	St. Louis, MO, USA
Glycerol	Merck KGaA	Darmstadt, GER
Glycogen (20 mg/mL)	Thermo Fisher	Waltham, MA, USA
HEPES	Merck KGaA	Darmstadt, GER
Hydrochloric acid (HCl)	Merck KGaA	Darmstadt, GER
Iodacetamide (IAA)	Sigma Aldrich	St. Louis, MO, USA
Imidazole	SERVA	Heidelberg, GER
IPTG	Sigma Aldrich	St. Louis, MO, USA
Kanamycin (Kan)	Carl Roth GmbH	Karlsruhe, GER
Magnesium acetate ($MgAc_2$)	Merck KGaA	Darmstadt, GER
Magnesium chloride ($MgCl_2$)	Merck KGaA	Darmstadt, GER
Mechlorethamine (NM)	ApexBio Technology	Taiwan, TW
Methanol (MeOH), LiChrosolv	Merck KGaA	Darmstadt, GER
Potassium chloride (KCl)	Merck KGaA	Darmstadt, GER
ReproSil-Pur 100 C18-AQ 5 µm	Dr. Maisch GmbH	Ammerbuch, GER
ReproSil-Pur 120 C18-AQ 1.9 µm	Dr. Maisch GmbH	Ammerbuch, GER
ReproSil-Pur 120 C18-AQ 5 µm	Dr. Maisch GmbH	Ammerbuch, GER
Silver nitrate ($AgNO_3$)	Sigma Aldrich	St. Louis, MO, USA
Sodium acetate (NaAc)	Merck KGaA	Darmstadt, GER

Sodium carbonate (Na_2CO_3)	Merck KGaA	Darmstadt, GER
Sodium chloride (NaCl)	Merck KGaA	Darmstadt, GER
Sodium thiosulfate) ($Na_2S_2O_3$)	Merck KGaA	Darmstadt, GER
Titansphere^{TM} 5 µm (TiO_2)	GL Sciences	Tokyo, JP
Trisaminomethane (TRIS)	VWR chemicals	Darmstadt, GER
Urea ($CO(NH_2)_2$)	Merck KGaA	Darmstadt, GER
Water (H_2O), LiChrosolv	Merck KGaA	Darmstadt, GER
X-Gal Solution^{TM} (ready-to-use)	Thermo Fisher	Waltham, MA, USA

Table 3: List of pre-formulated gels, buffers and solutions.

Product	Company	City/Country
2xYT medium powder	Sigma Aldrich	St. Louis, MO, USA
InstantBlue^{TM} Protein Stain	Expedeon	Cambridge, UK
LB (lysogeny broth) medium powder	MP Biomedicals	Eschwege, GER
NuPAGE Antioxidant solution	Invitrogen	Karlsruhe, GER
NuPAGE LDS sample buffer (4x)	Invitrogen	Karlsruhe, GER
NuPAGE 20 x MOPS SDS running buffer	Invitrogen	Karlsruhe, GER
NuPAGE Novex 4-12% gradient Bis-Tris Mini Gels (1 mm)	Invitrogen	Karlsruhe, GER
NuPAGE Sample reducing agent (10x)	Invitrogen	Karlsruhe, GER
PCI solution (25:24:1)	Roth	Karlsruhe, GER
Protease Inhibitors (EDTA free)	Roche	Mannheim, GER
SOB medium (super optimal broth)	MP Biomedicals	Eschwege, GER
T4 polynucleotide kinase reaction buffer	New England Biolabs	Frankfurt, GER
Triethanolamine buffer solution pH 8.0	Sigma Aldrich	St. Louis, MO, USA
Trypsin resuspension buffer	Promega	Madison, WI, USA

Table 4: List of commercially available kits, markers, and standards.

Commercially available kits, markers, and standards		
Pierce BCA Protein Assay	Thermo Fisher	Waltham, USA
Protein marker Precision Plus, All Blue	Bio-Rad	Munich, GER
Protein marker Precision Plus, Xtra	Bio-Rad	Munich, GER
Protein marker Precision, Dual Color	Bio-Rad	Munich, GER

Table 5: List of enzymes.

Product	Company	City/Country
Benzonase	MilliporSigma	Burlington, USA
Chymotrypsin	Roche	Mannheim, GER
RNase A	Ambion	Darmstadt, GER
RNase T1	Ambion	Darmstadt, GER
T4 polynucleotide kinase	New England Biolabs	Frankfurt, GER
TEV protease	produced in-house	Göttingen, GER
Trypsin (modified, sequencing grade)	Promega	Madison, WI, USA

Table 6: List of buffers and solutions.

Buffers and solutions	
1 M DTT	154.25 mg/mL DTT in water (stored at -20 °C)
1 M glucose solution	180.2 mg/mL in water, sterile-filtered
1 M Tris/HCl buffer, pH 7.4	121.14 mg/mL Tris base in water, pH adjusted with 37 % (v/v) HCl
8 M urea	0.48 g/mL urea in water
Collodial coomassie staining solution	0.08 % (w/v) Coomassie Brilliant Blue G-250 solution
Coomassie Brilliant Blue staining solution	20 % (v/v) methanol, 1.6 % (v/v) orthophosphoric acid, 8 % (w/v) ammonium sulfate
HEPES crosslinking buffer	25 mM HEPES/HCl, 100 mM NaCl, pH 8.0

NiNTA buffer A	20 mM Tris/HCl, 500 mM urea, 500 mM NaCl, 25 mM imidazole, 5 mM β-ME, 5 % (v/v) glycerol, pH 7.4
NiNTA buffer B	20 mM Tris/HCl, 500 mM urea, 500 mM NaCl, 500 mM imidazole, 5 mM β-ME, 5 % (v/v) glycerol, pH 7.4
HPLC buffer A	0.08 % FA (v/v)
HPLC buffer B	80 % ACN (v/v), 0.08 % FA (v/v)
Protein storage buffer	25 mM HEPES/NaOH, 200 mM NaCl, 10 % (v/v) glycerol, pH 8.0
LC-MS injection buffer	5 % (v/v) acetonitrile, 0.1 % (v/v) formic acid
TEA crosslinking buffer	25 mM TEA/HCl, 50 mM KCl, 1 mM magnesium acetate, 1 mM β-ME, pH 7.5

Table 7: List of miscellaneous consumables.

Consumable	Company	City/Country
Amicon^{TM} Ultra Centrifugal filter units Ultra-15 MWCO 10 kDa	Merck KGaA	Darmstadt, GER
C18 Column Material Reprosil-Pur basic C18-HD (120 Å, 5 µm)	Dr. Maisch GmbH	Ammerbuch-Entringen, GER
Eppendorf Safe-Lock Tubes (0.5 mL, 1.5 mL, 2 mL)	Eppendorf	Hamburg, GER
Greiner 96 well plates polypropylene	Greiner Bio-One	Frichenhausen, GER
HiTrap HP	GE Healthcare	Chalfont St. Giles, UK
HiTrapTM Heparin HP	GE Healthcare	Chalfont St. Giles, UK
Microplate 96K black	Greiner Bio-One	Frichenhausen, GER
MicroSpin Columns G-26	GE Healthcare	Chalfont St. Giles, UK
NuPAGE Novex 4-12% Bis-Tris gradient Mini Gels (1mm)	Invitrogen	Karlsruhe, GER
PIPETMAN DIAMOND^{TM} Tips	Gilson	Middleton, WI, USA
Slide-A-Lyzer^{TM} MINI Dialysis Device 3.5 K MWCO, 0.1 mL	Thermo Fisher	Waltham, MA, USA
Storage Phosphor Screen	GE Healthcare	Chalfont St. Giles, UK
Whatman 3mm CHR paper	GE Healthcare	Chalfont St. Giles, UK

Table 8: List of live cells.

Cell line	Company	City/Country
Bacillus subtilis	AG Fischle	MPI-bpc, GER
Escherichia coli BL21 DE3	Thermo Fisher	Waltham, MA, USA
Escherichia coli BL21 DE3 Rosetta^{TM}	Merck KGaA	Darmstadt, GER
HeLa cells	Bioreactor	MPI-bpc, GER

Table 9: List of cell culture materials.

Media/ solution	Composition
2xYT medium	31 g/L 2xYT powder in water, autoclaved
Inoue solution	10.9 g MnCl, 2.2 g $CaCl_2$, 18.7 g KCl, 20 mL PIPES solution (0.5 M) in 1 L water
LB medium	25 g/L LB powder or 4 capsules in water, autoclaved
SOB medium	31 g/L SOB powder in water, autoclaved
SOC medium	50 mL SOB medium, 1 mL 1 M glucose solution

Table 10: List of technical instruments.

Instrument	Company	City/Country
8W lamps 254 nm G8T5	Sankyo Denki	Hiratsuka, JP
ÄKTAxpress	GE Healthcare	Chalfont St. Giles, UK
Bandelin Sonorex RK 510 H Ultrasonic Unit	Reichmann	Hagen, GER
Bioruptor^{TM} Sonication System UCW-201TM	Diagenode	Seraing, BEL
BP211D Analytical Balance	Sartorius AG	Göttingen, GER
BP41000 Precision Balance	Sartorius AG	Göttingen, GER
CPA423S Lab Balance	Sartorius AG	Göttingen, GER
Clean Bench HeraSafe & Heraeus	Thermo Fisher	Waltham, MA, USA
EmulsiFlex-C5	AVESTIN	Ottawa, ON, Canada
Eppendorf Centrifuge 5415R	Eppendorf	Hamburg, GER

Eppendorf Concentrator 5301	Eppendorf	Hamburg, GER
Epson Perfection V700 Photo	Seiko Epson K.K.	Nagano, Jp
Gel dryer model 538	Bio-Rad	Munich, GER
Heraeus Fresco 17	Thermo Fisher	Waltham, MA, USA
Heraeus Biofuge Pico	Thermo Fisher	Waltham, MA, USA
Heraeus Megafuge 1.0 R	Thermo Fisher	Waltham, MA, USA
Heraeus Pico 17	Thermo Fisher	Waltham, MA, USA
Herasafe HSP12 Hood	Heraeus	Hanau, GER
KMO 2 basic IKAMAG$^{\text{TM}}$	IKA	Staufen, GER
Lambda Scan 200e	MWG-Biotech	Ebersberg, GER
Mini-6K Centrifuge	Hangzhou Allsheng	Zhejiang, CHN
Multitron Standard Incubator	INFORS HT	Bottmingen, CHE
NanoDrop$^{\text{TM}}$ ND-1000	Thermo Fisher	Waltham, MA, USA
NESLAB RTE-7 Circulation Bath	Thermo Fisher	Waltham, MA, USA
Phosphoimager Typhoon 8600	GE Healthcare	Chalfont St. Giles, UK
PIPETMAN Classic (P2, P10, P20, P100, P200, P1000)	Gilson	Limburg, GER
Scintillation Counter Tri-Carb 2100TR	Packard PerkinElmer	Waltham, MA, USA
SorvallTM LYNXTM Superspeed Centrifuge	Thermo Fisher	Waltham, MA, USA
SpeedVac Concentrator Savant SPD121P	Thermo Fisher	Waltham, MA, USA
SureLockTM Mini-Cell Electrophoresis system	Thermo Fisher	Karlsruhe, GER
Thermomixer Comfort	Eppendorf	Hamburg, GER
ThermoStat plus	Eppendorf	Hamburg, GER
UV-light crosslinking apparatus	built in-house	MPI-bpc, GER
Vortex Genie 2	Roth	Karlsruhe, GER

Table 11: List of LC-Mass spectrometers and FAIMS.

Instrument	Company	City/Country
FAIMS Pro^{TM}	Thermo Fisher	Waltham, MA, USA
Orbitrap Fusion^{TM} Lumos^{TM} Tribrid^{TM} Mass Spectrometer	Thermo Fisher	Waltham, MA, USA
Orbitrap Fusion^{TM} Tribrid^{TM} Mass Spectrometer	Thermo Fisher	Waltham, MA, USA
Orbitrap QExactive^{TM} HF-X Mass Spectrometer	Thermo Fisher	Waltham, MA, USA
Orbitrap Exploris^{TM} Mass Spectrometer	Thermo Fisher	Waltham, MA, USA
UltiMateTM 3000 RSLCnano System	Thermo Fisher	Waltham, MA, USA

Note that details about R and RStudio, including the different packages with their specific version numbers can be found in table 14

Table 12: List of software used.

Software	Company	City/Country
Adobe Illustrator CS6 v.14	Adobe Incorporated	San Jose, CA, USA
Adobe Illustrator CS6 v.16	Adobe Incorporated	San Jose, CA, USA
Freestyle 1.6	Thermo Fisher	Waltham, MA, USA
Microsoft Office 2010	Microsoft	Redmond, WA, USA
OpenMS 2.4.0	Kohlbacher et al. 2016 [46]	Tübingen, GER
Proteome Discoverer^{TM} Software 2.1	Thermo Fisher	Waltham, MA, USA
Proteome Discoverer^{TM} Software 2.4	Thermo Fisher	Waltham, MA, USA
PyMOL 2.1	Schrodinger Center	New York, USA
RNP$_{xl}$	Kohlbacher et al. 2016 [46]	Thübingen, GER
Nuxl	Kohlbacher et al. (in preparation)	Thübingen, GER
Xcalibur 3.0	Thermo Fisher	Waltham, MA, USA
Xcalibur 4.1	Thermo Fisher	Waltham, MA, USA

5.3 Methods

5.3.1 Hsh49 & Hsh49:Cus1 expression and purification

Both Hsh49 alone, as well as in combination with Cus1, were expressed in bacterial cells, similar to Van Roon *et al.* to yield adequate amounts of purified protein [144]. First, His-tagged Hsh49 protein alone was overexpressed in *Escherichia coli* BL21 DE3 cells growing in LB broth supplemented with Kanamycin. Protein production from the pET vector containing His$_6$-Hsh49 (pET:Hsh49) was induced with 0.5 mM IPTG at an OD$_{600}$ of 0.6 and grown for an additional 4 hours at 30 °C. To lyse the bacterial cells, lysis buffer (20 mM Tris/ HCl, pH 7.4, 500 mM urea, 500 mM NaCl, 25 mM imidazole, 5 mM β-ME, 5 % glycerol, and one tablet of cOmplet^{TM} Mini EDTA-free Protease Inhibitor Cocktail) was added and the lysate was thoroughly resuspended. Cells were lysed by pressurizing the suspension multiple times in a Canadian Press. To remove cellular debris, the lysate was centrifuged at 8000 rpm for 30 min.

The supernatant was subsequently loaded on two tandem HisTrap HP columns. Both columns were equilibrated with Ni-NTA-A buffer (20 mM HEPES/ KOH, pH 7.4, 500 mM urea, 500 mM NaCl, 25 mM imidazole, 5 mM β-ME, and 5 % glycerol). Unbound proteins and DNA/RNA remnants were washed away in a high salt wash using Ni-NTA-A with 1 M NaCl. A linear gradient ranging from 0-100% Ni-NTA-B (Ni-NTA-A with 500 mM imidazole) eluted the His-tagged Hsh49 in 1 mL elution fractions. Hsh49-containing fractions were identified by SDS-PAGE analysis and pooled. Pooled eluates were TEV-digested (dialysis against Ni-NTA-A with a Hsh49 to TEV ratio of 1:40) overnight at 4 °C. Removal of the His-tagged TEV enzyme from the pooled eluates was achieved by a second Ni-NTA column analogous to the first Ni-NTA column, leaving highly purified Hsh49 in the flow-through. Upon collecting, the flow-through was concentrated and buffer was exchanged into the appropriate storage buffer (20 mM HEPES/ KOH, pH 7.4, 200 mM NaCl, 5 mM β-ME, 10 % glycerol) using 10 kDa MWCO Amicon Ultra Centrifugal filters. Hsh49:Cus1 proteins were expressed and purified simultaneously from the same pETDuet vector akin to the purification strategy employed for Hsh49 alone. The identity of the target proteins was confirmed using SDS-PAGE and tandem mass spectrometric analyses.

5.3.2 Determination of Protein concentration

Out of the many ways to determine protein concentrations, the Bradford Protein Assay and the BCA-Assay (bicinchoninic acid) are used routinely, owing to their robustness and reproducibility [162, 163]. In principle, the manufacturers' manuals (BioRad, Thermo Fisher) were followed with slight variations. Both assays require a standard curve to be generated from which the unknown protein concentration was calculated.

The protein standard was comprised of 5-7 data points of known BSA protein concentrations ranging from 0 - 2 µg. A serial dilution of BSA stock solution, 0.2 mg ml^{-1}, was used to generate such BSA standard samples to which either Bradford or BSA reagent is added. Both unknown sample and known standards were measured in triplicates for robust and accurate protein concentrations.

5.3.3 5'-labeling of RNA oligonucleotides with γ-^{32}-ATP

Synthesized stretches of poly-RNA oligonucleotides were incubated with γ^{32}P-ATP and T4 PN kinase in T4 PNK buffer. The mixture was incubated for 1 hour at 37 °C and the labeled poly-nucleotides were isolated from the T4 enzyme using MicroSpin^{TM} G-25 columns in combination with standard ethanol-chloroform extraction. The labeled RNA was resuspeded in RNAse-free water. Radioactive signal intensities were quantified by liquid scintillation counting.

5.3.4 PCI extraction

Phenol-chloroform-isoamylalcohol (PCI) extraction allows for the separation of suspended proteinaceous components from nucleic acids. Both RNA and DNA can be separated from proteins using PCI extractions. First, 1 volume of PCI solution and 1 µl of 1 µl glycogen were added to the sample. Second, the mixture was incubated whilst vigorously shaking for 15 min. The sample was subsequently centrifuged for 5 min, 13000 rpm, at room temperature to produce a two-phased liquid-liquid mixture. Nucleic acids amassed in the upper aqueous phase, while proteins concentrated in the organic lower phase. The aqueous phase was transferred into a fresh Eppendorf tube to further purify to nucleic acids from the aqueous phase by adding one volume of chloroform. Analogously to the first incubation, the mixture was shaken and then centrifuged again. Purified nucleic acids were then extracted from the second aqueous phase by carefully aspirating, followed by ethanol precipitation.

5.3.5 Ethanol precipitation

Mixtures containing nucleic acids and residual proteins from PCI extractions were usually further purified in an ethanol precipitation step. Alternatively, cross-linked protein-nucleic acid complexes were usually stored as an ethanol precipitate at -20 °C. To precipitate either nucleic acids or cross-linked complexes, three volumes of ice-cold ethanol (-20 °C) and 1/10 volume of 3 M NaOAc, pH 5.3, were added to the mixture. A short centrifugation step at 8000 rpm for 1 min ensured all liquid was collected into one uniform body. The mixture was incubated at -20 °C for at least 2 h. Alternatively, the Eppendorf tube may be dipped into liquid nitrogen for 3 s, followed by placing the tube on ice until further use. In any case, precipitated samples should appear as white flakes in the Eppendorf tube, which were centrifuged for 30 min at 13000 rpm, 4 °C, to precipitate all protein components of the solution, and collect them in a pellet. Aspirating the supernatant with long thin tips left the pellet at the bottom of the tube. Washing the pellet with 2 volumes of 80% ice-cold (-20 °C) ethanol was followed by an additional analogous centrifugation step. Aspirating the supernatant with long thin tips left a purified pellet that was dried by vacuum centrifugation until the solvent was completely removed.

In the case of Bradford assays, the protein sample was first diluted with deionized water to a final volume of 800 µL, to which 200 µL of Bradford solution (BioRad) were added. The diluted mixture was incubated at room temperature in the dark for 10 min. Bradford assays require measuring the absorbance at 595 nm and unknown concentrations are calculated using the Lambart-Beer Law. BCA assays can be performed exactly to the manufacturer's protocol, if starting amounts are not limiting, or scaled down to a mini-BCA assay. For a mini-BCA assay, the BCA working reagent was prepared 50:1 and 100 µL of working reagent were added to 5 µL of either sample or BSA standard. The mixture was incubated at 37 °C and absorbance was measured at 562 nm. The mini-BCA assay only required 2 µL of incubated mixture to be measured photometrically.

5.3.6 SDS-PAGE using the NuPAGE system

Proteins can be separated by molecular weight using (native or denaturing) gel electrophoresis. The NuPAGE system (Thermo Fisher Scientific) constitutes an optimized inclusive system for SDS-PAGE. Following the manufacturer's protocol, 100 µL of protein samples constitute a reasonable starting volume, to which 10X NuPAGE sample reducing agent (500 mM DTT) and 4X NuPAGE sample buffer were added in corresponding portions. The mixtures were heated in a heating block for 10 min at 70 °C before being loaded onto a pre-cast 4-12% gradient Bis-Tris 1.0 mm gel. The gel was run in the appropriate chamber in a MOPS SDS running buffer, complemented with NuPage Antioxidant, for about 50 min at 200V.

5.3.7 SDS-PAGE with radioactively labeled RNA

Reconstituted protein-RNA complexes were cross-linked chemically as described above. Cross-linked samples were subsequently analysed by SDS-PAGE and autoradiography. First, samples were run on pre-cast Bis/Tris NuPAGE SDS gels following the manufacturer's protocol, stained with colloidal Coomassie and destained with water. Wet gels were scanned and used with phosphoimage screens to yield wet gel images by placing the wet gel on the screen for 2 h. Subsequently, the gel was vacuum-dried at 80 °C for 1 h on Whatman filter paper. Dried gels were placed on films and stored at -80 °C overnight and developed as described by the manufacturer.

5.3.8 (Chemical) crosslinking of protein-RNA complexes

For optimal results, LC-MS/MS analyses require 90 ug of protein and 1-2 nmol of RNA in general. Crosslinking radioactively labeled RNA with protein, however, was performed using 3 ug of protein and 1 pmol of hot RNA. Protein and RNA were incubated on ice for 30 min to reconstitute the protein-RNA complex. Subsequently, UV-crosslinking or chemical crosslinking was used. The different chemical crosslinkers were used as follows:
a fresh working stock of 500 mM DEB was made from 12.5M stock solution and immediately used. The protein-RNA complex was reconstituted in storage buffer lacking the 10% glycerol to which DEB was added at a final concentration of 50 mM. The mixture was incubated at 37 °C for 1 h to run the crosslinking reaction to completion. NM crosslinking experiments were also performed in said storage buffer, lacking the 10% glycerol. Similar to DEB crosslinking, a 7.6 mM final NM concentration containing reaction mixture was incubated at 37 °C for 1 h.

Successful 2-iminothiolane crosslinking first required the protein to be exchanged into 2IT buffer (25 mM TEA/HCl pH 7.8, 100 mM KCl, 5 mM Mg-acetate, 6 mM β-ME). The RNA was then added to reconstitute the protein-RNA complex on ice for 30 min. 2IT powder was dissolved in 500 mM TEA buffer and added to the reconstituted protein-RNA sample at various final concentrations. A final concentration of 2.7 mM was chosen if the 2IT powder has been used within the last four weeks for the first time. Alternatively, 20 mM were used if the 2IT powder has been used for more than four weeks priot to the experiment. Upon incubating the reaction mixture for 20 min at 20 °C, the cross-linked sample was UV-irradiated at 254 nm for 3 min. Any excess 2IT was quenched by adjusting the β-ME to 3 %. This reductive quenching step was performed at 30°C for 30 min, 300 rpm.

Lastly, the canonical UV crosslinking method was performed as a gold standard to compare chemical crosslinking to it. The method is described in detail by Kramer *et al.* [46]. Briefly, protein and RNA coalesced into a reconstituted complex on ice for 30 min and then UV-irradiated at 254 nm for 10 min using an apparatus built in-house. The apparatus emitted UV light at the defined wavelength of the five lamps at a distance to the sample of roughly 3 cm. The sample was irradiated in a polystyrene 96-well plate. Total energy emitted by the apparatus in this setting was estimated to be 81 Jcm^{-2} based on formula 9. A detailed derivation of that formula can be found in appendix A, section 13.1.1 (182). Following the crosslinking reaction, samples were acetone-precipitated overnight.

5.3.9 Acetone precipitation

Solutions containing either cross-linked multi-protein complexes or cross-linked whole-cell lysates were typically acetone precipitated with ice-cold acetone, stored at -20 °C akin to the ethanol precipitation mentioned above. Briefly, the protein solutions were precipitated by adding 4 volumes of ice-cold acetone and the entire mixture was kept at -20 °C overnight. Precipitated protein was centrifuged at 13,000 rpm for 30 min, and then washed in 80% ice-cold acetone. Following another round of centrifuging at 13,000 rpm for 10 min, the protein pellet was air-dried at 37 °C and subjected to a protein-RNA heteroconjugate clean-up and enrichment described in detail within [10, 46].

5.3.10 In-solution digestion of cross-linked protein complexes and lysates

100 ug protein complexes or whole cell lysates were dried by vacuum centrifugation as described above and resuspended in 50 µL 4 M urea (Tris-HCl, pH 7.9), at 600 rpm for no longer than 10 min at 25 °C. 150 µL of Tris:HCl, pH 7.9 were added to the resuspended pellet to dilute the urea concentration to 1 M. Any residual precipitates were fully resuspended before adding $MgCl_2$ to 1 mM, as well as benzonase to initiate nucleotide digestion. The mixture was incubated for 2 h at 37 °C before adding 1 µL of RNaseA and 1 µL RNase T1 stock enzymes to the mixture. The mixture was incubated for another 2 hours at 37 °C before cross-linked protein complexes were digested with trypsin in a 1:20 (w/w) ratio overnight. Whole-cell lysates from *E. coli*, *B. subtilis* or HeLa were additionally pre-digested with LysC in a 1:200 ratio before the overnight trypsin digest (1:50 (w/w)).

5.3.11 Desalting with C18 columns (Harvard apparatus)

C18 mini Desalting Columns from Harvard apparatus were used to remove excess salts, following the manufacturer's suggestions. Briefly, the C18 columns were conditioned with 500 µL 100% methanol, and all centrifugation steps were carried out at 110 x g. The column was equilibrated in successive rounds with 60 µL of various solutions: one round of (95% ACN, 0.1%), one round of (80% ACN, 0.1 % FA), one round of (50% ACN, 0.1 % FA), and one round of (0.1% FA). The trypsin-digested samples were adjusted to roughly 5% ACN and 0.1 % FA and incubated at 37 °C for 10 min prior to loading the samples in portions of 60 µL onto the equilibrated C18 columns. The columns were washed with two rounds of (0.1 % FA) before the sample was eluted into a fresh Eppenorf tube with one round of (10% ACN, 0.1% FA), two rounds of 50% ACN, 0.1 % FA), and one round of (80% ACN, 0.1 % FA). The eluted peptides were subsequently dried by vacuum centrifugation.

5.3.12 TiO_2 enrichment of nucleic-acid cross-linked peptides

Cross-linked peptide-nucleotide heteroconjugates were specifically enriched using $TiO2$ as previously described [10, 102, 103]. Dried enriched cross-linked sample pellets were ideally used immediately after the enrichment step, or kept at 4 °C for up to 2 days until subjected to LC-MS/MS. Pellets were resuspended in 2 µl 50% ACN, 0.1% FA. The sample was vigorously vortexed for 2 min before being diluted to a final concentration of 2% ACN, 0.05% TFA. In the case of purified complexes, 7 µL were injected for a single LC-MS run. In the case of highly complex samples from whole-cell lysates, only 4 µL of resuspended sample were injected into the LC-MS system. Peptide-nucleotide cross-linked samples were always measured on an OT mass spectrometer (Lumos, Fusion, QExactiv, Exploris) with top speed methods on MS2 described in section 5.3.20.

5.3.13 Anisotropy measurements for Hsh49

Fluorescence anisotropy measurements were performed as described by Kretschmer *et al.* and were performed by Gerald Aquino and Katharine Sloan in the Bohnsack laboratory (UMG, Göttingen) [164]. Briefly, Recombinant Hsh49 was first dialysed against anisotropy buffer (30 mM HEPES pH 8.0, 100 mM NaCl) overnight. Increasing concentrations of protein were incubated with 20 nM of fluorescently labelled RNA (poly-A, 5'-A_{11}-ATTO4488-3'; poly-C, 5'-C_{11}-ATTO488-3'; poly-G_5U, 5'-GGGUGGGUGGU-ATTO488-3'; poly-U, 5'-U_{11}-ATTO488-3') in anisotropy buffer for 5 min at room temperature. The samples were then transferred to a Quartz SUPRASIL®10x2 mm High Precision Cell cuvette (Hellma Analytics) for measurement. Anisotropy measurements were performed at 30 °C on a FluoroMax-4 spectrofluorometer (Horriba Scientific) with the FluorEssenceV3.5 software. The excitation and emission wavelengths were set to 500 nm and 520 nm respectively. The excitation slit width was 5 nm and emission slit width was 10 nm. The G factor for the polyA, polyC, polyG, and polyU RNAs was determined as 0.79, 0.78, 0.78, and 0.82 respectively. The integration time was 1 sec, the maximal trials per sample were set to 10 and the target standard error was 2%. The data were fitted with formula 1 described by Kretschmer *et al.* and dissociation constants were calculated using Origin 8.2 [164].

$$r = r_0 + \left(\frac{D_{max}}{[RNA]_{total}}\right)\left(\frac{([protein]_{total} + [RNA]_{total} + K_d)}{2}\right.$$
$$\left. - \sqrt{\left(\frac{[protein]_{total} + [RNA]_{total} + K_d}{2}\right)^2 - [protein]_{total}[RNA]_{total}}\right) \tag{1}$$

$$r_0 = \text{anisotropy of unbound RNA}$$

$$D_{max} = \text{amplitude}$$

$$[protein]_{total} = \text{total protein concentration}$$

$$[RNA]_{total} = \text{total RNA concentration}$$

$$K_d = dissociationconstant$$

5.3.14 NELF cloning, protein expression, and purification

The four-subunit NELF complex was cloned, expressed, and purified as previously described by Seychelle Vos from the Cramer lab, MPI-bpc, Göttingen [150]. The construct contains NELF-D, which is identical to NELF-C, but lacks the first nine amino acid residues. The Hi5 cell line (Expression Systems, Davis, CA, USA) was used to express the NELF complex. Hi5 cells expressing NELF were harvested by centrifugation, resuspended in Lysis buffer (300 mM NaCl, 20 mM HEPES pH 7.4, 10% (v/v) glycerol, 30 mM imidazole pH 8.0, 1 mM DTT, 0.284 µg/mL leupeptin, 1.37 µg/mL pepstatin A, 0.17 mg/mL PMSF, and 0.33 mg/mL benzamidine), flash-frozen, and stored at -80 °C until purification.

NELF was purified as previously described [150]. Briefly, Hi5 cells expressing NELF were lysed by sonication and clarified by centrifugation. The clarified lysate was loaded onto a 5 mL HisTrap column (GE Healthcare Life Sciences) equilibrated in Lysis buffer. The column was washed with 10 CV of Lysis buffer followed by 5 CV of High Salt buffer (800 mM NaCl, 20 mM HEPES pH 7.4, 10% (v/v) glycerol, 30 mM imidazole pH 8.0, 1 mM DTT, 0.284 µg/mL leupeptin, 1.37 µg/mL pepstatin A, 0.17 mg/mL PMSF, and 0.33 mg/mL benzamidine). The column was then washed with 5 CV of Lysis buffer followed by 5 CV of Low Salt buffer (Lysis buffer with 150 mM NaCl). A 5 mL HiTrap Q column equilibrated in Low Salt buffer was then attached to the base of the HisTrap column. The HisTrap column was developed over a gradient into Nickel Elution buffer (150 mM NaCl, 20 mM HEPES pH 7.4, 10% (v/v) glycerol, 500 mM imidazole pH 8.0, 1 mM DTT, 0.284 µg/mL leupeptin, 1.37 µg/mL pepstatin A, 0.17 mg/mL PMSF, and 0.33 mg/mL benzamidine) and was removed after being fully developed. The HiTrap

Q column was washed with 5 CV of Low Salt buffer and was developed over a gradient into High Salt buffer. Fractions were analyzed by SDS-PAGE and Coomassie staining.

Appropriate fractions were mixed with 1.5 mg of TEV protease and 416 µg of lambda protein phosphatase and transferred to a 3-12 mL 10 kDa MWCO Slide-A-Lyzer cassette. The protein dialyzed overnight against 1 L of Lysis buffer supplemented with 1 mM $MgCl_2$. The protein was then applied to a 5 mL HisTrap column equilibrated in Lysis buffer to remove TEV protease and uncleaved protein. The flow-through was concentrated in 30 kDa MWCO Amicon Ultra Centrifugal Filters and applied to a HiLoad S200 16/600pg column equilibrated in 150 mM NaCl, 20 mM HEPES pH 7.4, 10% (v/v) glycerol, and 1 mM DTT. Peak fractions were analyzed by SDS-PAGE and Coomassie staining. Pure fractions were concentrated in a 30 kDa MWCO Amicon Ultra Centrifugal Filters. The protein was aliquoted, flash-frozen, and stored at -80 °C.

5.3.15 NELF:TAR RNA sample preparation

The following section was performed in collaboration by Seychelle Vos from the Cramer lab, MPI-bpc, Göttingen. An RNA oligo corresponding to the HIV-1 TAR stem loop was synthesized by Integrated DNA Technologies (sequence: 5'-CCA GAU CUG AGC CUG GGA GCU CUC UGG-3') with and without a 5' -6 FAM label. The TAR stem loop was resuspended in water and stored at -80ºC. The TAR stem loop was folded (50 µM RNA, 100 mM NaCl, 20 mM HEPES pH 7.4, 3 mM MgCl2, and 10 % (v/v) glycerol) at 95ºC for 3 min, followed by ice incubation for 10 min. NELF and the folded TAR RNA were mixed at a 1:1.5 molar ratio, corresponding to 0.5-1 mg of NELF:TAR complex. The complex was dialyzed against 100 mM NaCl, 20 mM HEPES pH 7.4, and 3 mM $MgCl_2$ in 20 kDa MWCO Slide-A-Lyzer MINI Dialysis Unit for 16 hrs at 4ºC. Membranes were pre-wetted for 20 min-10 hrs in water prior to dialysis to remove residual PEG on the membrane surface.

5.3.16 Protein-Protein crosslinking of the NELF:TAR complex

Protein-protein crosslinking was performed in a similar manner described elsewhere [3]. Briefly, 200 ug of NELF:TAR complex resuspended in 50 mM HEPES buffer, pH 7.9, were cross-linked at a final concentration of 1 mM BS3. The reaction mixture was incubated at room temperature for 30 min before being quenched with 2.5 µL 2 M Tris, pH 8.1. The sample was dried by vacuum centrifugation and then redissolved in 6 M urea solution. Following incubation at room temperature for 20 min, the sample was alkylated with 20 µL of 60 mM IAA and the mixture was incubated at room temperature for 30 min. Lastly, the alkylated complex was trypsin-digested at 1:20 (w/w) overnight at 37 °C. Digested cross-linked peptides were then subjected to sepharose-based size exclusion chromatography (SEC) analagous to the Äkta settings mentioned elsewhere [3, 165]. The first fractions containing cross-linked peptides were pooled, dried by vacuum centrifugation, resuspended in MS injection buffer, and subsequently analyzed via LC-MS on an Orbitrap mass sepctrometer. Spectral data were analyzed using pLink2 and cross-links were visualized using XiNet as previously described [102, 166].

5.3.17 Expression of *Saccharomyces pombe* Dmnt2 and tRNAAsp

The sequence of the *S. pombe* tRNA$^{\mathrm{Asp}}$ (sptRNA$^{\mathrm{Asp}}$) was identified using gtRNAdb (locus chrI:4532857-4532927(+)30). sptRNAAsp was cloned into a pRAV23 vector and transcribed from the linearized vector template in run off in vitro transcription described in [5]. The reaction contained T7 Polymerase and 10 mM rNTPs each in 1× HT buffer (30 mM HEPES pH 8.0, 25 mM MgCl2, 10 mM DTT, 2 mM Spermidine, 0.01 % Triton X-100). After incubation for 3h at 37 °C, solids were removed by centrifugation and the reaction was stopped using 50 mM EDTA, pH 8.0. The transcription mixture was concentrated in Amicon Ultra centrifugation filter units with 3 kDa cutoff and run on a denaturing 10 % urea-containing Polyacrylamide gel. Target tRNA was excised from the gel, extracted (10 mM Tris-HCl pH 8.0, 1 mM EDTA, 300 mM NaCl) and ethanol precipitated overnight at -20 °C using 1/10 volume of 3 M sodium acetate pH 5.5 and three volumes of ethanol. RNA was pelleted and washed with a 70 % ethanol solution. The tRNA was then dissolved in RNAse-free water and stored at -20 °C until further use.

Saccharomyces pombe Dnmt2 (spDnmt2) was expressed and purified in collaboration by Sven Johansson as described within [5]. In brief, spDnmt2 was expressed from a pGEX 6P-3 vector as GST spDnmt2 fusion protein in *Escherichia coli* BL21(DE3) cells using autoinduction. Cells were disrupted by microfluidization (M-110S Microfluidizer) in 50 mM Tris-HCl pH 7.5, 150 mM NaCl, 1 M LiCl, 2 mM DTT. Debris was removed by centrifugation and the lysate was loaded onto Glutathione Sepharose FastFlow cartiges (50mM Tris-HCl pH 7.5, 150 mM NaCl, 2 mM DTT). GST-spDnmt2 was eluted with reduced glutathione and the tag was cleaved by addition of PreScission Protease. GST was removed subsequently by heparin-sepharose (GE Healthcare) purification (50 mM Tris-HCl pH 7.5, 100 mM-2M NaCl, 2mM DTT). In a final step, spDnmt2 was subjected to size exclusion chromatography (50 mM Tris-HCl, 150 mM NaCl, 2 mM DTT). spDnmt2 containing fractions were concentrated, flash frozen in liquid nirogen and stored at -80 °C until further use.

5.3.18 *In vivo* crosslinking of *Escherichia coli, Bacillus subtilis*, and HeLa cells

Escherichia coli BL21 DE3 cells and *Bacillus subtilis* cells were grown in LB medium to an OD_{600} of 0.6 before being cross-linked using either UV-irradiation or DEB treatment. HeLa cells were obtained from the Bioreactor facility in suspension (1,000 cells ml^{-1}) and about 3 x 10^6 cells were cross-linked en block either by UV-irradiation in petri dishes or by 50 nM DEB for 10 min. DEB-treatment was quenched by adding 50 mM Tris-HCl. *B. subtilis* and HeLa cells were only cross-linked with DEB and total cell contents were enriched and analyzed. *E. coli* Bl21 DE3 cells, however, were further fractionated as described below. UV-crosslinking is described in detail elsewhere [46]. In brief, mercury-based UV lamps emitting at $\lambda = 254$ nm were switched on 15 minutes prior to crosslinking to ensure homogenous irradiation. Cells were irradiated for 10 min at 4 °C followed by the sample preparation outlined above. Similarly, cells were DEB-treated with 50 mM DEB for 10 minutes at 37 °C whilst shaking at 300 rpm before subjected to S30/S100 fractionation by ultracentrifugation described elsewhere [167]. The same sample preparation workflow for isolated complexes outlined above was used to analyze complex *in vivo* cross-linked proteins. A growth kinetic for DEB-treated *E. coli* BL21 DE3 cells and untreated cells was generated from treating a 1:10 dilution of the same bacterial stock (LB medium, $OD_{600} = 0.6$) with 50 nM DEB. Optical densities were measured at various time points to assess the toxic effect of DEB on the growth of *E. coli* BL21 DE3. Time points included 0, 20, 40, 60, 80, 100, 120, 140, 160, 180, 200, 220, 240, 260, 280, and 300 min.

5.3.19 Mass spectrometry (MS/MS) and MS data analysis

Enriched samples were loaded onto a self-packed C18 column, and mounted to a Dionex Ultimate 300 UHPLC+ focused. Column dimensions included: 3 µm pore size, 75 µm in diameter, 30 cm in length (Reprosil-Pur^{TM} 120C18-AQ). Peptides were separated by reversed-phase chromatography on a 58 min multi-step gradient. The flow rate was set to 0.3–0.4 µL/min, and peptides were analyzed by either OT Fusion (Lumos) or OT HF(-X). MS1 spectra were recorded with a resolution of 120k (profile mode), whereas MS2 spectra were recorded with a resolution of 30k (centroid mode). The isolation window was set to 1.6 Th, preferred ions were set to have peptide-like isotopic distributions, and the dynamic exclusion was set to 9 s. Raw data of chemically connected RNA-protein hetero-conjugates were analysed and manually validated using the OpenMS pipeline RNP_{xl} and the OpenMS TOPPASViewer, as well as the KNIME platform, integrating openMS pipelines. The following sections describe instruments specific parameters set for certain experiments.

5.3.20 LC-MS methods

A self-packed C18 column was routinely used to separate peptides via reversed-phase chromatography. The LC method included the following specifications:

1. sample pickup: volume = 3 µL, flow = 20 µL / min
2. sample loading: volume = 8 µL, max. pressure = 600.00 bar
3. solvents: A = water, B = 80 % ACN
4. analytical column equilibration: volume = 5 µL, max. pressure = 600.00 bar
5. autosampler: flush volume = 100 µL

All samples were measured on an Orbitrap Fusion, Lumos, or QExactive HF-X from Thermo Fisher. The following section enumerates all instrument settings and parameters:

Method settings Application mode: peptide, method duration (min): 120

Global parameters Ion source type: NSI, spray voltage: static, positive ion (V): 2200, negative ion (V): 600, ion transfer tube temp (°C): 300, use ion source settings from tune: False

MS global settings Infusion mode: liquid chromatography, expected LC peak width (s): 30, advanced peak determination: True, default charge state: 2, internal mass calibration: Off

Experiment 1 (MS) Start time (min): 0, end time (min): 120, master scan: full scan, Orbitrap resolution: 60000, scan range (m/z): 375-1575, RF lens (%): 10, AGC target: custom, normalized AGC, target (%): 100, maximum injection time mode: custom, maximum injection time (ms): 104, microscans: 1, data type: profile, polarity: positive, source fragmentation: disabled

Filters MIPS, monoisotopic peak determination: peptide, relax restrictions when too few precursors are found: True, intensity, filter Type: intensity

Threshold Intensity threshold: 1.0e4

Charge State Include state(s): 2-7, include undetermined charge states: False

Dynamic Exclusion Mode: custom, exclude after n times: 1, exclusion duration (s): 6, 9, or 21

Mass Tolerance Unit: ppm, low: 10, high: 10, exclude isotopes: True, perform dependent scan on single charge state per precursor only: False

Data dependent Data dependent mode: cycle time, time between master scans (sec): 1.7,

ddMS^2 Multiplex ions: False, isolation window (*m/z*): 1.6, isolation Offset: Off, collision energy mode: fixed, collision energy type: normalized, HCD collision energy (%): 28, Orbitrap resolution: 30000, scan range mode: define first mass, first Mass (*m/z*): 100, AGC target: custom, normalized AGC target (%): 100, maximum injection time Mode: custom, maximum injection time (ms): 70, microscans: 1, data type: centroid

5.3.21 LC-MS FAIMS method

A 25 cm Aurora UHPLC series column (*ion opticks*) was used to separate peptides via reversed-phase chromatography. The LC method included the same specifications as mentioned above. All samples were measured on an OT-Exploris mass spectrometer with the FAIMS device mounted. In the following section the different parameters from the method outlined above are enumerated and FAIMS specific settings were written in bold :

Global parameters Ion source type: NSI, spray voltage: static, positive ion (V): 2000, negative ion (V): 600, ion transfer tube temp (°C): 300, use ion source settings from tune: False, FAIMS mode: standard resolution, FAIMS gas: static, **FAIMS gas (L/min): 0**

Experiment 1 (MS) Start time (min): 0, end time (min): 120, master scan: full scan, Orbitrap resolution: 60000, scan range (*m/z*): 375-1575, **FAIMS voltages: On, FAIMS CV (V): -35 (or any other desired CV)**, RF lens (%): 40, AGC target: custom, normalized AGC, target (%): 200, maximum injection time mode: custom, maximum injection time (ms): 50, microscans: 1, data type: profile, polarity: positive, source fragmentation: disabled

ddMS^2 Multiplex ions: False, isolation window (*m/z*): 1.6, isolation Offset: Off, collision energy mode: fixed, collision energy type: normalized, HCD collision energy (%): 30, Orbitrap resolution: 30000, TurboTMT: Off, scan range mode: define first mass, first Mass (*m/z*): 100, AGC target: custom, normalized AGC target (%): 100, maximum injection time Mode: custom, maximum injection time (ms): 70, microscans: 1, data type: centroid

Experiment 2 (MS) Start time (min): 0, end time (min): 120, master scan: full scan, Orbitrap resolution: 60000, scan range (*m/z*): 375-1575, **FAIMS voltages: On, FAIMS CV (V): -45(or any other desired CV)**, RF lens (%): 40, AGC target: standard, maximum injection time mode: custom, maximum injection time (ms): 50, microscans: 1, data type: profile, polarity: positive, source fragmentation: disabled

5.3.22 Targeted HCD LC-MS method

To hone the mass spectrometer in on potentially cross-linked peptides and remeasure potential MS2 spectra that contain a cross-linked peptide, OT Lumos instruments were either equipped with the FAIMS device or without it in a targeted LC-MS/MS run. LC-settings and FAIMS settings remained as described above to ensure comparability of the mass spectrometric runs. In addition, though, a targeted mass list containing marker ions (MI) of specific nucleosides and nucleotides was written into the method to specifically trigger a second MS2 event when a targeted mass was found. Specific *m/z* values for each nucleoside/nucleotide are listed in table 13. Up to 20 different initial MS2 scans could be re-trigged for an additional recording event. Details of the method included:

Mass Tolerance: ppm, Low: 15, High: 15, Use Groups: False, Trigger Only with Detection of at Least N Ions from the List: True, n: 1, Only Ion(s) Within Top N Most Intense: True, n: 20, Only Ion(s) Above the Threshold (Relative Intensity, %): False, Fail trigger if conditions are met: False

Targeted mass list: See table 13

Table 13: List of nucleotide specific marker ions present in MS2 spectra. Both nucleoside and nucleotide specific *m/z* values were used in a targeted HCD method and written into the specific mass target list that triggers a second MS2 recording event. The method does not utilize the top speed option previously used in MS2 recordings, but rather an upper limit of 20 initial MS2 spectra that may be re-triggered if a targeted MI mass was found.

Targeted mass list	
compound	*m/z*
uracil	113.0351
uridine	245.0774
cytosine	112.0511
cytidine	244.0933
guanine	152.0572
guanosine	284.0995
adenine	136.0623
adenosine	268.1046

5.3.23 RNP$^{\text{xl}}$ pipelines, settings, and manual annotation

Raw data files in their vendor format (.raw in the case of Thermo instruments) containing mass spectra are incompatible with the RNP$^{\text{xl}}$ pipeline. They need to be converted into the open access format .mzml that can be fed into the OpenMS node RNP$^{\text{xl}}$. MSConvert from the proteowizzard pipeline converts vendor specific files reliably into .mzml and has been used as the first step of data conversion [168, 169]. Next, converted data have been either fed into a pipeline of several different OpenMS nodes analogous to the approach described in Kramer *et al.* or have been used as direct input for an updated RNP$^{\text{xl}}$ node used successfully in recent publications [3, 5, 10, 46, 102]. The updated workflow encompassing the three essential nodes (RNP$^{\text{xl}}$Search, PercolatorAdaptor, and IDFilter) is depicted in figure 11.

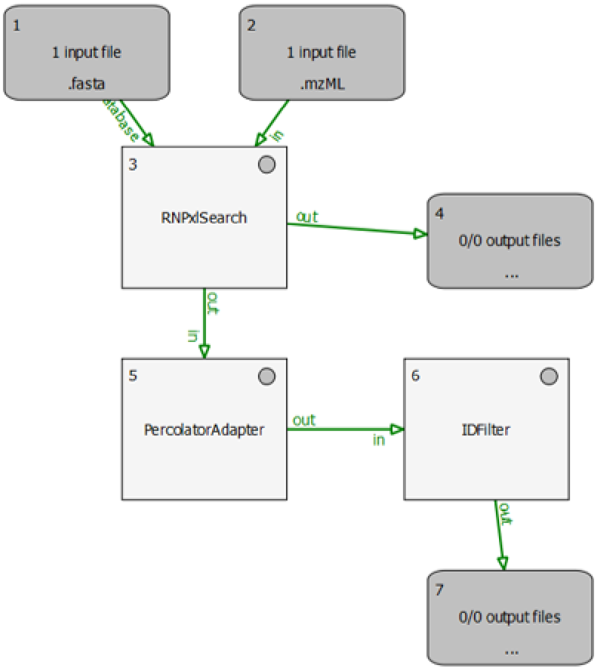
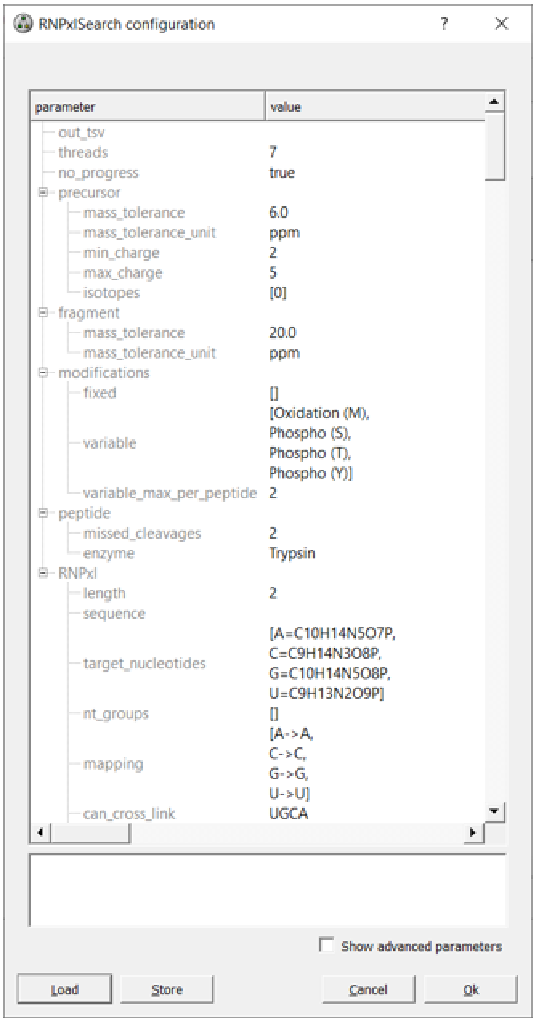


Figure 11: RNP$^{\text{xl}}$ Pipelines used to annotate spectra. Input data in .mzml format, as well as protein sequences in .fasta format are fed into the RNP$_{\text{xl}}$Search node to identify potential crosslinking spectra. Search results are both directly exported as a .csv file and queued into the PercolatorAdapter node that reranks and rescores spectral hits. The generated output is filtered against linear peptides and phosphopeptides, resulting in an output .idxml file that can be used to to annotate spectra using TOPPView. Note that the IDfiltered output does not contain any linear peptides or simple phosphopeptides to reduce the number of spectra to analyze by manual annotation. Information on linear peptides and phosphopeptides that can still be found in the RNP$^{\text{xl}}$Search output file (.idxml).

Most importantly, the following RNP$^{\text{xl}}$ Search node settings, shown in figure 12 were applied to search spectral data. To remove ambiguous phosphopeptides later, the search node was configured to include phosphorylations on serine, tyrosine, and threonine. Additionally, oxiations on methionine were also permissible. Allowing two missed cleavages in tryptic peptides in addition to limiting the number of cross-linked nucleotides to two dramatically reduced the amount of possible spectra in conjunction with a strict ppm deviation of 6 ppm on MS1 and up to 20 ppm on MS2. Also, precursors of charge state +1, as well as any charge state higher than +6, were excluded from the search. The cross-link specific parameters to be searched against as an adduct on the peptide were quite extensive and are tabulated in Appendix A, section 13.1.2 (page 184). Precursor modifications included neutral losses of water, phosphate, and phosphoric acid. To reduce the number of potential spectra, only methionine oxidation was considered as a variable peptide modification besides phosphorylation on serine, tyrosines, and threonines. Other identified mono crosslinking adducts identified from single crosslinking events were not used in the RNP^{xl} search, but are shown in figures 60 and 61 of appendix A for DEB and NM, respectively.



```
parameter                        value
out_tsv
threads                          7
no_progress                      true
precursor
    mass_tolerance               6.0
    mass_tolerance_unit          ppm
    min_charge                   2
    max_charge                   5
    isotopes                     [0]
fragment
    mass_tolerance               20.0
    mass_tolerance_unit          ppm
modifications
    fixed                        []
                                 [Oxidation (M),
                                 Phospho (S),
    variable                     Phospho (T),
                                 Phospho (Y)]
    variable_max_per_peptide     2
peptide
    missed_cleavages             2
    enzyme                       Trypsin
RNPxl
    length                       2
    sequence
                                 [A=C10H14N5O7P,
                                 C=C9H14N3O8P,
    target_nucleotides           G=C10H14N5O8P,
                                 U=C9H13N2O9P]
    nt_groups                    []
                                 [A->A,
                                 C->C,
    mapping                      G->G,
                                 U->U]
    can_cross_link               UGCA
```

Figure 12: Settings of the RNP^{xl} search node used to identify CSMs. Notice that the MS1 precursor variance was set stringently to 6 ppm to ensure accurate precursor masses. Precursors of charge 2-5 were triggered for an MS2 recording event. MS2 precursor variance was set conservatively to 20 ppm. Additionally, phosphorylation on S, T, and Y, as well as methionine oxidation were allowed as peptide modifications. Lastly, specific adduct masses to peptide fragments corresponding to cross-linked nucleotides were tabulated according to their respective chemical crosslinker in tables found in section 13.1.2.

Lastly, manually validated suggestions from the RNP^{xl} nodes follow a set of base criteria. These fundamental rules have been laid out as absolute minimal requirements for potential CSMs to be validated. The rules are as follows:

1. The signal to noise ratio (S/N) should be greater than 5% for peaks to be counted towards sequencing the peptide or indicating a cross-linked peptide with an adduct.
2. At least three shifted ions need to present in the spectrum to reliably indicate a cross-link spectrum match.
3. Marker ions of the cross-linked adducts need to be present in the spectrum.
4. Sequence coverage of the peptide should exceed 80%.
5. Corresponding survey scans on MS1 should display less than 5% interference form co-isolated species.
6. Cross-linked phospho-peptides were excluded.
7. The number of cross-linked adducts should not exceed two nucleotides.

5.3.24 Data analysis post RNP^{xl}/ NuXL output and machine learning

Output data in .csv format from either RNP^{xl} or NuXL served as the master table in subsequent data cleaning, wrangling, mining, and supervised learning approaches. Those data analyses were performed using R, version 3.6.2, embedded into RStudio, version 1.3.820-v1. Various different R packages have enabled and facilitated data analyses, most of which are mentioned above in regards to their specific function, but baseline analyses most heavily relied on the following packages, all updated to their newest releases from the Comprehensive R Archive Network (CRAN) (February, 2021):

Table 14: List of R packages used for different analyses. Various R packages were used to meta-analyse the collected data. Every package is listed with its corresponding version number, as well as its original publication. Different R scripts for disparate analyses contain various combinations of such packages and can be found on the accompanying CD.

R packages		
Package	Version	Citation
tidyverse	version 1.3.0	[170]
tinytex	version 0.19	[171]
kableExtra	version1.1.0	[172]
Peptides	version 2.4.1	[173]
caret	version 6.0.85	[174]
ggpubr	version 0.2.3	[175]
corrplot	version 0.84	[176]
readxl	version 2.0.0	[177]
pander	version 0.6.3	[178]
devtools	version 2.2.1	[179]
Hmisc	version 4.3.0	[180]
RColorBrewer	version 1.1.2	[181]
knitr	version 1.28	[182]
magrittr	version 1.5	[183]
ggcorrplot	version 0.1.3	[184]
gridExtra	version 2.3	[185]
car	version 3.0.6	[186]
RMarkdown	version 2.1	[187]

Machine learning strategies revolved around supervised machine learning approaches, focusing strongly on K-nearest neighbors (k-NN) and decision tree models. 150,000 manually evaluated spectra from both UV and DEB-crosslinking experiments were used to built an extensive data base used for both training, tuning, and validation purposes. Upon harmonizing the heavily skewed data set of roughly 830 true positives and 830 randomly selected negatives, classification of spectral data into true positive CSM hits or false positives was built on the exhaustive numeric data computed from the RNP$_{xl}$ output. Computation of additional sequence-driven features such as peptide length, pI, hydrophobicity, mw, and aliphatic indices were used to classify data [188].

In total, 39 numeric features were chosen to classify the data along multiple dimensions to check for dependencies and most correlating features. 31 features were taken from the RNP^{xl} output csv file, whilst 8 features were computed from the identified peptide sequences using the appropriate R script. The original 31 features from RNP^{xl} were based on an improved version of the original RNP$_{xl}$ used previously. [5, 46, 102] The following numeric metrics were calculated in addition to the RNP$_{xl}$ output: peptide length, pI of the peptide, aliphatic index (amino acids A, V, I, and L), number of charged amino acids (R, K, D, and E), number of polar amino acids (Q, N, H, S, T, Y, C, and W), number of hydrophobic amino acids (A, I, L, M, V, P, and G), the molecular weight of the peptide, and the percent composition of each amino acid class (charge, polar, hydrophobic). All 39 numeric features are listed in table 20 and all R scripts can be found in appendix C, section 13.3.

Next, the supervised learning algorithm of k-NN, published in [174], was applied and the predicted results were checked against their reference outcomes. Numeric data was min-max-normalized prior to applying the algorithm to re-scale all data points of the same feature into the same dimension. Additionally, C.5 and RIPPER classification algorithms for decision trees from [189] were utilized in secondary classification rounds to compare best model performance. Possible penalties for either true or negative classifications were set to fine-tune model parameters in the case of C5.0 to boost classification accuracy. Specifically, a false positive was penalized by a factor of 4 compared to a possible false negative outcome to reduce the false discovery rate of background noise being fitted and misclassified. Additionally, an upper limit of nine rules was set not to overbranch the decision tree. A comprehensive example with inclusive lines of codes can be found as a pdf on the accompanying CD.

Built models from the k-NN supervised machine learning approaches were further used to classify unknown RNP$_{xl}$ output features that were augmented through computed sequence-driven parameters mentioned above. Potential positives were then manually validated to ensure spectral integrity and quality to accurately confirm peptide-RNA-cross-links. Indiscriminate spectra were discarded and resulting lists of cross-linked proteins were functionally curated using Uniprot, GenePanther, and StringDB. Selected examples were structurally analyzed using PyMOL and published crystal structures as indicated in the text.

6 Results

Obtained results can be grouped into three main blocks in accordance with the aims of thesis. First, proof of concept studies involving only one type of protein (Hs49) and synthetic oligo-nucleotide RNAs were used to systematically investigate crosslinking potential of the chemical crosslinkers DEB, NM, and 2IT and to deduce adduct formulae and adduct masses with corresponding shifts from mass spectrometric data. Upon identifying the correct adduct masses and expected shifts of fragment ions in MS2 spectra, DEB-mediated crosslinking sites in Hsh49:Cus1 with the snRNA U2, Dnmt2 with its tRNA$^{\mathrm{Asp}}$, and the NELF complex with its HIV TAR RNA were analyzed *in vitro*. In that way, the RNA-binding interface in Hsh49 could be better characterized, a previously published model of the Dnmt2 protein binding its tRNA$^{\mathrm{Asp}}$ could be refined based on novel crosslinking sites from chemical crosslinking, and the importance of flexible regions within the NELF complex could be explained based on crosslining sites. Extending chemical crosslinking with DEB to *in vivo* systems posed the next challenge, piloting *in vivo* crosslinking in *E. coli* before applying it to *B. subtilis* and HeLA cells. A marker ion-centric approach was devised to improve confidence in obtained spectra displaying true crosslinked hits by retriggering MS2 events of potentially crosslinked ion species and to reduce the copious number of sectral hits. In an attempt to deal with hundreds of thousands spectra per set of experiment, a supervised machine learning approach was devised to classify unknown spectral data into true hits and false positives before the novel NuXl algorithm was available. Lastly, chemical crosslinking coupled with mass spectromtetry was combined with FAIMS seperation of ion species to enhance crossliking site identification in *in vivo* crosslinked samples from *E. coli*. Various FAIMS settings were scrutinized for increased crosslink output to identify possible preferences in settings in this preliminary FAIMS study.

6.1 Hsh49:Cus1 Purification

The original protocol to purify Hsh49 from Van Roon *et al.* states that Hsh49 was purified by the authors in a three step purification process, comprised of two affinity purifications with NiNTA and one heparin column [146]. The original protocol was adapted so that Hsh49 was purified solely in an abridged version of the three-step process to reduce loss of material [144]. In this shortened purification process, a single NiNTA column is used consecutively. *E. coli* BL21 DE3 cells harboring pET-Hsh49-His$_6$ are lyzed by high pressure homogenization. Upon His-tag purification during the NiNTA I step, the purified Hsh49:His$_6$ is subjected to a TEV dialysis to remove the His$_6$-tag. Dialysed Hsh49 is then loaded onto NiNTA II to remove highly basic proteins. The flow-through containing Hsh49 is then collected, buffer exchanged, and concentrated. The Hsh49:Cus1 complex was co-purified in a similar fashion. While combining both NiNTA I and NiNTA II with a third purification step (a heparin affinity column) yielded the purest sample, yields were greatly increased once the second NiNTA step was omitted (data not shown). Loading the pooled eluates from NiNTA I directly onto the heparin column resulted in improved yields and a substantial decrease in purification time. Figure 13 illustrates both purification strategies for Hsh49 alone and for the Hsh49:Cus1 complex.

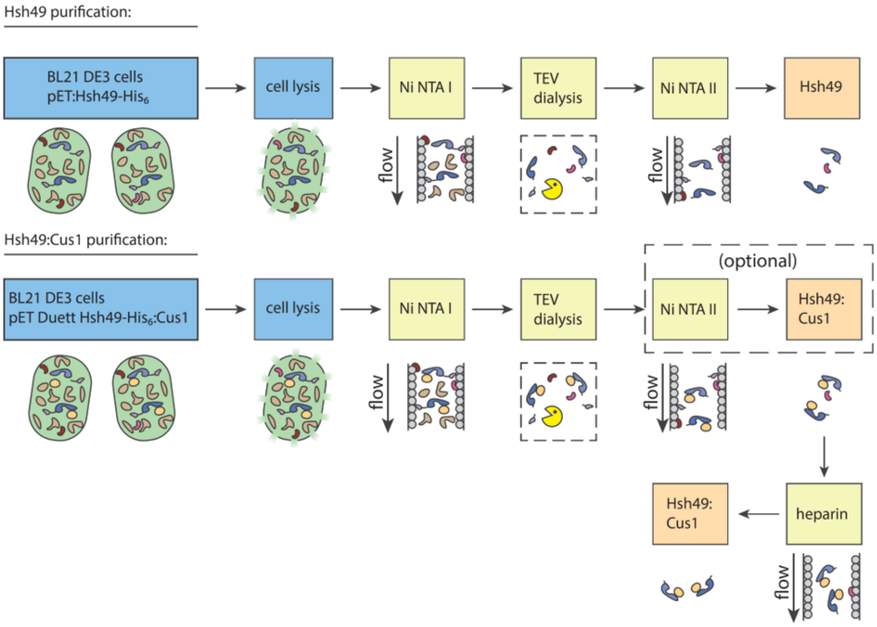


Figure 13: Hsh49 and Hsh49:Cus1 purification strategy. *E. coli* Bl21 DE3 cells expressing Hsh49 are lysed mechanically and subjected to a three-step purification process based on affinity chromatography. In the first step, the crude lysate is loaded onto a NiNTA affinity column, followed by a second NiNTA column after an overnight TEV digest. His-tagged Hsh49 is initially retained during the first step, but elutes close to the void volume after the TEV-digest during the second step. Hsh49:Cus1 complex is analogously purified via NiNTA I, and the second NiNTA may be regarded as optional, as it increases the purity of the sample at the cost of time and material. The polishing step concludes with a heparin column similar to the Hsh49 purification.

SDS-PAGE analyses of the sequentially purified protein illustrated the increase in protein purity, as depicted in figure 14 for Hsh49 alone. The sequential increase in purity from each step, ranging from crude lysates in figure 14 A), followed by the first NiNTA purification step was most drastic as only proteins with sufficiently large numbers of positively charged histidines should have allowed for efficient binding to the NiNTA column. Figure 14 A) depicts the later eluting fractions of the first NiNTA gradient that included a high salt wash to remove other nucleic acids such as DNA unspecifically bound to protein. In B), however, the first fractions after the void volume contained the target protein, as its His-tag was previously abscised and thus failed to bind to the column. The polishing step included a heparin column that increased protein purity dramatically as very few contaminating proteins could be identified during SDS-PAGE analysis and subsequent LC-MS analysis. This increased purity was achieved at the cost of a lower purified protein yield, however.

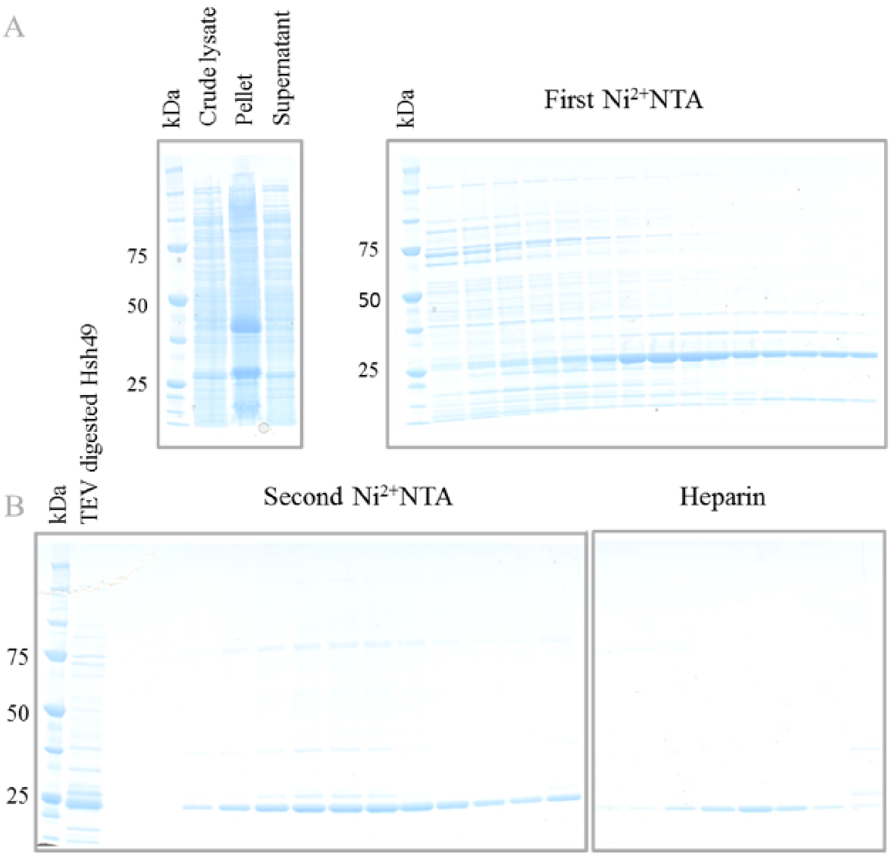


Figure 14: SDS-PAGE of sequentially purified Hsh49 protein A) Crude lysate from *E. coli* expressing Hsh49 was cleared by centrifugation, yielding a pellet and supernatant. During the first NiNTA affinity chromatography, Hsh49 elutes with already reasonable purity from the affinity column, yet an overnight TEV dialysis with subsequent separation on a second NiNTA column reduced the amount of contaminating proteins dramatically (**B)**. The final step involves a Heparin column, assuring highest purity in Hsh49 protein, but suffering from moderate loss in purified protein amounts.

6.2 Hsh49 crosslinking

To investigate the suitability of the crosslinking reagents DEB, NM, and 2IT for chemical crosslinking coupled with mass spectrometry in the context of protein-RNA crosslinking, a simple system was first utilized: one protein bound to a defined synthetic RNA. Hsh49 alone was chosen for its simple structural composition of two RRM domains joined by a short flexible linker region, and well-defined RNA binding properties.

6.2.1 Hsh49 crosslinked to synthetic poly-oligonucleotides

Chemical crosslinking reagents potentially display varied crosslinking behavior to different nucleotides based on nucleophilicity of the bases. This idea was tested by SDS-PAGE analyses of DEB, NM, 2-IT and UV-light cross-linked Hsh49 with [γ^{32}P]-labeled poly(A)$_{25}$, poly(C)$_{25}$, poly(G)$_{25}$ or poly(U)$_{25}$. Adhering to a simple one-protein-RNA-complex system initially, buffer compositions, crosslinker concentrations, amount of protein used, and length of the crosslinking time were all optimized (data not shown). The model protein Hsh49 was used to investigate potentially different crosslinking preferences of the crosslinkers towards certain nucleotides systematically using the optimized conditions in a feasibility study using ^{32}P radioactively labeled RNA with Hsh49 alone.

Figure 15 depicts SDS-PAGE analyses of Hsh49 protein reconstituted with various poly-nucleotide RNAs, cross-linked either chemically with different crosslinkers or using UV-irradiation. Figure 15 A shows residual uncut His-tagged Hsh49 and prominent cut Hsh49 bands around 25 kDa with faint higher molecular aggregates appearing in cross-linked samples. Autoradiograms for each individual poly-nucleotide are depicted in figures 15 B-E with radioactive signals around the Hsh49 band appearing in cross-linked samples. The prominent lower band corresponds to unbound radioactive RNA and the upper band appearing at 50 kDa is most likely attributed to two Hsh49 proteins bound to one RNA. UV-cross-linked samples usually yielded stronger signals than chemically cross-linked samples. To evaluate Hsh49 binding to poly-nucleotides quantitatively, fluorescence anisotropy experiments as shown in figure 15 F were performed. Here, poly-nucleotides of adenine, cytosine, and uracil did not resulted in good signals when bound to Hsh49 to evaluate a K$_\mathrm{d}$. A purely guanosine-containing poly-G-RNA is proven impossible to synthesize as G nucleobases form a G quadruplex easily *in vitro* which would impair synthesis of the polynucleotide, as well as potentially protein binding to the RNA. To circumvent this problem, a poly nucleotide of GGGUGGGUGGU (poly-G$_5$U) was used instead. In the case of poly-G$_5$U, however, a K$_\mathrm{d}$ of 2.4 µM was derived from a calculated fit.

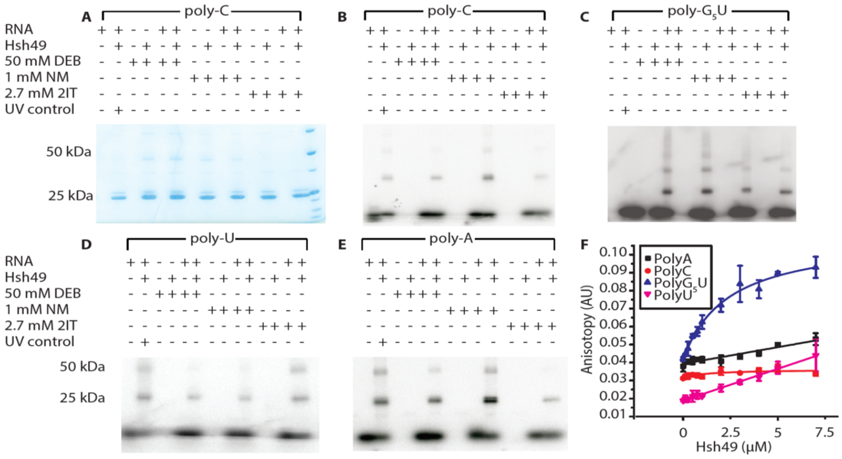


Figure 15: SDS-PAGE analysis of Hsh49 incubated with radioactively labeled poly-nucleotides. A slight smear in lanes containing cross-linked protein and poly-C RNA appears in the Coomassie-stained gel shown in **A**, indicating a shift to higher molecular mass. This crosslinking event could easily be detected radioactively as shown in **B-E** by the presence of a radioactively labeled RNA signal appearing at the molecular weight of Hsh49. Additionally, faint bands of higher molecular weight indicate cross-linked Hsh49 multimers such as dimers, all cross-linked chemically and by UV-irradiation to the radioactively labeled RNA. Quantifiable signal from flourescence anisotopy experiments confirms binding of poly-G_5U to Hsh49 (Kd = 2.4 µM), but other poly-nucleotide RNA failed to provide sufficient signal to calculate Kd-values (**F**). *Flouresence anisotropy experiments performed by Gerald Aquino and Kate Sloan.*

To confirm findings from the highly sensitive autoradiograms with labeled poly-nucleotides mass spectrometrically, an established MS-based preparative and analytical workflow for UV-cross-linked samples was adopted to fit chemically cross-linked samples, as shown in figure 16. The workflow is based upon subjecting cross-linked samples to RP-chromatography to remove excess RNA followed by selectively enriching for cross-linked peptide-RNA heteroconjugates using titanium dioxide as described in detail in sections 5.3.11 and 5.3.12. Following LC-RP chromatography separation, the cross-linked peptides were subjected to mass spectrometric analyses, followed by subsequent identification of potential hits by RNP_{xl}, and validated manually.

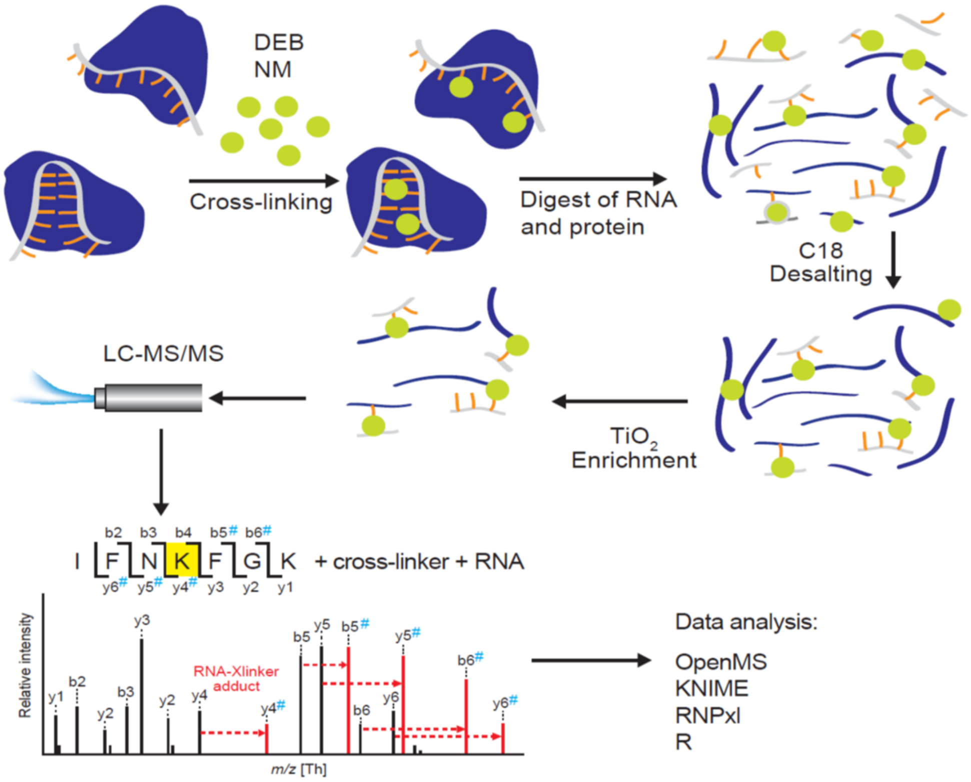


Figure 16: Schematic workflow of the chemical crosslinking process coupled with mass spectrometry. Enrichment strategy for protein-RNA heteroconjugates. First, RBPs are reconstituted and crosslinked. Next, both RNA and protein are digested enzymatically, followed by removal of non-crosslinked nucleotides through C18 chromatography. Next, linear peptides are removed by TiO2 enrichment, leaving only crosslinked peptides to be subjected to ESI-LC-MS/MS in the subsequent step. Lastly, data are analyzed using OpenMS, KNIME, RNP$_{\text{xl}}$, and R to identify, manually validate, and summarize crosslinked peptides.

Specific mass shifts, corresponding to different nucleotide fragments and chemical crosslinker were manually evaluated and are listed in appendix A table 17, table 18, and table 19 for 2IT, DEB, and NM, respectively. In these proof of concept experiments, prominent cross-links with good spectral quality to assess interacting amino acids in the case of DEB crosslinking for poly-urdine, poly-guanidine, and poly-cytidine RNA were found. Representative spectra are shown in figure 17 A-D. The output from the RNP$^{\text{xl}}$ software was manually curated by carefully inspecting potential spectra and verifying specific mass shifts. DEB-mediated histidine cross-links such as the one found in figure 17 A) were identified quite often and expand the panel of crosslinking amino acids reliably. This histidine cross-link to a uracil nucleotide was particularly well supported by multiple different nucleotide fragments on the peptide. Ample evidence from multiple different fragment adducts was commonly observed with uracil nucleotides, featuring entire nucleotides, nuceosides, and nucleobases attached to the peptide with different neutral losses. Canonical lysine crosslinking sites such as the one found in figure 17 B that were also identified in classical UV-crosslinking with a U nucleotide demonstrates the sensitivity and accuracy of the approach, although the residue was found to be cross-linked to guanine in this instance. In general, guanosine cross-links seemed to fragment rather specifically with only the nucleobase attached to the peptide via DEB. Cytosine cross-links were usually impossible to discern when adducts only contained the nucleobase with an ammonia loss since it is equivalent to an uracil nucleobase losing a water moiety. However, using only poly-C nucleotides allowed for exact identification of both nucleotide and peptide as shown in figure 17 C, as adduct masses were usually comprised of nucleobase cytosine with a neutral loss of water or nucleotide cytidine crosslinked via DEB. Only in the case of poly-adenine RNA, the exact interacting amino acid could not be located unambiguously, even though strong signals corresponding to the precursor attached to adenoside nucleotides were present (fig. 17 D). All spectra contain multiple annotated peaks corresponding to peptide fragments, shifted uniquely based on the added nucleotide fragment. A table of possible fragments can be found in appendix A, table 18.

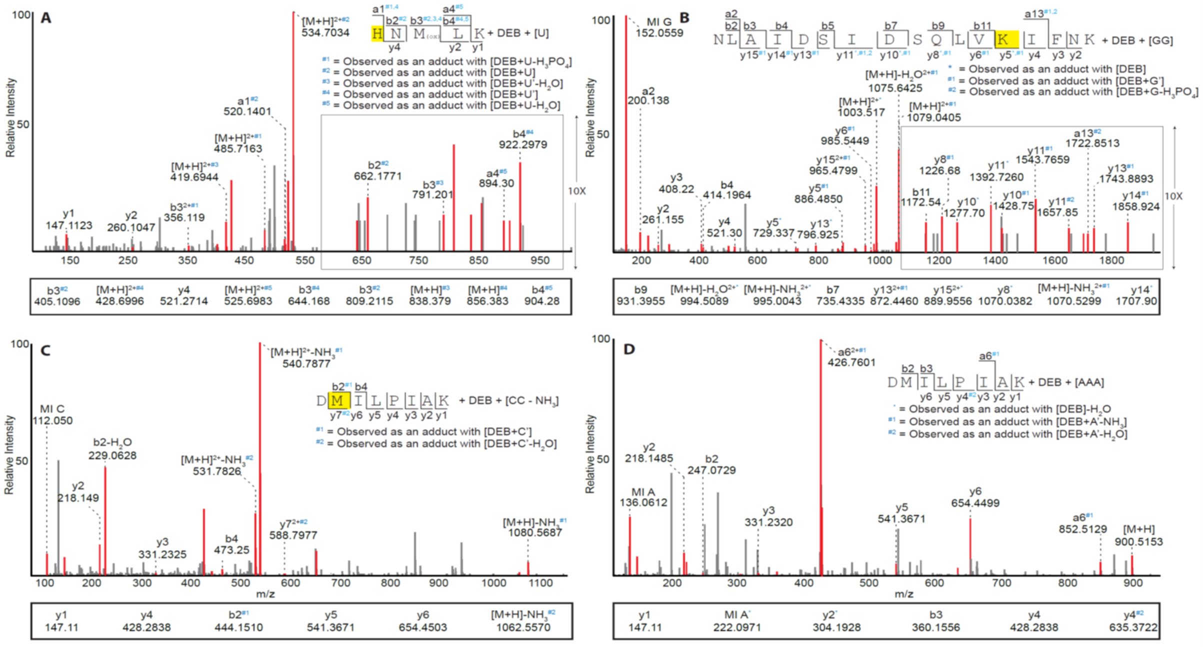
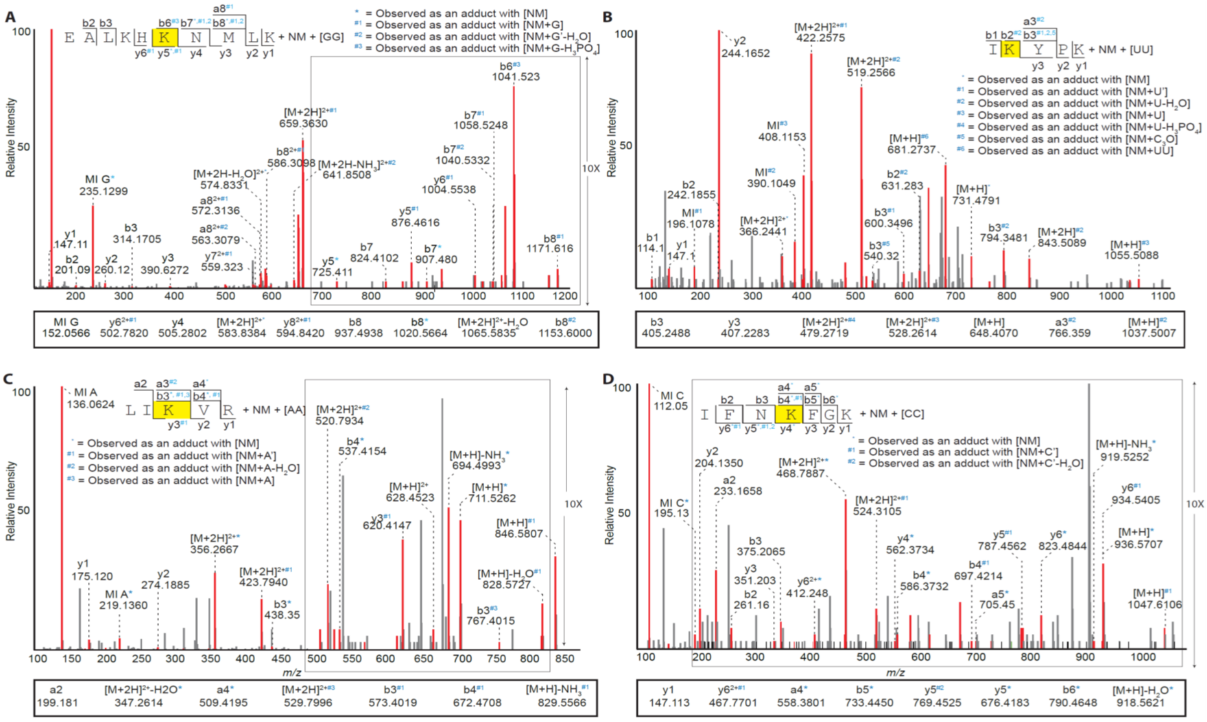


Figure 17: Exemplary manually curated MS2 spectra of Hsh49 crosslinked to poly-nucleotides via DEB. Different peptides were identified to be crosslinked to U, G, and C nucleotides (A-C), as indicated by specific adduct masses, locating the crosslinking amino acid specifically and unambiguously. In the case of poly-A (D), however, the exact crosslinking amino acid could not be determined unambiguously, as the peptide-RNA heteroconjugate remains largely unfragmented. The crosslinking amino acid is highlighted in yellow.

Similarly, high-quality spectra for NM-cross-linked Hsh49 for all four poly-nucleotides were obtained, allowing for manual validation of crosslinking amino acids unambiguously (figure 18 A-D). Here, a surprising number of lysine cross-links emerged from the data, suggesting that NM prefers lysines crosslinking sites given those conditions. The identified poly-G cross-links such as the one found in figure 18 A usually comprise of both nucleotides and nucleobases attached to the peptide, differing from the almost exclusive behavior of attaching a nucleobase observed in DEB. Uracil-containing fragments constituted a diverse group with nucleobases, nucleosides, and nucleotides and different neutral losses akin to DEB-mediated poly-U cross-links. Circumventing the ambiguity in cytosine cross-links due to neutral losses compared to uracil by using synthetic poly-C RNA, reliable crosslinking sites to cytosine adducts could be identified for NM. This way, both DEB and NM yielded crosslinking sites at the amino acid level due to stabilized nucleotides not undergoing ambiguous neutral losses. Surprisingly, even poly-A nucleotides could be reliably identified from the data, closing the gap of unlocalized crosslinking sites for all four nucleobases. crosslinking amino acids were generally nucleophilic in the cases of DEB and NM, extending crosslinking amino acids beyond lysines and tyrosines that were typically found in UV-crosslinking. Specific adduct masses for NM crosslinks are listed in appendix A, table 19.



MI G	y6$^{2+\#1}$	y4	[M+2H]$^{2+*}$	y8$^{2+\#1}$	b8	b8*	[M+2H]$^{2+}$-H_2O	b8$^{\#2}$
152.0566	502.7820	505.2802	583.8384	594.8420	937.4938	1020.5664	1065.5835	1153.6000

b3	y3	[M+2H]$^{2+\#4}$	[M+2H]$^{2+\#3}$	[M+H]	a3$^{\#2}$	[M+H]$^{\#2}$
405.2488	407.2283	479.2719	528.2614	648.4070	766.359	1037.5007

a2	[M+2H]$^{2+}$-H2O*	a4*	[M+2H]$^{2+\#3}$	b3$^{\#1}$	b4$^{\#1}$	[M+H]-$NH_3$$^{\#1}$
199.181	347.2614	509.4195	529.7996	573.4019	672.4708	829.5566

y1	y6$^{2+\#1}$	a4*	b5*	y5$^{\#2}$	y5*	b6*	[M+H]-H_2O*
147.113	467.7701	558.3801	733.4450	769.4525	676.4183	790.4648	918.5621

Figure 18: Exemplary manually curated MS2 spectra of Hsh49 crosslinked to poly-nucleotides via NM. Different peptides were identified to be crosslinked to U, G, C and A nucleotides (A-D) based on specific mass shifts corresponding to different RNA-adducts on peptide ions. Note that NM captures all nucleotides specifically at the amino acid residue level.

Lastly, 2IT-mediated cross-linked peptides to RNA could also be identified reliably on the MS2 level. As depicted in figure 19, 2IT always reacts on the lysine amino acid residue due to the chemical nature of converting primary lysines into reactive thiols. Other amino acids could not be reliably detected even when they were included in the search. While a restriction on the amino acid to solely lysine facilitates unambiguous identification of potential crosslinking sites, it is exclusively dependent on the presence of RNA-interacting lysines. To compound on this restriction, only lysine cross-links to uracil nucleobases could be detected even though all four synthetic poly-nucleotide RNAs were used.

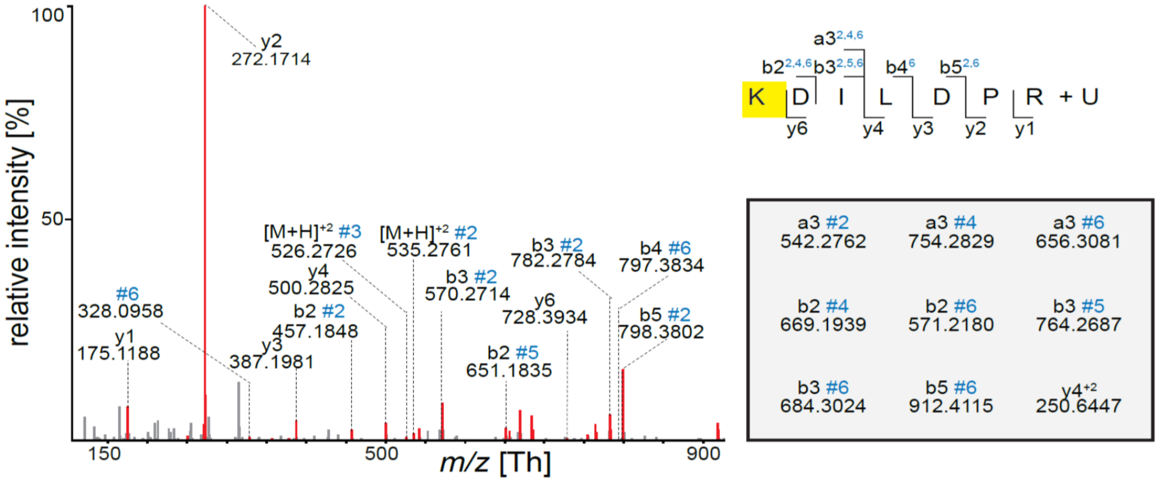


Figure 19: Exemplary manually curated MS2 spectra of Hsh49 crosslinked to poly-uracil nucleotides via 2IT. Akin to classical UV-crosslinking, 2IT is photo-activated and was only detected with uracil nucleotides. The cross-linked amino acid lysine is chemically converted into a reactive thiol that mediates the crosslinking reaction to the RNA, restricting the crosslinking amino acid severely to just lysines. *Figure adapted from master thesis "RNA-Protein crosslinking by 2-Iminothiolane and Nitrogen Mustard" by Luisa M. Welp.*

While the chosen test system constituted a highly artificial environment with one protein being exposed to different synthetic RNAs, it allows to first identify the specific mass shifts for each crosslinker. Having observed an incisive preference for uracil nucleotides in UV and 2IT crosslinking, DEB and NM show a wider chemical reactivity for nucleotides in general and even a preference for guanosine nucleotides. The above analyses with the RNA binding protein Hsh49 and non-specific RNA oligonucleotides showed that chemical crosslinking extends insights gained from UV crosslinking by expanding the pool identifiable nucleotides. Mapping the crosslinking amino acids to the crystal structure published by Van Roon *et al.* shows that crosslinks using artificial RNA and Hsh49 alone appear in both RRM domains as depicted in figures 20 for UV, DEB, and NM and in figure 21 for 2IT [144].

As depicted in figure 20, most crosslinks were identified in RRM2 with three regions being targeted by both UV-crosslinking and chemical crosslinking (a complete list of crosslinked peptides can be found in Appendix C, section 13.3). There is an additional region in RRM2 upstream from the last commonly identified region that was exclusively detected by chemical crosslinking in this instance. Also, the flexible linker regions that could be crystallized so far share multiple crosslinking sites with the two approaches. Congruency between the two approaches is lifted in RRM1, where only chemical crosslinking identified RNA-interacting amino acids (figure 20 A). Most surprisingly, NM outperformed any other crosslinking method in terms of identified crosslinking sites, even though DEB-crosslinking has been generating approximately 25% more CSMs per identified site than NM-crosslinking in previous pilot experiments. Lastly, chemically crosslinked amino acids generally favor nucleophiles within the side chains of K, C, R, Y, or S (figure 20 B-D).

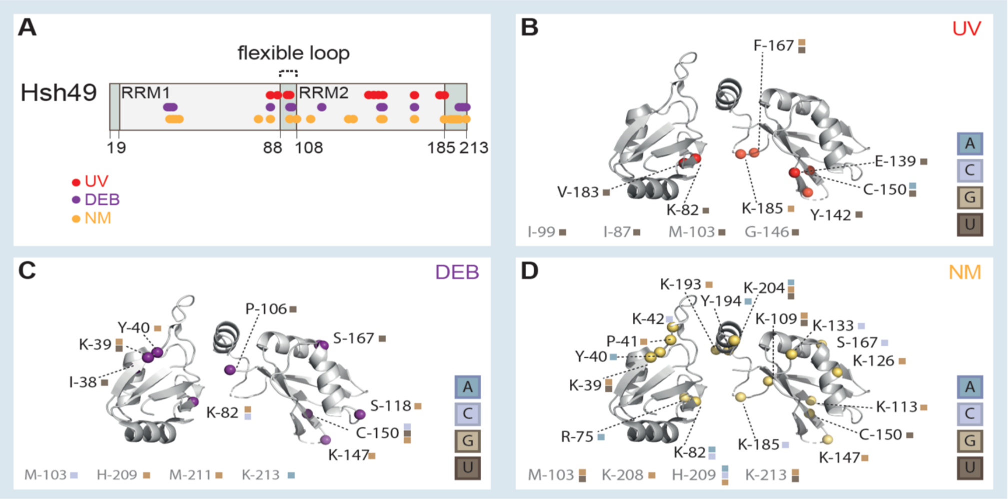


Figure 20: Structre of Hsh49 with mapped crosslinks from synthetic RNAs. A) Schematic depicting the positions of Hsh49 crosslinking sites within the Hsh49 protein. Note that most crosslinking sites fall into the RRM2 domain and a considerable amount of crosslinks was found in the flexible linker region. Interestingly, the two chemical crosslinker DEB and NM behave very similarly in regards to the actual crosslinking positions within the protein, both also hinting at a region in RRM1. Crosslinked sites identified from UV-crosslinking still correlate roughly with the sites identified with chemical crosslinkers. **B)** Mostly uracil containing crosslinking sites predominantly in RRM2 and the linker region that could not be crystallized due to its flexibility. **C)** DEB-mediated crosslinking identifies mainly nucleophilic residues in both RRMS to all four nucleotides. **D)** Extensive crosslinking sites to all four nucleotides from NM-mediated experiments using artificial RNAs. Similar to DEB crosslinking, regions in both RRMS were identified to be in direct contact with the RNAs, as well as the flexible linker region.

2IT cross-linked amino acids were by nature restricted to lysines due to the chemical transformation of the primary amine of the side chain. Even so, only uracil nucleotides were identified to be interacting with the chemically modified lysine side chain, as no other nucleotide was reliably detected. Also the number of CSMs per crosslinking site remained rather low compared to either UV-crosslinking or chemical crosslinking with DEB or NM. This finding was confirmed by multiple independent experiments for each crosslinker or crosslinking method (data not shown).

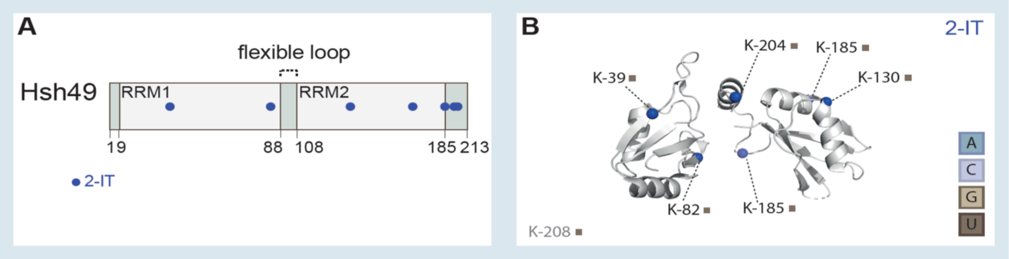


Figure 21: Structure of Hsh49 with mapped 2IT-mediated crosslinks to synthetic RNAs. A) Schematic depicting the positions of Hsh49 crosslinking sites within the Hsh49 protein from 2IT crosslinking. A sparse number of crosslinking sites were identified and mapped onto the structure, all of which were uracil containing nucleotides (**B**).

Having studied crosslinking of Hsh49 in combination with mass spectrometry extensively with synthetic poly-nucleotide RNAs comprising of a single different nucleotide for A, U, and C nucleotides, and poly-G_5U for a G nucleotide proxy, the benefits of chemical crosslinking started to become apparent. Identifying all four nucleotides unambiguously and reliably through the optimized workflow leads to the question of whether there also was a preference for a certain nucleotide for each chemical crosslinker as there is for canonical UV-crosslinking, exerting a uracil bias [3, 5, 46, 102, 103].

6.2.2 Preference in detecting nucleotides after (chemical) crosslinking

It was apparent that most crosslinks from chemical crosslinking experiments included either guanine or uracil nucleotides, raising the question of potential preferences in detecting certain nucleotides given the crosslinker used. To address these potential "crosslinking preferences" in detection, a strategy with dinucleotides comprising the poly-RNAs (CU, G_5U, GC, and GU) and the adapted workflow mentioned in section 3.10 was devised for model protein Hsh49. Upon manual validation of proposed hits from four experiments conducted in triplicates, crosslinker preferences in detection for either uracil or guanine were identified as shown in figure 22. 2-IT only cross-links Hsh49 to RNA in the presence of uracil-containing RNA at a lysine residue as mentioned above, precluding guanine from any detectable crosslinking event. Canonical UV-crosslinking was performed simultaneously to assess the preferences in detection of chemical crosslinkers against the "gold standard". About two-thirds of the identified cross-links contained uracil nucleobases, followed by guanine cross-links and very little cytosine cross-links as it was observed in these experiments. Contrary to the exclusive nature of 2IT, and the strongly preferring nature of UV-crosslinking for uracil nucleobases, DEB and NM behave indifferently towards it qualitatively speaking. In fact, guanosine cross-links were identified more often than uracil-containing crosslinks, indicating a preference in detection for guanosine over uracil in both DEB and NM. NM displayed a more homogeneous distribution between the three nucleotides G, C, and U. Cytosine, on the other hand, seemed to play a minor role in detectable crosslinking events, performing similarly in UV and DEB experiments, and slightly better in NM-mediated crosslinking. Moreover, the greatest number of validated hits (CSMs) were observed for DEB-cross-linked samples, followed by NM-mediated cross-links, and then UV-mediated cross-links (data not shown). A complete list of crosslinked peptides for all experiments can be found in Appendix C, section 13.3.

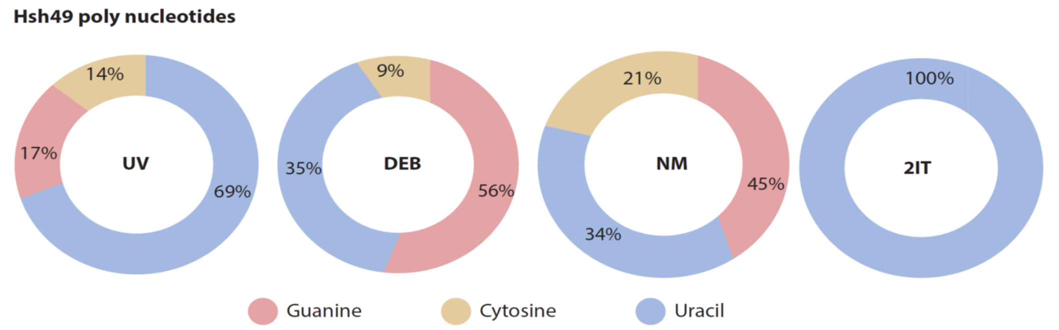


Figure 22: Detection preferences identified in controlled poly-nucleotide containing systems. Detection preferences identified in controlled systems. Observable preferences in detecting certain nucleotide adducts from manually curated CSMs of synthetic RNAs and Hsh49 were identified. UV-crosslinking is heavily uracil based with minor identifications for guanine and cytosine cross-links, but chemical crosslinker feature a broader range of crosslinked nucleotides. While DEB and NM demonstrate a greater preference in detection towards guanosine cross-links in comparable proportions, detection of cytosine is least prominent in any RNA containing either G or U.

Having reliably and unambiguously identified chemical crosslinks in DEB, NM, and 2IT experiments, the feasibility of using chemical crosslinkers as an alternative to conventional UV-crosslinking was proven successfully. Additionally, the known preference of uracil nucleotides for UV crosslinking could be confirmed with 2IT as no other nucleotide reacted with the activated thiol group detectably. This restriction could be circumvented by using DEB and NM that appeared to be favorably detected through guanosine nucleotides, at least in the restricted systems used. In the end, the restricted use of 2IT as an alternative crosslinking reagent to conventional UV-crosslinking lead to the decision to stop investigating 2IT as a potential chemical alternative since both methods identify uracil nucleobases. Reducing complexity as much as possible was initially necessary, but may not be tantamount to crosslinking protein-RNA complexes that bind their native RNAs. To address the issue of artificially forcing a preference that may not be present, Hsh49 was investigated under more native-like conditions by supplementing an RNA (two constructs of the U2 spliceosomal RNA(snRNA)) that has been shown to bind Hsh49 in the presence of an accessory protein (Cus1), inducing Hsh49 binding specificity to the U2 snRNA. Cus1 was purified by as instructed by Jana Schmitzova (Lührmann Lab) and the U2 snRNA was kindly provided by Olex Dybkov (Urlaub/Lührmann Lab, MPI bpc).

6.2.3 Hsh49:Cus1 U2 RNA

Analyses with the RNA-binding protein Hsh49 and non-specific RNA oligonucleotides above showed that chemical crosslinking extends insights gained from UV crosslinking, mainly by extending crosslinking nucleotides to all nucleotides. Mapping the crosslinking amino acids to the crystal structure published by Van Roon *et al.* shows that crosslinking events occur in both RRM domains when Hsh49 is crosslinked alone with synthetic RNA [144]. Wanting to investigate whether chemical crosslinking is comparable to UV crosslinking regarding its specificity of RNA interacting crosslink-able amino acids in RNA binding proteins in more native-like conditions, Hsh49 was complexed with the accessory protein Cus1. Synthetic poly-nucleotide RNAs may only bind randomly through electrostatic interactions promoted by proteinaceous surfaces and RNA structures that can accomodate RNA-binding surfaces on the protein. Conversely, a native RNA sequence folds into the proper secondary/tertiary structure that is most optimally suited to binding an RNA-binding protein through eons of biological evolution. As such, certain nucleobases may be prominently exposed and mediate RNA-binding to the protein and such interactions are of elevated interest to structural biologists studying protein-RNA complexes. Cus 1 and Hsh40 form a complex as part of the yeast spliceosomal U2 complex in which Cus1 promotes RNA sequence specificity in binding sepcifically in RRM1 of Hsh49, a crucial step in mRNA splicing [144]. Therefore, Hsh49/Cus1 protein complex bound to U2 snRNA shortened to 47 or 90 nt previously investigated by Van Roon *et. al.* was crosslinked chemically and using UV-irradiation [144]. Figure 24 depicts the structure of the U2 RNA, forming three major stem loops, the first of which (5' stem loop) is extremely important in Hsh49:Cus1 binding in RRM1 [144]. The full length RNA contains all three stem loops, while the shortened U2_47 RNA only contains the first 5' stem loop. To maximize protein-RNA interfaces, however, both the minimally required U2_47 and U2_90 were used in separate crosslinking experiments.

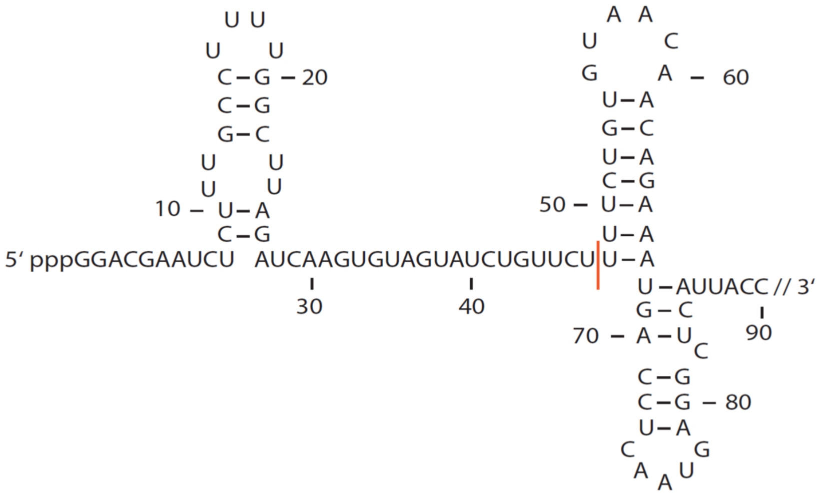


Figure 23: Structure of the U2 RNAs used in crosslinking Hsh49:Cus1. The full-length U2 RNA used in the Hsh49:Cus1 binding experiments contains 90 NTs and all three stem loops. The stem loop at the 5' end of the RNA was described to be extremely important in directing specific binding of U2 RNA to RRM1 of Hsh49, constituting the base U2 RNA used in the shortened U2_47 RNA [144]. To maximize protein-RNA interfaces, the full length U2_90 was also used.

Analysis of the crosslinked peptides showed that although Hsh49:Cus1 induces more identified crosslinking sites in RRM1, the identified crosslinking site in RRM2 did not vanish. Figure 24 depicts identified crosslinking sites in the crystal structures of the Hsh49:Cus1 complex when bound to either U2_47 or U2_90 RNA from three independent experiments. Akin to the isolated Hsh49 complex complexed with synthetic poly-nucleotides, mainly nucleophilic amino acids were identified to be crosslinked to uracil and guanosine-containing nucleotides. Also, both RRMs were identified to be involved in contacting the RNA, since both RRMs were found to be crosslinked, but a shift in the frequency of identified crosslinking sites in RRM1 was indeed observed in replicate experiments (n=5). Nonetheless, crosslinking sites in RRM2 did not abate. UV-induced crosslinking identified crosslinking sites in both RRMs but had a slight preference for RRM1 (55% of crosslinking sites) since most of the crosslinks in RRM2 were C-terminal near the end of the domain or even outside of it. This finding was confirmed for both types of RNA. Additionally, a few crosslinking sites were also identified in Cus1 that confers sequence selectivity and differences in the regions are minuscule between the two RNAs. DEB-mediated crosslinking yielded the largest number of crosslinking sites, most of which fell into the RRM2 domain. A few regions in RRM1 were also identified and coincided nicely with NM-mediated crosslinking sites, but fell short of the numbers provided by UV-crosslinking. Even so, all major nucleophilic amino acids in RRM2 that would normally be functionally binding RNA perfectly align in the center, almost visualizing a path the U2_90 RNA might be taking when binding. Lastly, NM crosslinking bolstered findings from DEB crosslinking, as the crosslinking sites matched often, but generally speaking did not provide additional information about the protein-RNA interface.

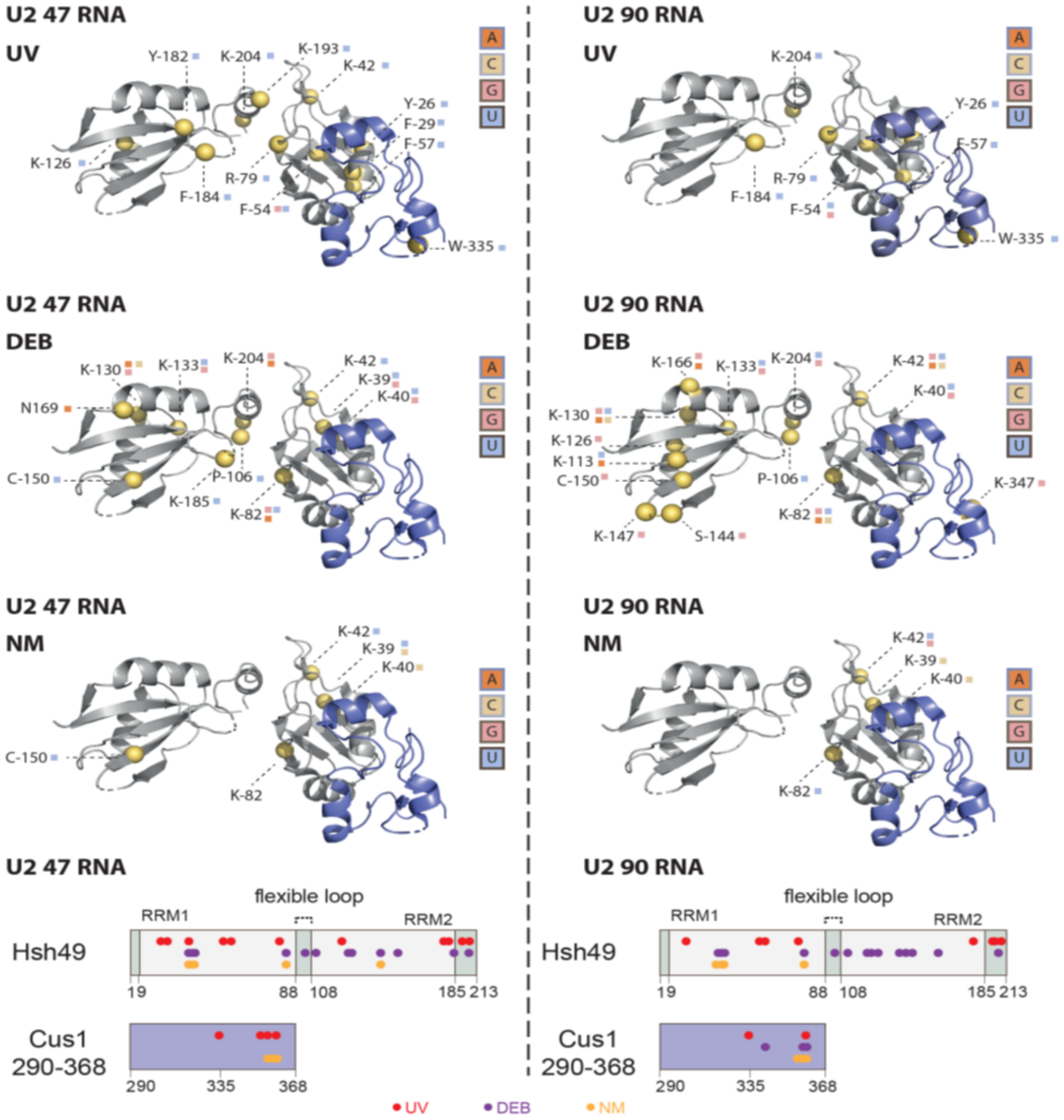


Figure 24: Crosslinked amino acids of both Hsh49 and Cus1 complexed with U2 RNAs.. Hsh49 is shown in grey, and Cus1 is depicted in blue. Crosslinking amino acids from U2 RNAs of length 1-47 bp and 1-90 bp are illustrated as yellow spheres. Notice the great variety of detected nucleotides in the case of DEB-crosslinked samples compared to the limited nature of UV cross-links to uracil nucleobases. NM crosslinks generally coincide with DEB-crosslinks. There is a slight increase in the frequency in identified crosslinking sites within RRM1, but sites in RRM2 did not abate upon Cus1 binding and conferring sequence specificity to the native U2 RNAs.

Estimating amino acid specificity for an individual nucleotide may be flawed when solely relying on an artificial system such as the one-protein-one defined RNA case of Hsh49. To investigate if a potential preference also exists on the amino acid residues depending on which crosslinking method was used, CSM data from the Hsh49:Cus1 complex bound to its native U2 RNAs were analyzed with respect to crosslinking amino acid. As shown in figure 25, UV crosslinking seemed to be returning CSMS of tyrosine residues most often given the specific conditions, followed by CSMs of phenylalanine, tryptophan, and lysine. Those three amino acids are also commonly identified in UV-mediated crosslinking [3, 5, 16, 46, 103]. CSMs from DEB-mediated crosslinking were found to be in stark contrast to the heterogeneity of UV-identified amino acid residues, as almost all CSMs were exclusively identified as lysine residues. Conversely, NM displayed a wide distribution of CSMs to histidine, lysine, and proline. Overall, CSMs of chemically cross-linked samples reveal nucleophilic amino acids to constitute the majority of crosslinked amino acids.

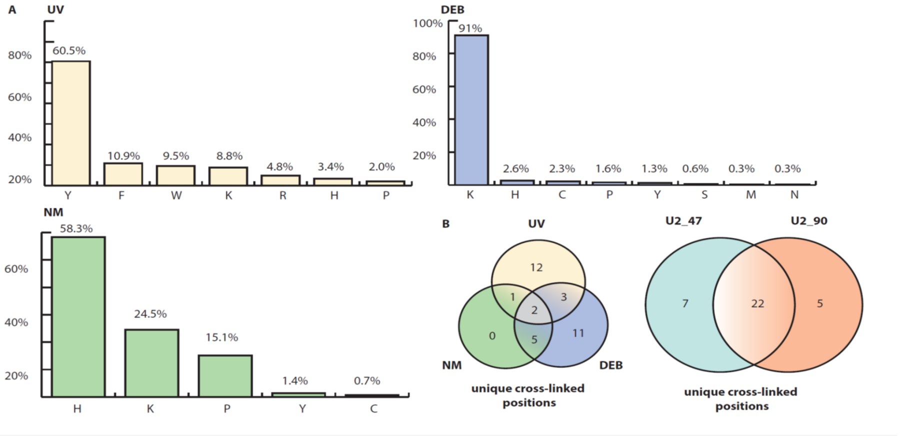


Figure 25: Different amino acid preference determined by CSM evaluation for different crosslinker used. Crosslinked amino acids from CSM data resulting in various Hsh49:Cus1 crosslinking positions. Unambiguous CSMs identifying a crosslinked amino acid from the Hsh49:Cus1 complexed with U2 RNAs data were extracted and analyzed. A) Different crosslinkers prefer slightly different amino acids for crosslinking. UV-crosslinking favors CSMs with tyrosine residues strongly, followed by phenylalanine, tryptophan, and lysine. DEB, conversely, is almost exclusively found in CSMs of lysine-mediated cross-links. NM features a broader amino acid distribution, ranking CSMs of crosslinked histidines and lysines as the favored residues. B) Complementary distribution of unique crosslinked positions. UV and DEB-identified cross-links are more unique to the individual crosslinker than shared, and NM shares almost all cross-links with DEB. C) Unique crosslinked positions identified in both U2 RNA-crosslinked experiments. Using the heavily truncated U2_47 RNA does not result in a vastly different number of unique cross-links compared to the longer U2_90 RNA. Overall, most crosslinked positions are found in both samples.

Scrutinizing the unique cross-linked positions from those CSM data revealed that only two crosslinking sites were shared between the different approaches (figure 25 B). One was identified to be lysine-42 that could be mapped onto the crystal structure, and the other one was found to be lysine-133 located in the flexible loop that could not be mapped. NM shares almost all of its crosslinked unique positions with DEB, corroborating the findings for chemical crosslinkers that appeared to be diametrical to UV-crosslinking. In fact, most of the unique crosslinking sites identified from UV experiments are unique to the crosslinking method, illustrating potential complement between the two approaches. Lastly, using either U2_47 or U2_90 RNA did not substantially change the identified unique crosslinking positions between the two RNA. The majority of crosslinking sites are shared between the RNAs as illustrated in the crystal structures above (figure 20) and in figure 25 B. Positions that are not shared most often fall within the same region, however, indicating few differences overall. A complete list of all identified crosslinked amino acid residues can be found in appendix C, section 13.3.

6.3 Dnmt2 crosslinking

To evaluate the chemical crosslinking approach in a more complex system, the methyltransferase Dnmt2 with its native tRNA G34RNA^{Asp} was analyzed and crosslinking locations were identified in collaboration with Sven Johansson from the Fitzner lab. While UV-crosslinking only returns a couple of crosslinking sites that led to a proposed conformation of the tRNA:protein complex, chemical crosslinking reliably detected RNA binding regions that were previously unknown [5]. Figure 26 shows that the majority of the crosslinking sites did fall within the nucleotide binding domains of Dnmt2, while a few chemical cross-links were found in the co-factor binding regions of Dnmt2, where S-adenosyl-L-homocysteine binds to donate a methyl group for the methylation of cytosine-38 of the tRNA^{Asp} [5]. The chemical crosslinkers again showed high selectivity in the regions they crosslink and two crosslinking sites coincide with all crosslinking methods, indicating that different modes of crosslinking still identify shared RNA-binding regions, albeit very small in this case. A complete list of all identified crosslinking sites can be found in Appendix C, section 13.3.

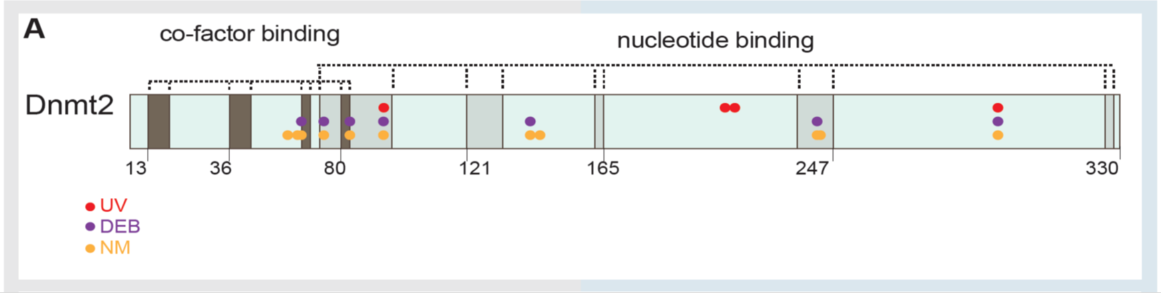


Figure 26: Schematic workflow of identified crosslinking sites in Dnmt2. The methyltransferase Dnmt2 features two main domains involved in co-factor binding (S-adenosyl-L-homocysteien (SAH)) and the nucleotide binding domain, binding the tRNA^{Asp}. Most crosslinking sites were found to be within the nucleotide binding domain with two of them coinciding perfectly between UV-mediated crosslinking and chemical crosslinking. Additionally, chemical crosslinking reveals additional sites N-terminal of the nucleotide binding site that overlap with the SAH binding site. DEB and NM generally locate the same complementary regions to regions identified by UV crosslinking.

6.3.1 Dnmt2 modeling

Following identifying crosslinking sites with both UV-crosslinking and chemical crosslinking, the suitability of the chemical protein-RNA crosslinking LC-MS approach was evaluated using data from Dnmt2 experiments. The methyltransferase Dnmt2 was analyzed in complex with its native tRNAs G34RNA^{Asp}. Recently, a study was conducted to identify amino acid residues in Dmnt2 that are in contact with the tRNA and from which a 3D structural model of the quaternary arrangement of both the components was generated [5]. In contrast to the initial UV-crosslinking studies that were performed, chemical crosslinking with DEB or NM revealed a multitude of crosslinked amino acids to tRNA as described above. The chemical crosslinked amino acids and their position in the 3D structure of Dmnt2 were then used to reconfigure the current model. ROSETTA-based ridged-body modeling of the tRNA structure in complex with Dmnt2 in collaboration with Piotr Neumann from the Ficner lab revealed a more favored quaternary arrangement supported by the chemical protein-RNA crosslinking sites in Dmnt2. These data demonstrate that chemical protein-RNA crosslinking is complementary to UV crosslinking and can aid in obtaining a better understanding when data points are sparse by adding more crosslinking sites in general and sites to different nucleobases than uracil.

Figure 27 shows combined cross-links from all experiments with Dnmt2 and its native RNA in two modeled conformations: In the left panel, model representations are based on atomic coordinates from Johansson *et al.* and are depicted as tRNA surface representations, while the Dnmt2 protein was rendered in ribbon representations [5]. The initial model was exclusively based on UV crosslinking data and suggest the tRNA handle to face downwards in this representation. Chemical crosslinking revealed additional sites indicated by gold and purple spheres in the right panel that were initially not considered in the model by Johansson *et al.*. Reanalyzing the model with all available cross-links leads to a new model with the tRNA flipped upwards at the handle that is also congruent with initial *in silico* rigid body modelling performed by Rosetta. In this new model, most crosslinking sites are in close vicinity to the tRNA and serve as a refinement of the old model.

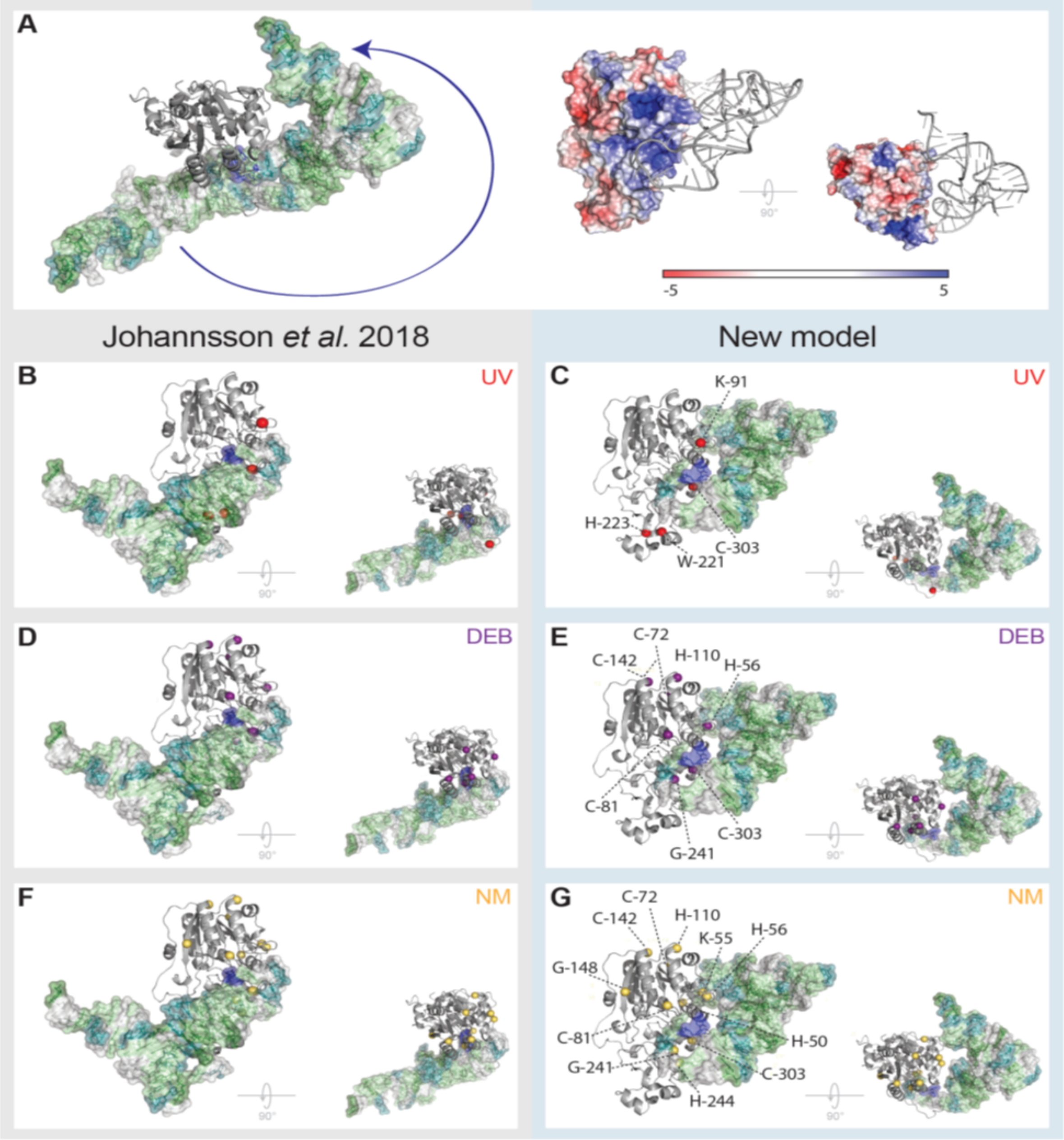


Figure 27: Combined cross-links from all experiments with Dnmt2 and its native RNA in two modeled conformations. Left panels) Model representations are based on coordinated from Johansson et al and are depicted as tRNA surface representations and protein ribbon representations. The initial model is based on UV cross-links and suggest the tRNA handle to face downwards in this representation. Chemical crosslinking reveals additional sites indicated by gold and purple spheres that were initially not factored in the Johansson model. Right panels) Reanalyzing the model with all available cross-links leads to a new model with the tRNA flipped upwards at the handle that is also congruent with initial *in silico* rigid body modelling performed by Rosetta. In this new model, most crosslinking sites are in very close vicinity to the tRNA. *Original figure adapted from master thesis "RNA-Protein crosslinking by 2-Iminothiolane and Nitrogen Mustard" by Luisa M. Welp.*

6.4 NELF complex protein-protein crosslinking

The NELF complex comprises of four proteins (NELF-A/B/D/E) that are interacting with each other. BS3 crosslinking coupled with mass spectrometry of the NELF complex bound to its target TAR RNA resulted in a surprising number of protein-protein crosslinks. RP-fractionated samples from the crosslinked NELF complex gave rise to 4038 CSMs that aggregated into 1668 unique crosslinking sites. This validates two important propositions upon which protein-RNA crosslinking is successfully based. First, all four subunits were present, detectable, and in direct contact with each other. Second, the prodigious number of protein-protein cross-links, especially stemming from NELF-E, assumed a highly flexible superstructure. That flexibility may stem from highly flexible tentacle regions of both NELF-E and NELF-A that are intrinsically disordered and are important in RNA binding as described below in section 5.3.14. Having confirmed a functional RNA-binding NELF complex that can be used reliably for protein-RNA crosslinking coupled with mass spectrometry, the NELF:Tar complex was crosslinked both chemically and by UV-irradiation. Crosslinking identified the protein-RNA interface.

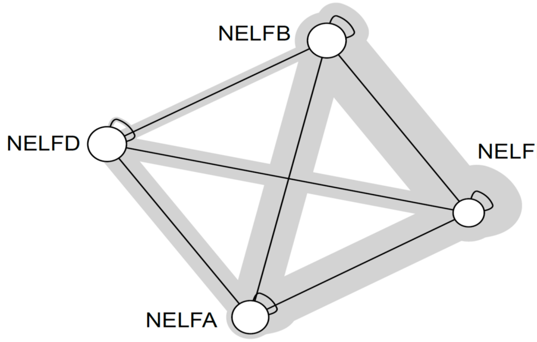


Figure 28: Protein-protein crosslinking of the NELF complex. BS3-crosslinking of the NELF complex shows extensive interactions between the different subunits with 4038 CSMs, collapsing to 1668 unique crosslinking sites. The thickness of the gray lines connecting the individual subunits correlate to the number of identified cross-links. Most cross-links originate from NELF-E that contains highly disordered regions in addition to the RRM domain at the end of the flexible tentacle region. Similarly, NELF-A contains a tentacle region that is highly disordered and flexible.

6.5 NELF complex nucleic acid crosslinking

As stated, the NELF complex comprises of four proteins (NELF-A/B/D/E) and its native HIV TAR RNA, forming a stable complex as depicted in figure 29 A. UV crosslinking returned a limited number of crosslinking sites, mainly associated with NELF-A, however, chemical crosslinking displays copious crosslinking sites across the complex. DEB crosslinking identified cross-links in the NELF-C protein and generally shares crosslinking regions with NM. Interestingly, there were many cross-links located in the NELF-A/E tentacle regions that could not be crystallized thus far. Moreover, the NELF-E RRM domain that could be crystallized was found to include a single crosslinking site of cysteine-300 identified through chemical crosslinking. Fluorescence anisotopy experiments with mutants of the NELF complex performed by Seychelle Vos from the Cramer lab showed significantly reduced TAR RNA binding upon deleting either NELF-A or NELF-E tentacle regions, and an even greater loss in binding affinity when both tentacle regions were deleted (figure 29 B).

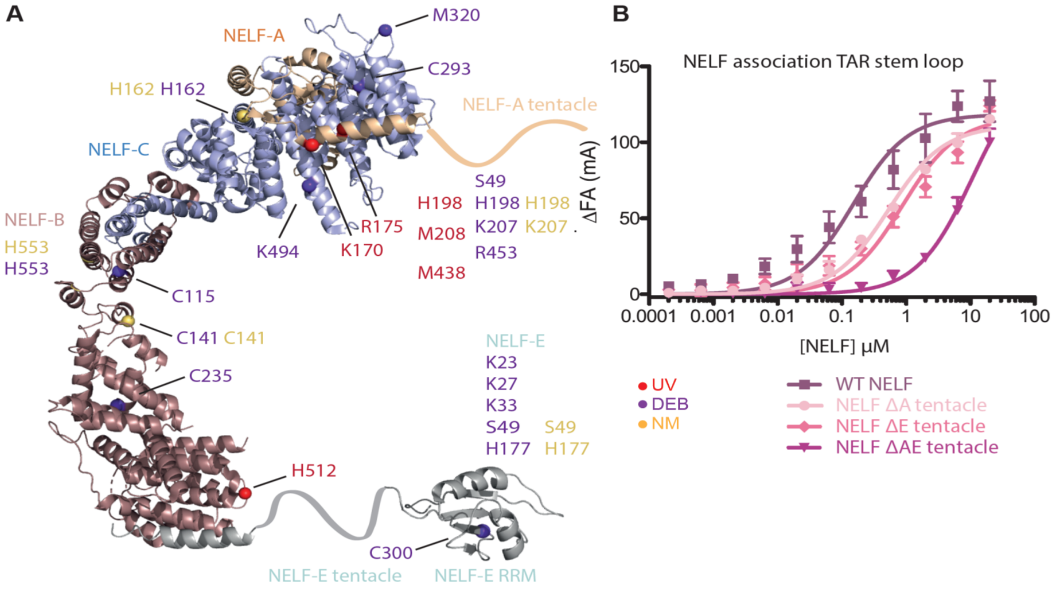


Figure 29: Identified crosslinking regions in the NELF complex are crucial to RNA binding. A) Combination of results from UV-crosslinking and chemical crosslinking with DEB and NM. UV crosslinking sites shown in red are least common and mainly restricted to NELF-A, whereas chemical crosslinking sites display a heterogeneous distribution across all members of the NELF complex, albeit focused on NELF-A and NELF-E. The vast majority of chemical cross-links are located in the tentacle regions of both NELF-A and NELF-E. Single-letter abbreviations of the crosslinked amino acid are given with their position in the sequence. Red=UV cross-link, purple = DEB, yellow = NM. B) Fluorescence anisotopy measurements with different NELF complexes, including deletion mutants of the tentacle region in NELF-A/E. Note how RNA binding is progressively impaired upon deleting tentacle regions.

6.6 Piloting chemical crosslinking of *E. coli* cells

To prove applicability of the chemical crosslinking approach and crosslinking site localization at amino acid resolution in an *in vivo* setting, *E. coli* BL21 DE3 cells were first treated with varying concentrations of DEB to assess at which concentrations visible crosslinking events would be detectable by SDS-PAGE. Concentrations of 1, 5, 10, 25, 50, 100, and 500 mM DEB were systematically analyzed by crosslinking at the indicated concentrations followed by SDS-PAGE analyses of whole cellular lysates. Figure 30 illustrates detectable crosslinking events indicated by a smearing of protein bands starting to emerge at 25 mM DEB, and 15'. Keeping incubation times short, but long enough for sufficient cytosolic and periplasmic crosslinking events to occur, a set concentration of 50 mM and 10 min incubation time was chosen.

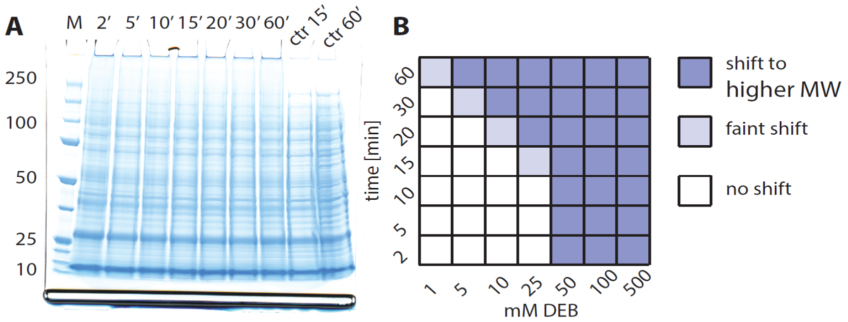


Figure 30: SDS-PAGE of *E. coli* BL21 DE3 cell lysates for various times of DEB treatment. **A)** Various time lengths of 50 mM DEB treatment, ranging from 2 min to 60 min. Compared to the last two control lanes of 15 min, and 60 min mock treatment, even a 2 min incubation with 50 mM DEB results in a shift to higher molecular weights in SDS-PAGE experiments that increases slightly over one hour with apparent blurring of the protein bands. **B)** Summary of multiple SDS-PAGE experiments with varying DEB concentrations. Crosslinking events causing band shifts and blurring start to appear at 15' for 25 mM DEB treatment, and are clearly visible at 50 mM.

Having found detectable crosslinking events at SDS-PAGE level for incubation times 15 min at 25 mM DEB and beyond, a set concentration of 50 mM DEB was chosen for *in vivo* crosslinking because the time frame was still short, but longer than the identified 2 min to ensure crosslinking reactions would occur in the cytosol at sufficient rates. Next, the viability of the bacterial cells needed to be evaluated to rule out the possibility that the treatment was too bactericidal and visible crosslinking at SDS-PAGE level only reflected proteins interactions of dead bacteria. A comprehensive growth kinetic was recorded to estimate the cellular damage occurring after DEB treatment that would be reflected in diminished growth rates. Figure 31 illustrates *E. coli* BL21 growth rates with and without DEB treatment over a 3h period. OD_{600} measurements roughly correlate with an increase in cell mass, as *E. coli* cells begin to propagate by binary fission until the end of the exponential phase is reached. DEB-treatment at starting point should slow down the growth rate compared to untreated cells. Luckily, it was found that 50 mM DEB is not bactericidal enough to kill off the cells upon exposure, but a bacteriostatic effect was observed over a 3h period. DEB-treated cells grew slower compared to untreated control samples, presumably due to cellular damage that needed to be repaired to resume exponential growth. The kinetic illustrates that DEB-treatment within the first 10 minutes does not cause the bacterial cells to die and thus constitutes a valid time frame for chemical crosslinking.

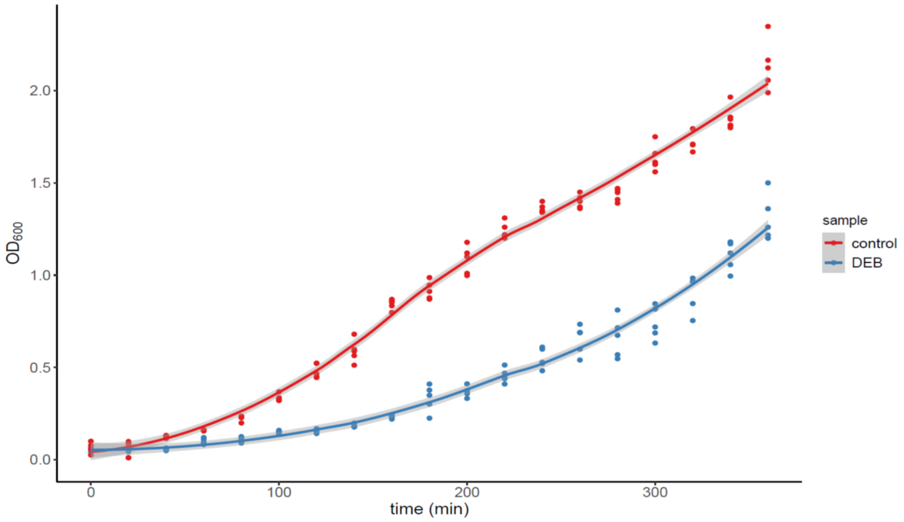


Figure 31: Growth kinetic of *E. coli* BL21 DE3 cells under DEB treatment. Bacterial cells were treated with 50 mM DEB and incubated over a 3h time frame, in which OD_{600} measurements estimated bacterial growth at various time points. DEB treated cells exhibit a lag in their growth compared to untreated control cells upon initial DEB treatment. This lag lasts approximately 90 min. in which cellular damage is presumably repaired before exponential growth can be resumed. Importantly, a 50 mM DEB treatment is not bactericidal enough to completely halt bacterial growth and can hence be used as a suitable concentration for chemical crosslinking.

Upon defining crosslinking conditions for *in vivo* studies, four data sets, corresponding to both S30 and S100 subcellular fractionated *E. coli* cellular lysates from DEB-treated or UV-irradiated bacterial cells were analyzed for CSMs using RNP_{xl}. RNP_{xl} output spectra (n=14,226) were manually evaluated in this pilot experiment, and combined to compare outcomes of DEB treatment vs. UV-irradiation. Figure 32 depicts the number of identified CSMs and corresponding crosslinked proteins. There is a large increase in CSMs stemming from the S30 UV-RNA sample that are comprised of multiple spectra for the same major cold shock proteins CspC and CspB that collapse into two proteins (A and B). Surprisingly, DEB-identified crosslinked proteins are comparable to UV-identified crosslinked proteins in terms of numbers. Comparing cross-linked proteins from each sample, both S30 and S100 subfractions share some proteins with their respective treatment group. UV-irradiated samples share 35 crosslinked proteins that were identified in both samples, whereas DEB-treated samples only share 22 cross-linked proteins across S30 and S100 fractions. Combining both fractions for each crosslinking class allows for direct comparison (figure 32 C)), where two distinct subsets with scant overlapping crosslinked proteins emerge. Given the total numbers of unique crosslinked proteins, it seems that DEB-treatment identifies other proteins that are not identified through UV-irradiation, making it an orthogonal approach to identifying protein-RNA crosslinked proteins. A complete list of all identified CSMs can be found in Appendix C, section 13.3.

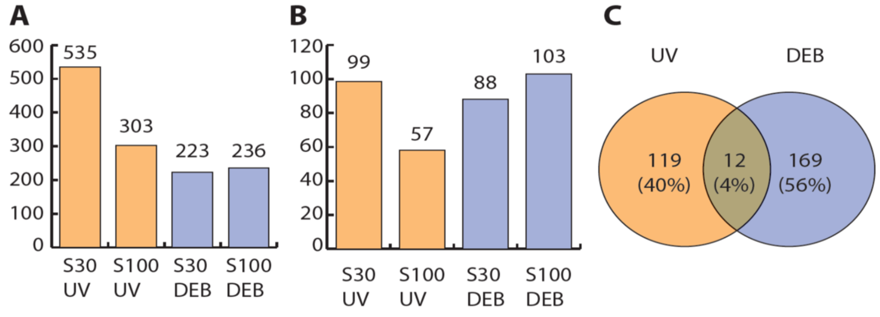


Figure 32: CSMs and crosslinked proteins of *E. coli* BL21 DE3 cells lysates from the pilot experiment. A) Number of manually validated CSMs from both UV-mediated and DEB-mediated crosslinking experiments. While DEB-identified CSMs do not vary considerably between samples, UV-identified CSMs vary considerably and outperform DEB-crosslinking in this pilot experiment. Differences are due to the fact that S100 samples should constitute a subset of S30 whole-cell lysates. **B)** Number of collapsed cross-linked proteins from true CSMs identified in A). Even though UV-crosslinking resulted in higher numbers of CSMs, the number of corresponding proteins is slightly lower than the number of DEB-identified crosslinked proteins. This is mainly the result of a few very prominent proteins that were identified with a higher number of CSMs (cold-shock proteins, CspC, and CspA). Again, a substantial difference between the samples was found in S30 and S100 samples in the case of UV, but not DEB. **C)** Venn Diagram of combined crosslinked proteins from both S30 and S100 samples that shared some proteins amongst them (35 proteins between S30 and S100 for UV-crosslinking, and 22 crosslinked proteins between S30 and S100 for DEB). The overlap between DEB-mediated crosslinked proteins and UV-crosslinked proteins is rather small, hinting at potentially complementary approaches.

6.7 30S ribosomal protein S7, 50S ribosomal protein L2, and NusB from the pilot experiment

To illustrate the validity of the chemical crosslinking approach using DEB as crosslinking reagent, three examples of the 300 cross-linked proteins from the pilot experiment using live *E. coli* BL21 cells were chosen as examples. Upon fractionation into an S30 and S100 subfraction, we identified 30S ribosomal protein S7 amongst the DEB crosslinked proteins in addition to many other DEB-crosslinked peptides. In the case of S7, the sequence GTAVKK was found to be crosslinked at the lysine residue K-136 to a cytidine nucleotide, as shown in figure 33 A). Additionally, we could identify 50S ribosomal protein L2 from the S30 subfraction to be crosslinked to a guanosine nucleotide via DEB. The peptide sequence HIGGGHK was found to be crosslinked at the histidine H-53, which is in direct contact with nearby RNA nucleotides (figure 33 B)). Both examples demonstrate that well-known RNA binding proteins are identified through chemical crosslinking. As a last example, we identified transcription antitermination protein NusB in our S100 fraction to be crosslinked via DEB to a guanosine nucleotide. The peptide SFGAEDSHK was found to be crosslinked at S-113 in a hyperflexible region of the protein, as depicted in in figure 33 C). Overall, these findings support the notion that DEB-mediated crosslinking identifies proteins fall within well established research on their biological function.

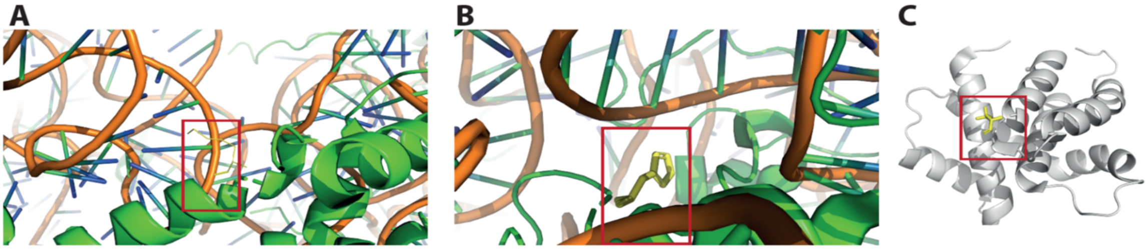


Figure 33: Example structures of chemically cross-linked proteins from *E. coli* BL21 DE3 cells lysates from the pilot experiment. Side chains of crosslinked amino acids are colored in yellow and boxed in red. The RNA backbone is colored in orange and the protein is colored in green (A and B) or gray (C). **A)** 30S ribosomal protein S7 was found to be chemically crosslinked to a cytosine nucleobase of 16S rRNA at position K-136. This lysine residue is in direct proximity to the RNA in the crystal structure validating the finding in a biological context. **B)** 50S ribosomal protein L2 was also reliably identified to be crosslinked to a guanosine nucleobase of 23S rRNA at position H-53 in proximity of nearby RNA nucleotides. Both RS7 and L2 proteins fit models based on crystal structures in the context of RNA binding in close proximity. PDB accession of the bacterial ribosome used in A) and B): [3J9Z] [190]. **C)** Transcription antitermination protein NusB was also unambiguously identified to be crosslinked to another guanosine nucleobase at a hyperflexible region of the protein (S-113). PDB accession: [3D3B] [191].

6.8 The problem with manual validation and big data sets

RNP^{xl} automates thousands of spectra that show peptide-RNA-heteroconjugates. It should be noted, however, that RNP^{xl} only returns possible CSMs. Owing to the ever-growing advances in technology, especially in the speed and resolution of mass spectrometers over the last few years, generated data files contain millions of spectra to sift through. A quick example illustrates the dimensions discussed here: one 120 min LC-MS OT run with a complex sample starts recording at minute zero until minute 120. The cycle time is set to 1.7 seconds, yielding about 4,235 MS1 spectra. From those survey scans, the top speed method allows maximal MS2 scans that can fit the cycle time of 1.7 seconds. Assuming the AGC target for every consecutive MS2 scan is reached in a tantamount fashion to conventional top20 methods used in data-dependent acquisition, there are 20 MS2 scans from one MS1 survey scan. Using this conservative assumption, 84,705 MS2 spectra are estimated from one complex run. In fact, eleven 120 min OT runs on complex *E. coli* cell lysates corroborate this crude exemplary calculation with a sample average of 86,973 MS2 spectra and a standard deviation of 4,520 spectra. These numbers amount to a 5.2% difference in about 68% of all sample means, indicating similar performance and numbers of MS2 spectra in one single file. Comparing DEB treated cells with UV-irradiated cells in a statistically relevant setting of n=3 for two fractionations amounts to 12 runs, or approximately 1.2x10^{6} spectra to manually validate. Expert evaluation reduces the time needed to judge a CSM to be a true negative (TN) to 0.5 s, but a true positive (TP) may need up to 15s. Estimating the number of TPs in an RNP$_{xl}$ output dataset of potential crosslinks to be 5% for complex data, 1.2x10^{6} entries amount to 60,000 TPs and 1,140,000 TNs. Estimating the time to fully manually curate the data would lead to 15,000 min spent on TPs and 9500 min spent on TNs. A combined 24,500 min, or 408 h, would require 51 8-hour days to go through a complete set.

While RNP$_{xl}$ does automatically annotate spectra to be possibly cross-linked, its inherent settings were based off crosslinking data obtained from UV-crosslinking, leading to a potential bias. A meta-analysis of four complex datasets with good spectral quality, including both DEB-crosslinked and UV-crosslinked *E. coli* samples, accumulated 14,226 manually curated spectra as either true positive (TP) or true negatives (TN). That metafile was used to find commonalities between the RNP$_{xl}$ output data describing those spectra as either TP or TN. As expected, a UV-bias was readily detected when considering the two most indicative metrics used in pre-filtering spectra for manual validation: the combined RNP$_{xl}$ score, summarizing different subscores, and the localization score that reflects how accurately the software could locate the crosslinking amino acid (if at all). Figure 35 illustrates the distribution of combined RNP$_{xl}$ scores for a UV-cross-linked sample of *E. coli* cell lysate. The scores from TPs fall within a narrow window of data points close to a score of 0, with some outliers scoring as high as 0.8, but their numbers are relatively few compared to the TPs centered around 0 since more than 75% of all TPs have a score lower than 0.15. On the other hand, TNs disperse their score more widely but more than 50% of the scores stemming from TNs fall between 0.4 and 0.8. Hence, the combined score separates the TNs accurately from the TPs if a cut-off is defined lower than 0.2 with minimal interference from low-scoring TNs. Some high scoring TPs are also lost, but losses are negligible in comparison to the large number of spectra that would need to be manually curated.

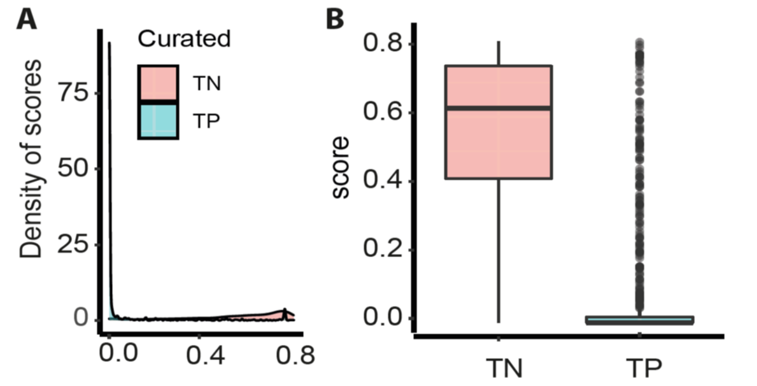


Figure 34: Boxplot and density plot of RNP$_{\text{xl}}$ score distributions for UV-crosslinking. A) Density plot of the total RNP$_{\text{xl}}$ scores for both TPs and TNs, falling into discrete regions. TP CSMs almost exclusively fall into a narrow region with low RNP$_{\text{xl}}$ scores, allowing them to be distinguished from TNs. **B)** Boxplot of scores from TN and TP CSMs, illustrating how scores fall into two discrete regions, where the TNs fall around a median of 0.62, and TPs cluster around a combined RNP$_{\text{xl}}$ score of 0. TNs spread 25% of their scores up to 0.42, causing minimal interference with TPs. Overall, the two distributions can be separated adequately by defining a cut-off value of 0.15 for CSMs from UV-crosslinking experiments.

Unfortunately, DEB-mediated crosslinking does not fair equally well compared to UV-crosslinking. The score distributions of the combined RNP$_{\text{xl}}$ scores from curated TPs (CSMs) are shown in figure 35 A. Both boxplots overlap in their distributions of more than 75% of the data, precluding any attempt to distinguish them by the combined score that worked reasonably well for UV-crosslinking. The density plot of both data distributions shows an almost even spread of DEB-associated TP scores across the range of scores from 0-0.6, allocating more than 50% of the TPs. Even scores past 0.8 are of little use to distinguish TNs from TPs, as TNs are drastically increasing simultaneously with TPs. This leads to the conclusion that the combined RNP$_{\text{xl}}$ score is not suited for separating DEB-associated TP CSMs, most likely due to the scoring settings being optimized for UV-crosslinking and not for DEB-mediated crosslinking. The only semi-valuable metric in distinguishing TPs from TNs is the matched ion current (MIC) shown in figure 35 B. Here, the density distribution of TPs seems to be bimodal, making it possible to separate some of them from the TNs if a strict cut-off of MIC score < 1.5 is enforced that removes the TNs, mainly distributed around low

MIC scores. The only problem with that is a dramatic loss in TP identification since 25% of them are also found in that region, leaving around 25% of TP CSMs to be correctly identified. Since only classifying a quarter of the data correctly is not good enough, other ways to include additional metrics were computed and tested.

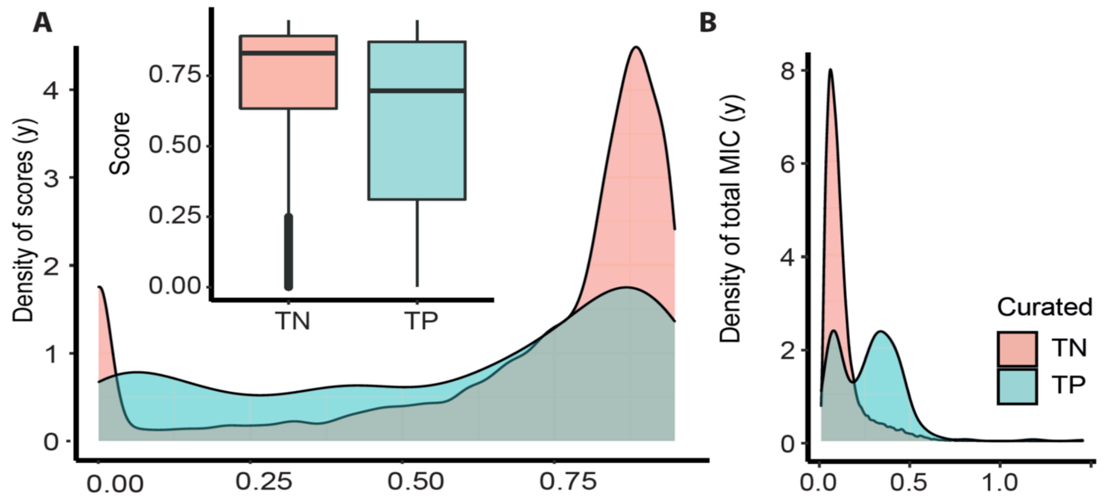


Figure 35: Boxplot of RNP$_{xl}$ score distributions for DEB-mediated crosslinking. A) Data distribution of combined RNP$_{xl}$ scores for both true positives (TP) and true negatives (TN) visualized in boxplots and desnisty estimations. TPs have a wider and more uniform spread than TNs, but occupy the same score range for more than 75% of the data. Overlapping significantly precludes the RNP$_{xl}$ score to be used as a distinguishing metric between TPs and TNs, albeit working well in the case of UV-crosslinking. **B)** Single most discerning metric of total matched ion current (MIC) score that has a bimodal distribution and separates TPs from TNs in values above an MIC score $>$ 1.5 with a considerable loss in TPs at lower scores (25% loss in TPs).

6.9 Additional metrics used to distinguish TPs CSMs from TNs

In addition to the RNP$_{\text{xl}}$ generated output data, additional metrics were computed as described in section 5.3.24. The combined 39 numeric features were then correlated against each other for TP CSMs. Unfortunately, cross-correlation against one numeric feature at a time did not yield new information that could be used to distinguish TPs from TNs. At best, it gave a weak correlation coefficient for knowledge that was already acquired through experience: a truly cross-linked peptide features spectra that have a highly intense precursor ion on MS1 (little interference during selection), a small ppm error for that precursor mass, a high MIC (most peaks are annotated to stem from the peptide and/or nucleotide adduct), a high localization score, a decent marker ion score (presence of nucleotide marker ions in spectrum), and a relatively short peptide length ($<$ 12 amino acids). A difference between a UV-induced CSM and a DEB-based CSM was not statistically significant other than a general suppression in all correlations for DEB, potentially hinting at the UV-bias in the settings and the heterogeneity of DEB-mediated crosslinks (data not shown). However, computers can quickly scan a plethora of different variables and detect underlying patterns much more reliably than manual inspection of heat maps and correlational matrices, so three classifying algorithms were chosen in an attempt to find underlying patterns in the dataset to distinguish TPs from TNs.

6.10 Building three classifiers based on the supervised machine learning algorithms k-NN, C5.0, and RIPPER

In preparation to set up machine learning algorithms, the numbers of TPs and TNs were harmonized not to mislead supervised machine learning approaches by splitting the complex *E. coli* cell lysate data into a training dataset and testing dataset in a 80%:20% ratio. TP CSMs from UV data encompassed 838 TPs and 838 TNs, totalling 1676 entries, whereas the DEB data set yielded 459 TPs and 459 TNs, totalling 918 entries. To maximize the number of TPs in each data set, TP CSMs from both S30 and S100 runs with exactly the same MS-method were combined. The k-NN lazy learning algorithm was chosen for its simplicity and robustness, even when faced with large datasets and multiple variables. Additionally, two rule-based algorithms were used to classify unknown data. The C5.0 decision-tree algorithm also performs well with large datasets and allows false positives (FP) to be penalized twice as much as a false negative (FN), potentially reducing the number of FPs. Lastly, the RIPPER algorithm also classifies data based on deduced rules and is easier to read than a branching C5.0 decision tree. This supervised machine learning approach is illustrated in figure 36. R-scripts for classifying RNP_{xl} entries from either UV-mediated or DEB-mediated crosslinking experiments in .Rmd format can be found on the accompanying CD.

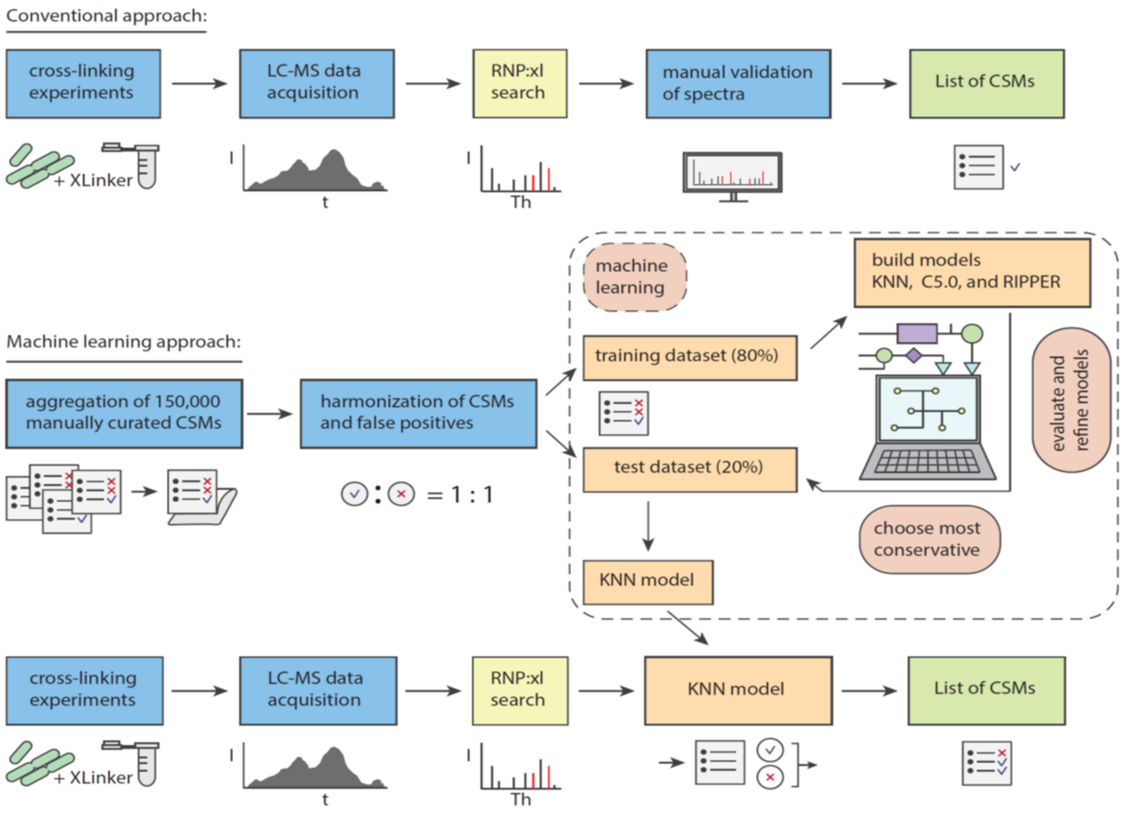

Conventional approach:
cross-linking experiments
LC-MS data acquisition
RNP:xl search
manual validation of spectra
List of CSMs
+ XLinker
I
t
I
Th
Machine learning approach:
aggregation of 150,000 manually curated CSMs
harmonization of CSMs and false positives
= 1 : 1
machine learning
build models KNN, C5.0, and RIPPER
training dataset (80%)
test dataset (20%)
evaluate and refine models
choose most conservative
KNN model
cross-linking experiments
LC-MS data acquisition
RNP:xl search
KNN model
List of CSMs
+ XLinker
I
t
I
Th

Figure 36: Supervised machine learning approach to classify spectra. The conventional approach to identifying protein-RNA crosslinking sites in protein-RNA complexes involves crosslinking experiments, LC-MS data acquisition and the laborious process of manually validating individual spectra to generate a list of CSMs. Supervised machine learning on a set of 150,000 spectra that were manually inspected to be TPs or TNs were harmonized to include equal numbers of TPs and TNs. This reduced dataset was split into a training data set (80% of the data) and testing data set (20% of the data) to build, train, and evaluate the three supervised machine learning algorithms k-NN, C5.0, and RIPPER. Upon validation, the k-NN model was chosen to classify spectra based on $\mathrm{RNP}_{\mathrm{xl}}$ output and generate a list of likely CSMs.

6.10.1 Comparing model performances and cross-validation against the testing data set

Having trained each model on the training data set, the models were evaluated using the test data set and 10 commonly used metrics. First, the rate of true positives (TPs), indicating where both prediction and actual outcome are the same, serves as a base measure. Similarly, the rate of true negatives (TNs) indicates how often a negative prediction was truly a negative CSM. Leading into the error rates, the rate of false positives (FPs), also known as “type I error”, points at the instances where a positive prediction was actually wrong. The type I error should be ideally low since it would reduce the number of falsely identified CSMs and thus reduce the error in consecutive meta-analyses that aggregate proteins and biological relevance from individual CSMs. The rate of false negatives (FNs) is also known as “type II error”, meaning the rate of predicting a CSM as false when it actually is true. In this instance, the type error is not important to the biological inference since losing misclassified entries is preferable to gaining fallacious ones that would skew interpretation. Nonetheless, auxiliary information that bolsters inferences that can be made from correctly classified data would potentially be lost.

Additional metrics can be used to evaluate and compare model performances. Sensitivity of a model describes how often a TP is correctly identified in regards to the numbers of TPs and FNs. Specificity on the other hand, relates the number of TNs to both TN and FP. The positive predictive value (PPV) indicates how often a TP was identified in regards to all positive predictions (TP and FP). Akin to that, the negative predictive value (NPV) describes how often TN was identified in regards to all negative predictions (TN and FN). Accuracy measures how often the classifier is correct in both positive and negative predictions. Lastly, the overall predictive value describes the mean of PPV and NPV. Table 15 summarizes those evaluating metrics for each model and crosslinking experiment. Overall, all three algorithms performed well in classifying unknown CSM data from the test data set. k-NN in particular performed with the highest overall predictive value (OPV) of 0.95 for UV classifications, and a decent OPV for DEB classification (OPV = 0.84). Both sensitivity and specificity were found to be 0.95 for UV-data, and 0.85/0.82 for DEB-data. It seems that k-NN misclassifies a greater number of FPs than FNs in DEB-related cases, alluding to potential problems with biological inferences from overestimated data that needed to be taken into account in the validation of the models.

The C5.0 algorithm is based on decision trees that can grow quite extensively as introduced in section 3.5.2. In the models, 10 trees were generated and compared for best performance. The trees contained 12-38 branches, classifying the data with an error of 5.1-11.9% (data shown in the classification reports on the accompanying CDs). Interestingly, the total $\mathrm{RNP_{xl}}$ score, the MIC of the adduct fragments, as well as the number of hydrophobic amino acids were most crucial in constructing the trees at early nodes for UV-based classifications, and the total MIC, marker ion scores, and number of charged amino acids for DEB-based classifications. These findings support earlier notions about potentially differentiating DEB-based CSMs by MIC scores (total and partial adduct fragments) mentioned above. Surprisingly, C5.0 performed exceedingly well on DEB data compared to k-NN or RIPPER, but suffered from twice as many FPs than k-NN, which should be avoided.

Lastly, the RIPPER algorithm performed well in general, but failed to reach the levels of k-NN in robustness or DEB-related sensitivity in C5.0. However, the RIPPER algorithm formulates easy-to-follow rules that coincides well with expert knowledge. Indeed, RIPPER formulates rules that align with previous observations: UV-mediated crosslinking events produce CSMs that have a total $\mathrm{RNP_{xl}}$ score < 0.07435 are most likely to be TPs. Conversely, a total $\mathrm{RNP_{xl}}$ score > 0.08391 is most likely to be a TN. DEB-mediated crosslinking displays different features that can be used to build a rule: an MIC adduct fragment score ($\mathrm{RNP_{xl}}$_pl_pc_MIC) > 0.1946, a retention time < 1658 s (the first 27 min of the gradient), and fewer than 4 hydrophobic amino acids per peptide. This makes perfect sense given a well annotated sequence has generally good scores as required and DEB-irradiation is known to induce crosslinks with hydrophilic/charged amino acids of hydrophilic peptides that elute early during the RP-C18 gradient.

Table 15: Summary of model performances. k-NN, C5.0, and RIPPER were used to classify CSMs from both UV-mediated and DEB-mediated crosslinking experiments as either true positive CSMs or true negative CSMs. Notice how all three models perform better for UV-experiments than for DEB experiments, as indicated by overall predictive values (OPVs) for UV-based classifications > 0.92, and OPVs for DEB-based classifications are < 0.93. All three algorithms displayed lower rates of TPs and higher FPs for DEB data, accumulating into a lowered sensitvity compared to UV-based classifications. The C5.0 algorithm is an exception, seemingly performing well with DEB data. Nonetheless, all three algorithms can accurately classify UV data, and perform decently with DEB-data.

	k-NN		C5.0		RIPPER	
Metric	UV	DEB	UV	DEB	UV	DEB
TP	0.94	0.78	0.96	0.94	0.92	0.83
FP	0.05	0.19	0.095	0.1	0.082	0.16
TN	0.95	0.9	0.91	0.86	0.92	0.84
FN	0.05	0.14	0.04	0.06	0.079	0.17
PPV	0.95	0.8	0.91	0.9	0.92	0.84
NPV	0.95	0.87	0.96	0.93	0.92	0.83
sensitivity	0.95	0.85	0.96	0.93	0.92	0.83
specificity	0.95	0.82	0.91	0.9	0.92	0.83
accuracy	0.95	0.84	0.93	0.92	0.92	0.84
OPV	0.95	0.84	0.94	0.92	0.92	0.84

6.10.2 Exploring supervised machine learning models on extended cell lysates from *E. coli*

To evaluate the suitability of the three algorithms, $\mathrm{RNP_{xl}}$ output data (potential CSMs) of a comprehensive 10-run data set were classified using the different algorithms. Figure 37 depicts the number of supposedly true CSMs (TPs) from the $\mathrm{RNP_{xl}}$ output data. Surprisingly, the RIPPER algorithm returned the highest number of potential CSMs in an exorbitant fashion. Scrutinizing those values, RIPPER suggests CSMs fall into a distribution with a min of 3573 CSMs, a median of 6828 CSMs, and a max of 10337

CSMs. Those numbers are exceedingly high and aggregate to an enormous number of crosslinked proteins that is unlikely to be true: min = 1756 proteins, median = 2728 proteins, and max = 3170 proteins. That would amount to 71% of the entire *E. coli* reference proteome, indicating a fallacious estimation. Similarly, the C5.0 algorithm probably overestimates the number of CSMs since its classification resulted in a median of 2876 CSMS with a min. of 1700 CSMs and a max. of 4219 CSMs. Collapsing the CSMs to crosslinked proteins resulted in 1728 crosslinked proteins (median) with min and max flanking the distribution with 924 and 2222 suppposedly crosslinked proteins, respectively. That median would correlate to 39% of the reference proteome to be crosslinked, which is lower than estimations made by RIPPER, but still unreasonable for an enriched sample. Lastly, the k-NN algorithm classified the same output data most conservatively with 453 CSMs as the median. Both min and max of 284 CSMs and 782 CSMs seemed reasonable and, although the max was found to be on the higher end of previously curated CSMs per data set, the numbers returned reflected manual curation closest. As depicted in fig. 37 B) k-NN-classified CSMs fall within a more narrow distribution with a median of 453, a min of 284, and a max of 782 CSMs. Those numbers reflect results from manual curation and seemed to be appropriate. The median number of potentially crosslinked proteins derived from those CSMs was found to be 261, and the min of 198 and a max of 408 potentially crosslinked proteins seemed far more reasonable than the outputs generated by C5.0 and RIPPER. Based on these findings, the k-NN model was chosen to classify additional data that were acquired through similar mass spectrometric analyses.

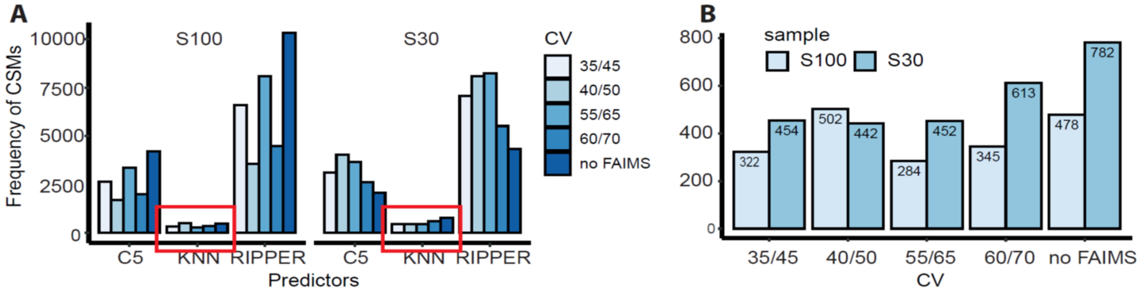


Figure 37: Comparison of k-NN CSM predictions. A) Number of classified true positive CSMs according to the three algorithms k-NN, C5.0, and RIPPER. The RIPPER algorithm yields the highest number of supposed true CSMs across all different samples, peaking at 10,337 supposedly true CSMs in one S100 sample, leading to questionable results that are far from any previously curated results. C5.0 also classifies an unrealistic number of CSMs to be true positives, albeit at a lower number than RIPPER. Lastly, k-NN performs relatively robustly with CSMs within normal ranges one would expect from manual curation. **B)** Classified true CSM entries from k-NN, fall within a narrow data range with median = 453 CSMs and reasonable variability for different experiments: range [322, 782].

6.11 Evaluating the k-NN model against the new NuXL software

The k-NN-model was used extensively to deal with a plethora of new *in vivo* data derived from crosslinking experiments with *E. coli*, *Bacillus subtilis*, and human HeLa cells to pre-select potential CSMs and reduce the amount of spectra to validate. This approach is imperfect since the cross-validation of the k-NN-model against the testing dataset returned an OPV of 0.95 for UV-classifications and 0.84 for DEB classifications. While those numbers do not allude to a failure of the algorithms, the prodigious amount of spectra to be classified at such predictive accuracy does pose a slight problem in regards to overestimating CSMs to be true positives when they are not. This type I error was estimated to be 5% for UV, and 15% for DEB-classifications which sum up to 500 false positives for UV and 1,500 false positives for DEB-classifications given a 10,000 CSM data set. Still, pre-selecting CSMs based on the k-NN model helped immensely until the new NuXL algorithm was developed by Dr. Timo Sachsenberg and Prof. Dr. Oliver Kohlbacher (Kohlbacher Lab, University of Thübingen). This new software was developed in close collaboration over the years as a successor to $\mathrm{RNP}_{\mathrm{xl}}$ and allows customization of crosslinking reagents and different nucleotides to be analyzed (manuscript in preparation). Most importantly, NuXL features an upper limit for the false discovery rate (FDR) that can be set quite strictly to 1%. Using the new NuXL software with similar settings in regards to ppm accuracy and nucleotide compositions as described in section 5.3.23, a strict 1% FDR was set and data were reanalyzed.

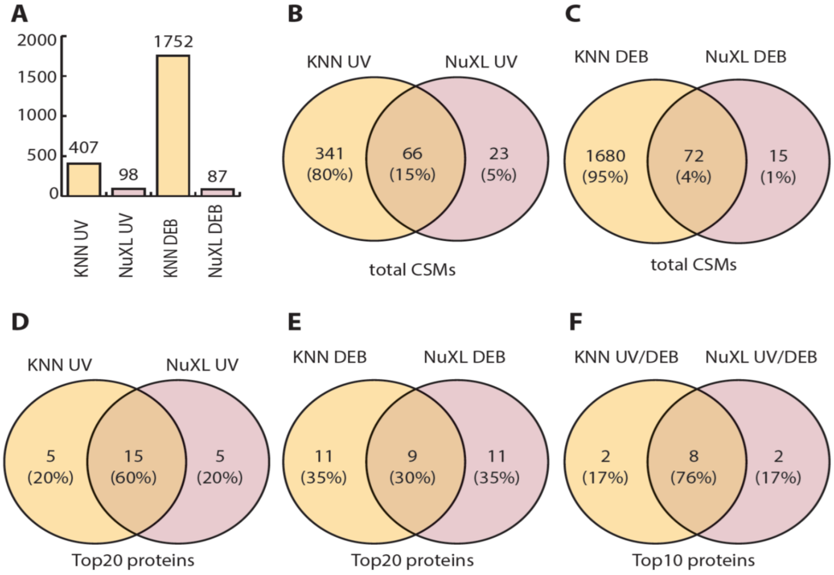


Figure 38: Comparison of k-NN CSM predictions with NuXL. A) Total numbers of classified CSMs by k-NN and NuXL. k-NN reports more true CSMs than NuXL for both UV-derived and DEB-derived samples. The ratio of CSMs for UV-samples equals 4.6, while the CSM ratio for DEB-samples equals 19.3, reflecting the proposed 5% error rate of k-NN classifying UV spectra, and an increased 20% error rate for DEB spectra. **B)** Direct comparison of UV classified CSMs. k-NN and NuXL overlap in 15% of total CSMs, equating to 74% of the NuXL data that is shared with k-NN. A large number of unique k-NN classifications skews the proportions. **C)** Direct comparison of DEB classified CSMs. k-NN and NuXL only overlap in 4% of total CSMs, equating to 83% of the NuXL data that is shared with k-NN. A large number of unique k-NN classifications skews the proportions dramatically, leading to the notion that k-NN may not be best for classifying DEB-derived data. **D)** Ranking supposedly crosslinked proteins by the frequency of CSMs per protein results in a direct comparison of "highly abundant crosslinked protein". The top 20 UV-crosslinked proteins are shared 60% of the time between k-NN and NuXL, with each contributing 20% of uniquely identified hits. **E)** Ranking potentially DEB-crosslinked proteins by the frequency of CSMs per protein generates a top 20 list of proteins that are shared 30% of the time between k-NN and NuXL, with each contributing 35% of uniquely identified hits. Again, DEB classifications are more diverse across the algorithms, probably due to a slight misfit in k-NN. **F)** Ascending the ladder of ranked proteins from both algorithms, 80% shared proteins can be obtained for either UV or DEB-classifications by inspecting the top 10 cross-linked proteins, validating top hits from k-NN to be most likely accurate.

Figure 38 shows a direct comparison between the k-NN model and the novel NuXL software, classifying CSMs from a set of 12 experiments (UV vs. DEB, S30 vs. S100, measured in triplicates) to be true hits. Figure 38 A) depicts total combined numbers of CSMs per algorithm. Here, k-NN repeatedly returns higher numbers than NuXL, especially in the case of DEB-treated samples. The CSM ratio of UV predictions k-NN vs. NuXL was calculated to be 4.6, perfectly reflecting the fact that NuXL operates at 1% FDR, while k-NN most likely scores around an error rate of 5%. However, the CSM ratio from DEB treated samples was found to be 19.3, suggesting that k-NN operates at an error rate approximately 20 times higher than NuXL, even though its type I error rate was estimated to be around 15%. Looking at the overlap between CSMs in figure 38 B) and C), k-NN classifies 80% and 95% of CSMs uniquely positively in the case of UV-treatments and DEB-treatments, respectively. Conversely, NuXL only reports 5% of CSMs to be uniquely positive in addition to a 15% overlap with k-NN's predictions in the case of UV-treated samples. An unrealistic 95% of CSMs were classified to be true hits based off k-NN, with a minor overlap with NuXL of 4% and a diminishing 1% uniquely NuXL-classified CSMs, indicating that k-NN fails to correctly identify reasonable amounts of CSMs for DEB-data. Ranking the supposedly crosslinked proteins by frequency of their corresponding CSMs allows for comparisons of the most likely crosslinked proteins based on frequency in each classification method instead. Here, proteins were ranked and the top 20 most abundant supposedly crosslinked proteins were compared (figure 38 D) and E)). For UV crosslinking, a substantial overlap of 60% was identified between the supposedly crosslinked proteins between k-NN and NuXL, only resorting to 20% of unique classifications to each algorithm (figure 38 D). When looking at the two sets, 75% of each set is shared, indicating reasonable confidence in results obtained for UV-crosslinked proteins based on computer classifications. DEB classifications differ dramatically once again, scattering across the data sets with a meager 30% overlap of the top 20 proteins and each algorithm uniquely identifying 35% of the top 20 proteins in each set. Restricting the top 20 list of proteins with the most CSMs corresponding to them to a top 10 list, however, increases the number of shared proteins to 76% (figure 38 E). When looking at the individual data sets, both k-NN and NuXL share 80% of their entries with each other, illustrating congruence among the proteins classified from abundantly identified CSMs. This finding was important as NuXL was only recently developed and k-NN was heavily used previously to identify crosslinking sites in most abundant proteins given their CSM counts. Top k-NN suggestions were only considered with at least 5 CSMs of more and manually curated afterwards, ensuring quality of the results despite an algorithm that evidently features a higher error rate, but still served greatly as a pre-filtering step in data analysis and validation.

6.12 HCD E. coli targeted

Reliable and most importantly unambiguous identification of crosslinked amino acids by mass spectrometry is in part due to serendipity in the ion source. One needs to have sufficient ions that also fragment the peptide backbone preferably in such a way that either b- or y-ions cover the complete sequence to maximize chances of locating the crosslinked amino acid. Oftentimes, crosslinks cannot be unambiguously located to a single amino acid, so increasing the number of high spectra provides a way to counteract the odds of only partially or insufficiently fragment the peptide. A marker ion (MI)-targeted approach was devised to increase the number of true CSMs by re-triggering an MS2 event once a MI has been detected in a previous MS2 scan. That way, a potentially crosslinked peptide is chosen for fragmentation multiple times, yielding hopefully multiple CSMs that aid in locating the crosslinked amino acid. To test this approach, S30 whole cell lysates and S100 subfractioned cell lysates from *E. coli* were extensively studied using both UV-crosslinking and chemical crosslinking wiht DEB. A complete list of CSMs from NuXL output data can be found in appendix C, section 13.3.

Figure 39 A) illustrates that UV-derived total CSMs still exceed DEB-derived total CSMs, independently of the sample. Perusing the list of NuXL-identified CSMs, an overwhelming preference for certain peptides appears to hold true for UV-derived samples. The cold-shock proteins CspA, CspB, CspC, CspG, for example constitute 51% of all total CSMs for UV-crosslinked samples, whereas DEB-mediated crosslinking features a broadened distribution of almost equal weights for the most abundantly identified CSMs of the ribosomal proteins. Collapsing the CSMs to crosslinked proteins, as shown in B) and C) for S30 and S100 samples, reduces the vast CSM discrepancy between the crosslinking modes and shows comparable outcomes that only overlap in 14% on average if both S30 and S100 samples are combined (D)). Scrutinizing CSMs from both experimental sets, the notion that the targeted MI approach works well for both UV-crosslinked and DEB-crosslinked samples as demonstrated by two selected examples for each setup (figure 39 E)-H)).

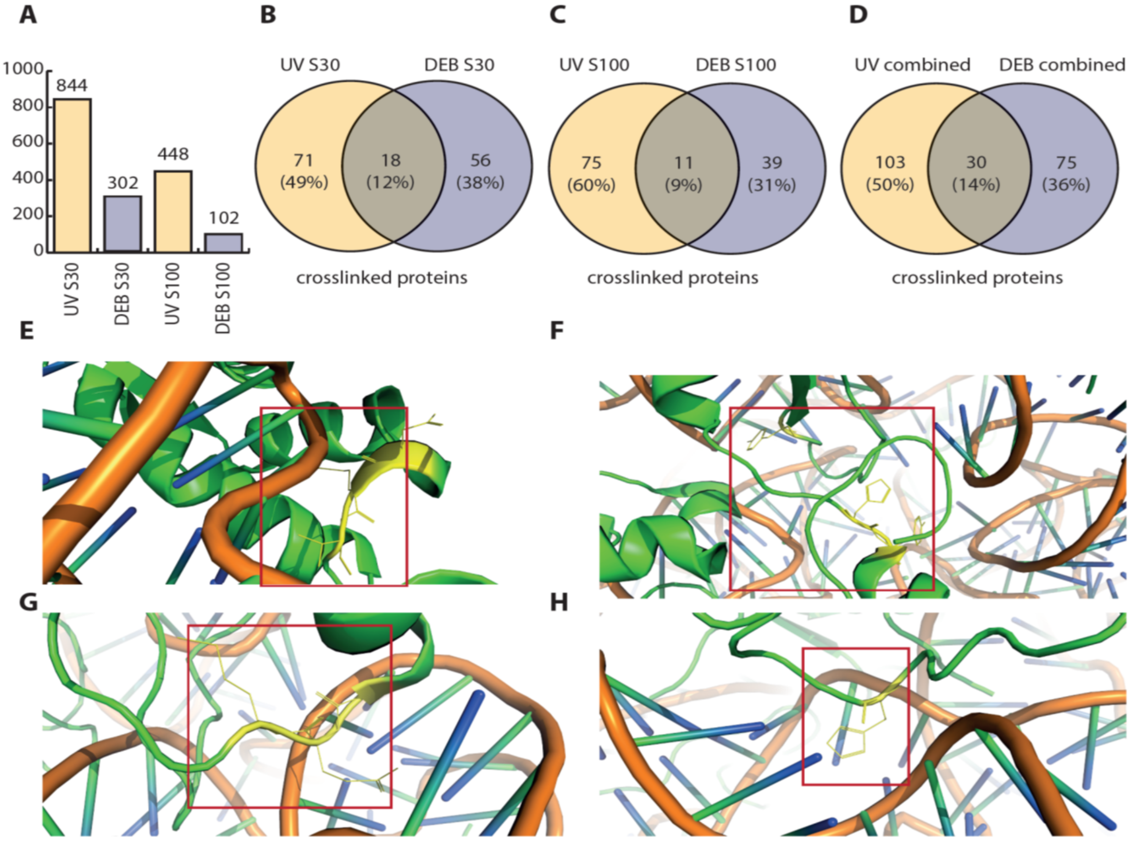


Figure 39: ***E. coli*** **crosslinks from the MI-targeted approach. A)** Total CSM counts from the MI-targeted approach. UV-derived CSMs exceed numbers of DEB-derived CSMs for both S30 and S100 samples as noted before, but DEB-crosslinked peptides can be identified reliably by the targeted MI approach. **B-C)** Collapsed crosslinked proteins from the S30/S100 samples, reiterating the complementarity of the two approaches in identifying different crosslinked proteins. Again, an overlap of about 10% could be identified between the samples. **D)** Combined S30 and S100 samples for both UV-derived crosslinked proteins and DEB-derived crosslinked proteins. In total, UV-crosslinking seemed to perfom slighly better than DEB-mediated crosslinking in terms of total numbers, but complementary information from DEB-crosslinking provides valuable insights. **E)** Three UV-crosslinked amino acids from 50S ribosomal protein L11 that are in close proximity to the adjacent RNA and only found in UV-irradiated samples. The peptide AADMTGADIEAMTR was found to be crosslinked at the three neighboring DMT residues (116-118) colored in yellow. **F)** Three UV-crosslinked amino acids from 30S ribosomal protein S7 that are also in close proximity to the adjacent RNA. The crosslinks are steming from two different peptides, only found in UV-irradiated samples: The peptide AFAHYR was found to be crosslinked at the two neighboring H and Y residues (153-154) colored in yellow and close in front. Additionaly, the peptide VGGSTYQVPVEVR was found to be crosslinked at the tyrosine residue 85 that can bee seen in the back. **G)** Three DEB-crosslinked amino acids from 50S ribosomal protein L1 that are in close proximity to rRNA. The peptide GEMNFDVVIASPDAMR was found to be crosslinked at the neighboring EMN residues (107-118) colored in yellow. Crosslinked nucleotides were identified to be mostly G, but reliable U adducts could also be found. **H)** Histidine crosslink of the peptide VPLHTLR from 30S ribosomal protein S3 at position 176. The corresponding G nucleotide is in close proximity and was only identified through DEB-crosslinking.

Some crosslinking sites were exclusively identified through UV-treatment, while others were only identified by DEB-mediated crosslinking. The 50S ribosomal protein L11, for instance, was identified to be crosslinked by UV-irradition in three specfic sites that are in close proximity to the adjacent RNA. The peptide AADMTGADIEAMTR shown in figure 39 E) was found to be crosslinked at the three neighboring D, M, and T residues (116-118) to an uracil nucleotide of the rRNA. Similarly, three UV-crosslinked amino acids from 30S ribosomal protein S7 were also identified and are in close proximity to the adjacent rRNA ash shown in F). Here, crosslinks are steming from two different peptides: The peptide AFAHYR was found to be crosslinked at the two neighboring H and Y residues (153-154) close to the front. Moreover, the peptide VGGSTYQVPVEVR was found to be crosslinked to uracil at the tyrosine residue 85 that can bee seen further in the back of figure 39. Both examples were corroborated by multiple CSMs. In the case of DEB, three crosslinked amino acids from 50S ribosomal protein L1 were identified and are in close vicinity to the rRNA, depicted in figure 39. Here, the peptide GEMNFDVVIASPDAMR was found to be crosslinked at the neighboring E, M, and N residues (107-118) to most prominently G nucleobases, albeit some U crosslinks could also be validated. Lastly, the histidine crosslink of the peptide VPLHTLR from 30S ribosomal protein S3 at position 176 was also uniquely identified through DEB-crosslinking coupled with mass spectrometry to a corresponding G nucleobase. These four examples showcase the versatility in identifying different crosslinking sites in their biological context and fit well with current crystal structures of the ribosome and thus the applicability of a MI-centric targeted approach [190].

While those examples were manually validated and mapped onto the crystal structures, a more general approach to crosslinking site localization by NuXL was investigated. Surprisingly, 87% of CSMs from UV-derived peptide spectra are assigned a localization by NuXl, exceeded by 93% of DEB-CSMs being assigned a localization. While these numbers are optimistic, manual inspection of both cases lowered the actual probability of correctly localizing the crosslinked amino acid to about 80% for DEB-mediated crosslinking experiments and 85% for UV. These numbers are still extremely valuable and prove high utlity for a MI-centric targeted approach to crosslinking mass spectrometry. Also, comparing CSMs and aggregated proteins to a regular non-targeted LC/MS run such as the pilot experiments (see fig 38 A), a substantial increase by a factor of 13.2 more total CSMs for UV crosslinking, and a factor off 4.6 for DEB-identified CSMs was identified for this MI-centric approach in this preliminary study. Additionally, ribosomal proteins dominated both shared overlapping proteins between the two approaches, as well as overall functional protein class. In total, mass spectrometric evidence of crosslinked peptides from 50S ribosomale proteins RL1, 2, 3, 4, 5, 6, 9, 11, 14, 15, 20, 22, 25, 28, and 32 were identified through UV-crosslinking and chemical crosslinking using DEB. Similarly, the 30S ribosomal protein RS 1, 3, 5, 7, 8, 10, 11, 12, 13, 18, and 19 were found to be crosslinked to RNA using both combined approaches.

Other proteins such as components of the RNA polymerase, transcription factors or elongation factors were also identified and resurfaced in confirmatory experiments described below in detail. Fischer-Exact-Tests on over-representation in different categories such as biological processes, molecular function and local network clusters reveal a coherent view on different RNA-interacting proteins. In particular, regulation of transcription in response to stress signals was significantly enriched ($p < 0.001$) due to the presence of cold-shock proteins CspA/C/E. Similarly, translation initiation was found to be enriched at $p = 0.014$ because of initiation factor 1 and 3 (IF-1/3), as well as methionyl-tRNA-formyltransferase. Functionally, RNA binding was found to be an underlying property owing to the larger number of ribosome-associated proteins (28 out of 57 proteins), in particular the large 50S subunit for which there were 17 out of 32 proteins identified. This encompassing rRNA-binding property and being constituents of the ribosome was found to be highly enriched with a p-value of 1.1 x 10_{-19}. RNA transcription by DNA-dependent RNA polymerase was also enriched at a p-value of 0.028 due to rpoA, rpoB and rpoC being identified. Lastly, a subnetwork including the protein translocase subunit SecY was identified and highly enriched at a p-value of 1.3 x 10^{-4}. SecY associates with two proteins of the 50S ribosomal subunit (L14 and L15), in addition to proteins of the small subunit (S8 and S13) to dock the ribosome to the translocon [192]. A complete list of CSMs for all experiments can be found in Appendix C, section 13.3.

6.13 FAIMS of *in vivo* DEB-crosslinked *E. coli* cells

Having shown that a MI-centric targeted approach can help raising confidence in identifying a truly crosslinked peptide by increasing the number of supporting CSMs, one might also think about other orthogonal ways to increase the sensitivity of crosslinked peptide detection by mass spectrometry. There are multiple ways of subfractionating a sample, reducing complexity sequentially and therefore allowing for more time to detect a crosslinked peptide in subsequent mass spectrometric analyses. Here, only subcellular fractionation into S100 and S30 *E. coli* samples was performed, but one can also fractionate those subcellular fractions orthogonally using RP-C18 basic reversed-phase chromatography. Another way would be to separate ions in the mass spectrometer to allow for time resolved detection of different ion species. FAIMS does seperate ions based on their behavior in an oscilating electric field, leading to an attempt to seperate crosslinked peptides from other ions species. To identify optimal counter voltages (CVs) for FAIMS, two metrices were identified and analyzed: Marker ion intensities and the presence of linear peptides. Others have reported high CV settings of counter voltages of -60V and beyond to be most optimal to maximize the number of unique peptides, and CVs of -55V for protein-protein crosslinked samples, serving as a starting point to investigate CV settings compatible with protein-RNA crosslinking [35, 193]. A complete list of all CSMs from NuXL output data can be found in appendix C, section 13.3 for all FAIMS runs.

6.13.1 Testing different CVs to identify optimal marker ion intensities

The presence of a diagnostic marker ions such as as singly charged nucleobases in MS2 spectra found after CID fragmentation of a precursor molecule, ideally a peptide-RNA heterogconjugates, is a fairly good measure to estimate the likelyhood of an RNA-containing molecule such as RNA-crosslinked peptides, or pure RNA, to be present. If a dignostic marker ion is detected, it is possible that the peptide is indeed crosslinked to RNA, and if diagnostic marker ions are absent, chances of the peptide being crosslinked converge to zero. As such, detection of marker ions in MS2 are potentially indicative of a crosslinked peptide to be present and if FAIMS resolves crosslinked nucleotide-peptide conjugates by applying proper CVs, one would expect to find time-resolved elution of different species that yield a diagnostic marker ion. Unfortunately, the immonium ion of tyrosine shares a diagnostic ion that yields a very similar m/z value to MI A. MI A ($C_5H_6N_5$) yields a peak at m/z = 136.0623 and the immonium tyrosine ion ($C_8H_{10}NO$) yields a peak at m/z = 136.0762. OT mass spectrometers can resolve that differences in masses (102 ppm) if properly calibrated, but seperation of the two species by FAIMS would still be preferable to aid in accurate classification of spectra later. Hence, all FAIMS runs including both double CVs and tripple CVs were analyzed for the presence of diagnostic marker ions, most prominently MI A and G that usually yield highly intense peaks, and their respective XICs were generated and compared. The presence of prominent marker ions for either A or G nucleotides reflected truly DB-crosslinked peptides that eluted at these specific times and were unambigously identified as crosslinked heteroconjugates. Figure 40 depicts two different FAIMS double CV runs at CVs = -40/-50V and -60/-70 V over a 118 min gradient that is used for complex samples. A) illustrates how marker ions A and G can be time-resolved at CV = -40/-50 V with both MI A and G eluting over the first 35 min of the gradient, while the IM-Y starts to elute from the RP-C18 column post 40 min. Interestingly, a second peak elutes for species fragmenting into a potential MI A around 70 min of the gradient that coelutes with IM-Y. In B), however, all three ion species elute over the first 40 min of the gradient, potentially interfering with each other and thus eliminating CV = -60/-70 V as a prime candidate for FAIMS-based seperation of certain ion species. In fact, comparison of all FAIMS runs indicates that CV = -40/-50 V is best suited to seperate MI A from IM-Y either in a double run, or in a triple FAIMS run, constituting CVs -40/-50/-60. Another important ion species that does not yield information of crosslinked peptides are linear peptides itself that could potentially be avoided using FAIMS.

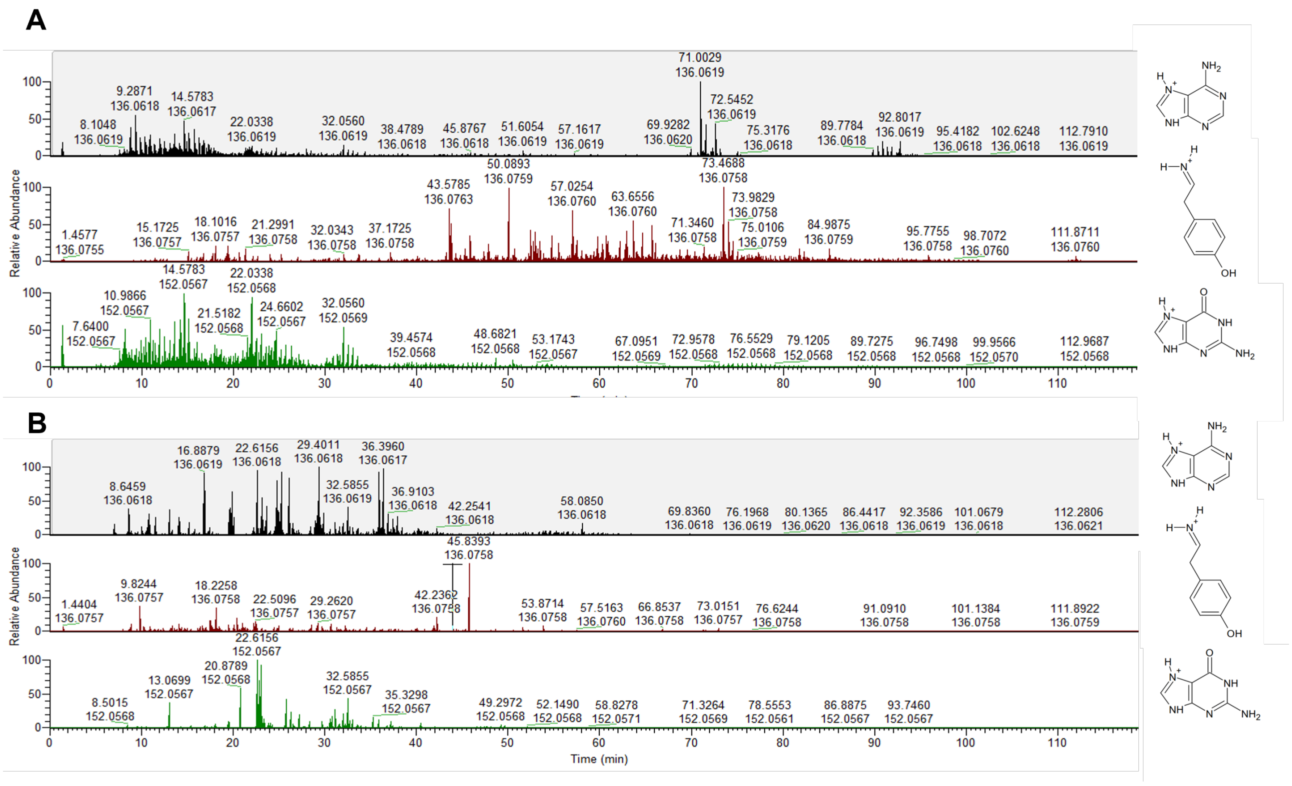


Figure 40: XICs of marker ions and immonium tyrosine at two different CVs. A) Extracted ion chromatograms (XICs) of the 118 min gradient used to analyze complex samples. At CV = -40/-50 V, adenine marker ions ($C_5H_6N_5$) at m/z = 136.0623 (top panel) elute mainly over the first 25 min of the 118 min gradient and around 75 min. The immonium tyrosine ion ($C_8H_{10}NO$) at m/z = 136.0762 elutes mainly after 40 min and is well seperated at CV = -40/-50 V, interfering little over the first 40 in with the MI A (middle panel). The MI G coelutes with MI A over the first 35 min and is also well sperated from the IM-Y ion (lower panel) **B)** XICs of the same RP-chromatographic method used in A). At CV = -60/-70V stronger MI A peaks appear over the first 40 min (top panel), but they coelute with the chromatographically inseparable IM-Y and MI G (middle panel, bottom panel).

6.13.2 Testing different CVs to identify settings least optimal for linear peptides

Since linear peptides are also of little interest and do not help in determining truly crosslinked peptides, certain FAIMS settings at which the greatest reduction in identifying linear peptides occurs, might also be a viable option to increase the chance of detecting crosslinked peptides. FAIMS was originally built to increase linear peptide identifications by increasing detection of differently modified linear peptides (PSMs). To do so, the most prevalent PTMs in a complex sample needed to be known. Identifying which PTMs were present in the complex bacterial lysates from all FAIMS runs, yielded a fairly small list of PTMs, illustrated in figure 41 A) with their respective relative frequencies. Methionine oxidation was found to be the most rampant peptide modification identified, ascending up to 35% of methionine-containing peptides to be oxidized. This finding illustrates the importance of knowing which PTM is found most prominently in a given sample because crosslinked peptides are equally likely to be modified on the peptide level as linear peptides are. Failing to include this PTM in the RNP_{xl} or NuXL search engine could potentially lead to a substantial loss in CSM identifications by 35%. Less frequently occurring peptide specificities included partially digested peptides, yieldig semitryptic peptides (15%), and the formation of pyroglutamate from glutamate or glutamine at the N-terminus (13%). Whilst modifications on peptides of 10% or less do influence outcomes, their effect is minor compared to a modification occurring 30% of the time. Additionally, including too many PTMs in the search for crosslinked peptides would increase the search space, as well as potential combinations that incidentally fit an obscurely modified peptide, resulting in their exclusion to be fed fed into RNP_{xl} or NuXl.

Having identified the major PTMs (methionine oxidation, semitryptic peptides, and pyroglutamate formation), all FAIMs runs from TiO_2-enriched samples were analyzed for remaining linear peptides that also bound to the affinity column. As depicted in figure 41 B), the reference proteome of *E. coli* BL21 (UniProt reference proteome *E. coli* BL21, accessed June, 2021) encompasses 4438 proteins from which 1505 proteins (34%) could still be identified from linear peptides. The expressed proteome is larger than the subset obtaining from enriched samples, but its exact number remains elusive as protein expression depends on environmental growth conditions and external stimuli, warranting expression of certain proteins. While bacterial growth was only temporarily impeded by DEB-treatment as explained above in section 6.6, minor proteomic changes are likely to occur during bacteriostasis that were eventually overcome after a lag phase as described above in section 6.6.

The identified proteins were aggregated from individual PSMs, stemming from different samples and at various CVs. Figure 41 C) and D) illustrate total numbers of PSMs obtained at different CVs from either double CV runs or triple CV runs. The total number of PSMs acquired at a CV = -40 V showed the highest difference regardless of other concurring CVs in either double or triple runs, indicating that more linear peptides were selected at CV = -40 V than any other CV. The number of total PSMs differed between double and triple runs, even though the CV was fixed because depending on what other concurring CV or CVs were applied, different ion species were selected for fragmentation. A CV of -50 V, on the other hand, always resulted in lower PSMs than -40 V. A CV of -70 V yielded surprisingly few linear peptides compared to a CV of -40 V in all double and triple FAIMS runs. Also, CVs -40 V and -50 V share the most linear peptides, indicating that those two CVs seem to be conducive to FAIMS separation of linear peptides. CV = -70 V was found to identify fewer PSMs on average across all different CVs, and also shared the least amount of PSMs with other CVs, indicating that FAIMS-separation at this CV leads to identification of more orthogonal ion species compared to other CVs. These data suggest that even though a CV of -40 V or -50 V separates diagnostic MIs from IM-Y well, they also lead to a selection of increased linear peptides that are fragmented and ID'd. Ultimately, direct comparison of CSMs in respect to different CVs can answer the question of optimal FAIMS settings for cross-link identification.

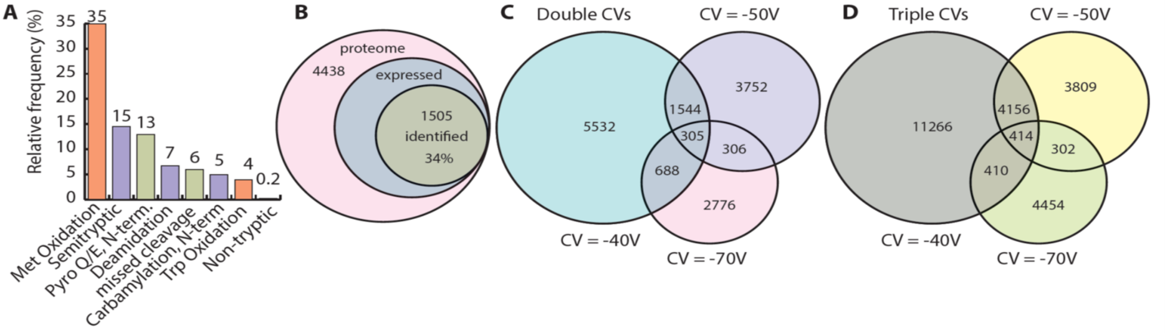


Figure 41: Different PTMs and modifications of bacterial linear peptides identified in *E. coli.* A) Frequency distribution of PTMs identified in all FAIMS runs using Byonic. Identification of linear peptides is enhanced when peptide modifications and specifications are known. In the case of *E. coli*, methionine Oxidation was identified to be the most prevalent PTM, identified in 35% of all methionine-containig peptides. **B)** Comparison of the identified proteins to the reference proteome. About 34% of proteins from the *E. coli* reference proteome, containing 4438 proteins could be identified by PSMs from linear peptides, although the sample was highly enriched using TiO_2. The expressed proteome encompases all different proteins expressed under the given growth conditions. The expressed proteome is thus smaller than the reference and larger than the sample, as it is highly dependent on growth conditions and potential stressors. **C)** Total PSM count of selected CVs from double CV runs. PSMs were highest for CV = -40 V alone and shared mostly with CV = -50 V, indicating that this combination lead to an increase in linear peptide identification. PSM counts refer to all PSMs obtained from different runs that included these specific CVs as one of the two CVs used in a double CV run. **D)** Total PSM count of selected CVs from triple CV runs. PSMs were again greatest for any triple FAIMS run containing CV = -40 V. Again, a large majority of the PSMs were shared between CV = -40 V and CV = -50 V, signifying that linear peptides are ineed most conducive to FAIMS sepeartion at those two CVs. Combinations including CV = -70 V identify PSMs that seem to be orthogonal to other CVs, only sharing a minor portion of PSMs with other CVs. PSM counts refer to all PSMs obtained from different runs that included these specific CVs as one of the three CVs used in a triple CV run.

6.13.3 Testing different CVs to identify the greatest yield of CSMs

As described above in section 6.13.1, there seems to be a differences in FAIMS settings that favor identification of one or two major species at various settings such as the separation of diagnostic marker ions at CVs = -40/-50 V from contaminating MI-Y ions, but also an increase in the detection of linear peptides at the same CVs. On the one hand, well separated marker ions might be beneficial, but one would rather focus on a decrease in linear peptides in hopes to identify crosslinked peptides. Lowest linear peptides were identified at higher CVs that fail to separate diagnostic marker ions from other ion species. Ultimately, only searching for crosslinked spectra indicative of protein-RNA heteroconjugates can answer questions regarding CSMs yields at different CVs. While initial analyses were conducted with the supervised machine learning approach described above, NuXL ultimately replaced the classifying supervised machine learning model because of its stringency and low FDR. While NuXL does return fewer potential hits, the overall ratios, as well as major examples described below did not change (data not shown). In any case, first baseline non-FAIMS run were analyzed for potential CSMs and collapsing proteins as described in figure 42. Here, UV-induced crosslinking resulted in a plethora of CSMs, especially in the case of S30 samples, outnumbering other types of samples by a factor of 10. Nonetheless, collapsed proteins from CSMs differed somewhat between the two sample types with numbers for UV-crosslinked proteins exceeding their DEB-counterparts in general. The huge increase in CSMs collapsed quickly due to hundreds of spectra referring to the same cold-shock proteins CspA/CspE/CspG. This finding bolstered earlier implications made from data acquired from a different mass spectrometer. Since FAIMS runs are tied to a specific setting on a specific mass spectrometer, re-establishing baseline identifications of non-FAIMS runs was indeed required to ensure comparability of the acquired results.

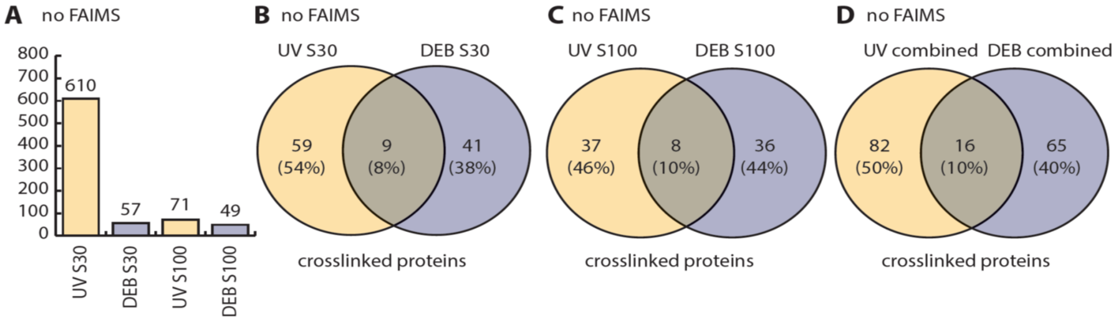


Figure 42: Baseline identification of cross-links from non-FAIMS runs. A) total CSM counts of different non-FAIMS measurements using complex crosslinked cellular lysates from *E. coli*. Similar to findings described above in section 6.6, UV-mediated crosslinking yielded the greatest number of CSMs. **B)-D)** Collapsed number of crosslinked proteins from both DEB-mediated crosslinking and UV-irradiation, confirming complementary identification of crosslinked proteins with an overlap of approximately 10%. The large number of UV-derived CSMs aggreavates into four proteins, namely the cold-shock proteins, resulting in comprarable numbers of crosslinked proteins for each approach. Again, overlapping proteins cluster into ribosomal proteins of both large and small subunit.

Since FAIMS can be applied with two CVs that alternate the electrical field in rappid succession or even with three CVs that lead to an even greater effect of ion transference through the quadrupoles, both double and triple CV runs were systematically investigated. First, double CV runs in combinations of CVs = -35/45 V, -40/-50 V, -55/-65V, and -60/-70 V were investigated. That way, the complete range of CVs from -35 V up to -60 V was covered in increments of 5 V, centered around a CV of -50 V. Figure 43 depicts the horizontal comparison of UV-crosslinking with chemical DEB-mediated crosslinking to systematically identify differences in FAIMS-separation that may be more conducive to one crosslinked moiety over the other at a given FAIMS setting. It is important to note that these FAIMS experiments were only conducted in single measurements to cover most FAIMS specific settings and thus constitutes preliminary data that generates a comprehensive overview. As depicted, double CV runs using CV = -35/-45 V yield substantially more CSMs for UV-derived samples than for DEB-derived samples, aggregating into a combined ratio of about 60:10:30% uniquely UV-identified, shared crosslinked proteins, and uniquely DEB-identified crosslinked proteins. This finding echoes the complementary aspect of chemical crosslinking in regards to UV-crosslinking, as it illustrates that about 30% crosslinked proteins can be identified through chemical crosslinking that would have been missed otherwise. While DEB-derived CSMs generally amount to half the number of crosslinked proteins than UV-identified CSMs, an interesting exception emerges at CVs = -40/-50 V, where DEB S30 samples, yields substantially more CSMs than its UV-counterpart (E)-G)). Looking at raw spectral data, errors in acquisition or sample composition leading to reduced signal intensities could partially be found to explain such drastic change in CSMs (data not shown). Congruent with NuXl findings, are reduced numbers classified by k-NN for the same sample, hinting at an error with this particular measurement and not necessarily erroneous classifications by software (Data not shown). Still, high quality data and numerous CSMs from the S30 sample mitigate the damage to such an extent that the aforementioned ratios still remain unchanged. Higher CVs of -55/-65 V or -60/-70V depicted in I)-L) and M)-P), respectively, restate the fact that UV-crosslinking generally results in more CSMs and consequently crosslinked proteins than chemically crosslinked samples, falling within a narrow window of data point for -55/-65V. The -60/-70 V double run, however, shows a sharp increase in UV-identifed CSMs that collapse into the known ration of 60:10:30%, primarily due to copious numbers of CSMs for cold-shock proteins Csps. Overall, these findings emphasize the complimentary aspect of chemical crosslinking to UV-crosslinking and applicability of double CV FAIMS runs to seperate different ion species at different CVs, favoring DEB-mediated and UV-crosslinked proteins at CV = -40/-50 in total numbers.

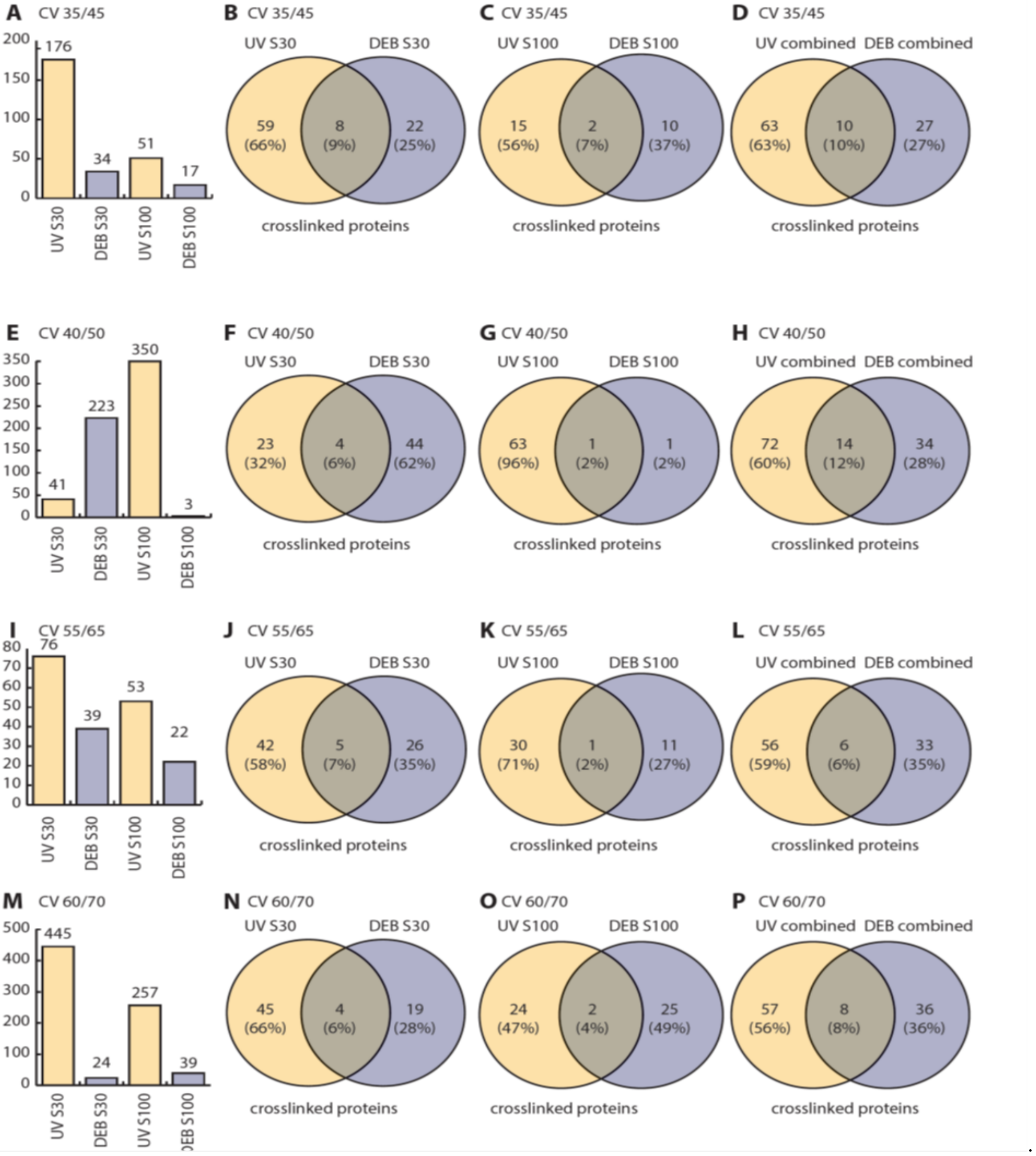


Figure 43: Comparison between UV-crosslinking and chemical crosslinking using double FAIMS runs. **A)** Total CSM counts of a double CV run, using CVs = -35/-45 V, resulted in a large population of UV-derived CSMs from the S30 sample and fewer CSMs for DEB-crosslinked samples. CSMs collapsed into proteins as illustrated in **B)-C)** for S30 and S100 samples, respectively. **D)** shows the overall crosslinked overlap of about 10% between the two approaches and about twice as many UV-derived crosslinked proteins than DEB-derived crosslinked proteins. **E)-H)** CV = -45/-50 V runs result in a different CSM distribution than expected. DEB S30 samples outperformed UV S30 samples in terms of CSMs and aggregated proteins, but DEB S100 samples grossly under-performed UV-derived CSMs as well as crosslinked proteins. Combined samples, however, hold true to the ratios of approximately 60:10:30 for uniquely UV-derived, shared, and uniquely DEB-derived crosslinked proteins. **I)-L)** Double CV run applying CV = -55/-65 V. Total CSM counts are higher for UV-crosslinked samples than chemically crosslinked samples using DEB. Collapsed crosslinked proteins also reflect this excess of identified UV-crosslinked proteins compared to DEB-mediated crosslinking for both S30 and S100 samples. The overlap of shared identified crosslinked proteins was also found to be small. Lastly, **M)-P)** CV = -60/-70 V runs depict a larger number of CSMs for UV-crosslinked samples that collapse into a small number of proteins. In particular, cold-shock proteins Csp-A/E were identified numerous times. The combined crosslinked proteins from both S30 and S100 samples reflect the 60:10:30% ration again, illustrating that while DEB-mediated crosslinking does not return the same number of CSMs, the inferred crosslinked proteins are mostly uniquely identifed via DEB, and not shared with UV-identifications.

Triple CV FAIMS runs add a third counter voltage to the oscillating electrical field, leading to a proposed filtering effect of greater magnitude. To test this, four triple CV FAIMS runs using CVs = -30/-50/-70 V, -35/-45/-55 V, -40/-50/-60 V, and -45/-55/-65 were sequentially tested using the same complex *E. coli* lysate. As depicted in figure 44, total CSM and crosslinked proteins counts are reduced compared to double CV runs, alluding to the enhanced filtering effect of using three counter voltages. Additionally, triple CV runs alter the 60:10:30% ratio considerably. While the ratio holds true for a -30/-50-70 CV run with copious CSMs for UV-derived samples (A)-D)), the ratio shifts to either UV-identified crosslinked proteins (E))-H)) or DEB-identified crosslinked proteins (I)-P)) at the expense of the shared overlap between the two approaches for triple runs -35/-45-55 V and -45/-55/-65 V. Interstingly, the -40/-50/-60 V triple FAIMS run seems to increase the shared crosslinked proteins by raising DEB-identified CSMs that also increase the total amount of identified crosslinked proteins via DEB to 45%. The double CV runs mentioned above also showed an increase in DEB-identified CSMs at CVs = -40/-50 V, leading to believe that those two CVs are indeed better suited to isolated DEB-crosslinked ions during FAIMS.

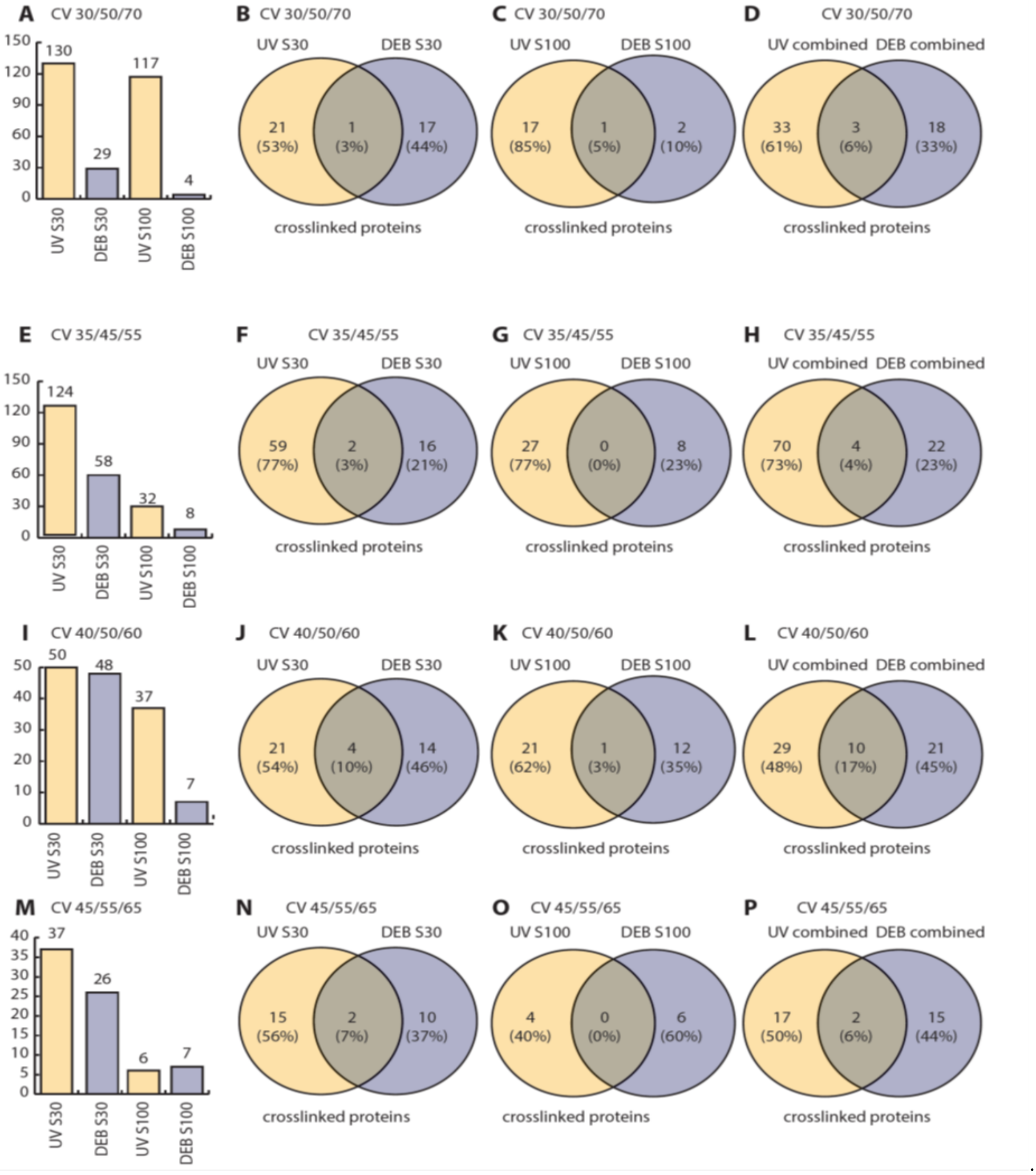


Figure 44: Comparison between UV-crosslinking and chemical crosslinking using triple FAIMS runs. A)-D) Triple CV run, using CVs = -30/-50/-70 V, shows a preference for UV-crosslinked peptides, as both CSMs and corresponding crosslinked proteins are greater in numbers than DEB-derived CSMs and corresponding crosslinked proteins which are extremely low in this run. The aforementioned 60:10:30% ratio holds still roughly true. **E)-H)** Triple CV run, using CVs = -35/-45/-55 V with decreasing numbers of CSMs from UV-crosslinked corresponding proteins in each S30 and S100 sample, correlating to a decrease in DEB-derived CSMs for both samples. Consequently, the overlap is smallest with only 4% of the crosslinked proteins being identifiable through both crosslinking approaches. **I)-L)** Triple CV run, using CVs = -40/-50/-60 V with almost equal numbers of CSMs and more UV-crosslinked corresponding proteins in each S30 and S100 sample. Additionally, the number of shared proteins is highest at these CVs. **M)-P)** At last, triple CV run, using CVs = -45/-55/-65 V yielding only few CSMs that aggregate into fewer crosslinked proteins in both S30 and S100 samples. Overall, UV-crosslinking yields slightly more CSMs/crosslinked proteins or does not differ in total number of crosslinked proteins, even though the overlap between the two approaches is fairly small.

Comparing the FAIMS and no FAIMS runs in regards to identified crosslinked proteins, independently of sample types allows for direct comparison between the different settings for the two crosslinking approaches in more broader terms. Figure 45 A) shows total CSM counts for each FAIMS run and the non-FAIMS measurements. Considering total CSMs counts alone, a double CV FAIMS run with CVs = -60/-70 V seems optimal, followed by a non-FAIMS run for UV-crosslinked samples. DEB-crosslinked samples, however, seem to be slightly better detected at a FAIMS run with two CVs, -40/-50 V. In B), the aggregated proteins from each measurement generally follow a trend that decreases with inceasing CVs for both UV-crosslinked and DEB-crosslinked samples. Still, non-FAIMS runs are outperforming FAIMS runs individually by the numbers of CSMs and resulting crosslinked proteins. This raises the question if FAIMS really aids in identifying protein-RNA heteroconjugates. If one would compare all eight FAIMS measurements to the non-FAIM run, as depicted in C), a resounding affirmation for FAIMS would be found. Both the numbers of UV-crosslinked proteins and chemically crosslinked proteins using DEB can be boosted by about 60% if FAIMS runs are included in the analyses of complex samples. Furthermore, the additional 60% are purely gained without losing other proteins to a different method, since FAIMS runs encompass the entire set of non-FAIMS proteins. While this direct comparison may be unfair in regard to measurements (one non-FAIMS run vs. eight different FAIMS runs), the likelihood of increasing the number of crosslinked proteins by 60% with replicate measurements of the same non-FAIMS method is highly unlikely. Therefore, one can conclude that FAIMS does indeed aid tremendously in identifying more UV-crosslinked and chemically crosslinked proteins if sample amounts are sufficient for multiple FAIMS runs.

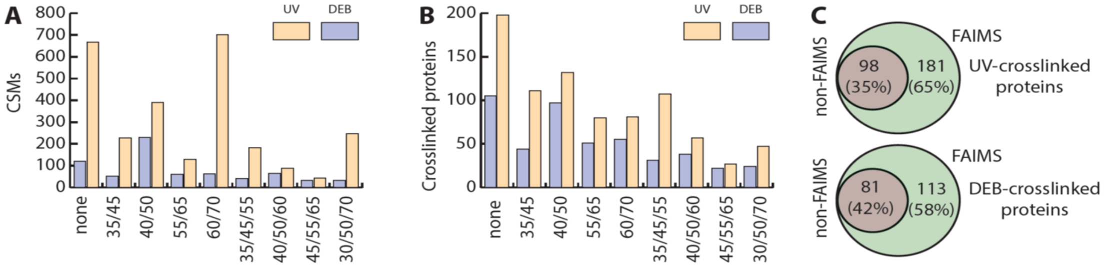


Figure 45: Vertical comparison of different CVs for identification of maximal CSMs and crosslinked proteins. A) total number of CSMs from all different FAIMS and non-FAIMS runs. Most CSMs could be identified through a double FAIMS run at CVs = -60/-70 V, namely 702 CSMs from UV-crosslinked samples (S30 and S100 combined), followed by the corresponding non-FAIMS runs (667 CSMs). The largest number of DEB-derived CSMs could be obtained from the double FAIMS run at CVs = -40/-50 V (391 CSMs), which also ranked third for UV-mediated CSMs (226 CSMs). **B)** Collapsed crosslinked proteins from CSMs show that even though UV-derived CSMs at CVs = -60/-70 V are bountiful, the corresponding protein numbers fall short to 81 crosslinked proteins. The largest number of identified crosslinked proteins could be obtained from a non-FAIMS run (198 crosslinked proteins). Moreover, most DEB-mediated CSMs also stem from the non-FAIMS run (105 crosslinked proteins), followed by 97 DEB-crosslinked proteins from a double FAIMS run at CVs = -40/-50 V. Increasing the counter voltages generally speaking also decreases the number of uniquely identified crosslinked proteins. **C)** Comparison of non-FAIMS runs and combined FAIMS runs demonstrates powerfully how additional FAIMS runs can increase the numbers of crosslinked proteins from 98 to 181 in the case of UV-crosslinking, and from 81 to 113 for chemical crosslinking using DEB. Since all proteins from non-FAIMS runs are included in the FAIMS superset, the information gain is purely additional without losing some proteins.

If sample amounts are limited, however, one would greatly benefit from knowing which FAIMS runs are most economical and return the highest value in terms of information gain. To identify that, FAIMS runs were compared among each other, both double and triple CV runs. Figure 46 depicts Venn diagrams of crosslinked proteins from each S30 and S100 sample in relation to their identifying double CV FAIMS runs. UV-crosslinked samples differ most dramatically in the two low CV runs -35/-45, favoring S30 whole cell lysates and a double run with CVs = -40/-50 favoring ribosome depleted S100 fractions (A)-B)). Combining both samples for UV-crosslinking demonstrates a double CV run with CVs = -40/-50 to be most comprehensive in covering both types of UV-crosslinked samples (C)). DEB-mediated crosslinking of S30 samples yields highest numbers for a double FAIMS run at CVs = -40/-50 V, whereas crosslinked proteins from S100 samples are best identified at CVs = -60/-70 V. Overall, though, the two double FAIMS runs fair comparably well when both samples are combined and given the fact that UV-crosslinked samples are also identified well at a double run with CVs = -40/-50 V, a single double CV FAIMS method can be used for both samples. These data support the notion that FAIMS settings can be used to reliably identify CSMs at different double CV settings, each of which seem to be orthogonal in nature as most proteins are uniquely identified for both UV-crosslinked samples, as well as DEB-mediated samples.

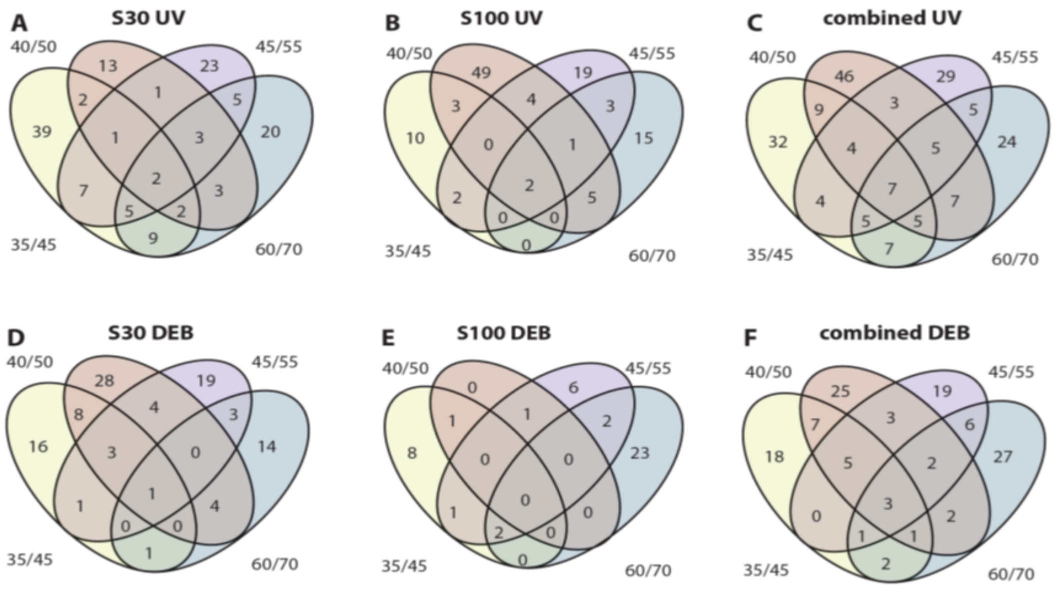


Figure 46: Comparison of different double FAIMS runs in regards to aggregated proteins. A)-C) All four double CV runs yield mainly unique identifications of UV-crosslinked proteins, probably due to isolation of different ion species at different CVs. Additionally, CV= -40/-50 V yields both the highest number of crosslinked proteins, as well as shares the most crosslinked proteins with other CVs, making it an excellent candidate for single FAIMS runs. **D)-F)** The double CV runs for DEB-mediated crosslinked samples, also yield mainly unique identifications. Promising double CV run at CV= -40/-50 V did yield the highest number of crosslinked protein compared to other CVS, in addition to covering most crosslinked proteins from other CVs, making it again an excellent candidate for single FAIMS runs.

Analogously to the double CV FAIMS runs described above, the triple FAIMS runs also have major commonalities and discrete distinctions in their ion-filtering effects. Figure 47 shows how each triple CV FAIMS run compares to each other considering the total number of crosslinked proteins that are returned by NuXL. Most importantly, triple FAIMS runs seem to identify fewer crosslinked proteins regardless of their settings. Lower CV settings including a triple run using CVs = -35/-45/-55 V seems to be optimal for UV-crosslinked samples (A)-C)), much as a triple CV FAIMS run with CVs = -40/-50/-60 V seems to be most suitable for DEB-crosslinked samples (D)-F)). The DEB S100 sample appeared to be rather void of identifiable crosslinked proteins in these triple runs experiments, as their overall yield was found to be surprisingly low and almost exclusive to each CV setting. However, the orthogonal behavior of amassing a set of rather unique proteins for each CV setting seems to be universal, as it was also found for double CV settings as described above.

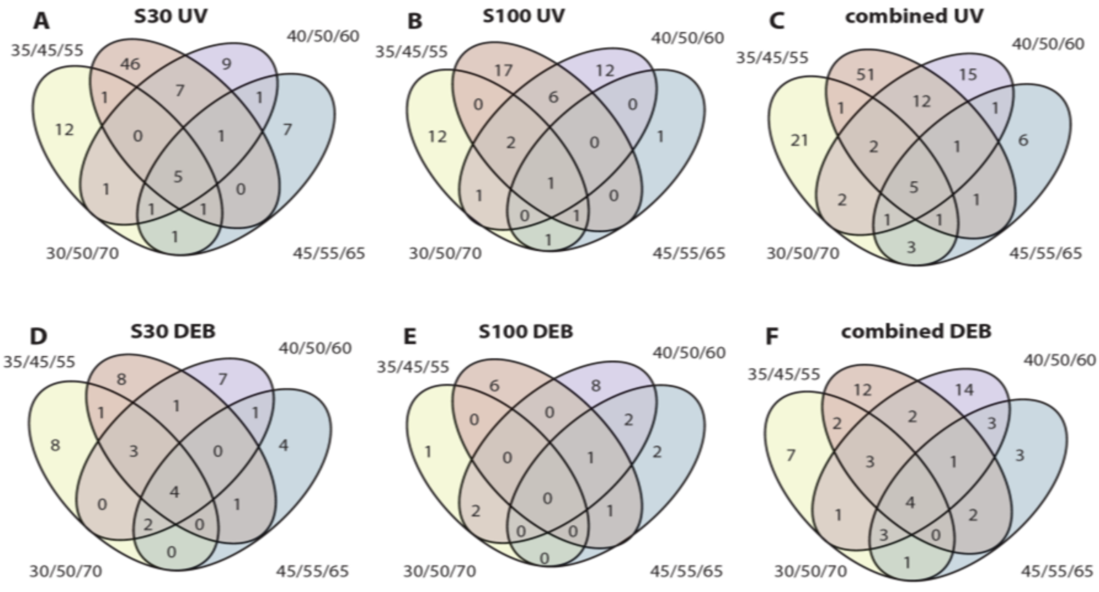


Figure 47: Comparison of different triple FAIMS runs in regards to aggregated proteins. A)-C) Triple CV runs for UV-crosslinked samples resulted in triple CV run -35/-45/-55 V to be identifying the greatest number of crosslinked proteins in both S30 and S100 samples, while covering most crosslinked proteins from other CVs, as well. This makes this triple CV run an excellent choice for analyzing UV-crosslinked data in a single shot experiment. **D)-F)** DEB-mediated crosslinked proteins inferred from CSM counts did not fall within the same triple CV run identified for UV-crosslinking. Most identified DEB-mediated crosslinked proteins stemmed from a triple CV run, comporised of CVs = -40/-50/-60 V. Moreover, coverage of shared crosslinked proteins with other CVs was identified to be high for this triple run, making it an excellent candidate for a single triple FAIMS experiment.

FAIMS separation of ion species boosts detection of certain ions by suppressing other ions. Classically, singly charged ions ($[M+H]^+$) that are not triggered for MS2 events but still enter the mass spectrometer and are part of the survey scan can be reduced dramatically by applying specific FAIMS settings. As such, the distribution of identified precursors that would later be successfully fragmented in MS2 and yield identifiable CSMs can be investigated for the different FAIMS setting explored. As such, figure 48 illustrates how many CSMs and proteins could be identified from differently charged precursor ions. In A)-B) for example, double charged precursor ions constitute the majority of CSMs, as well as precursors to identified crosslinked proteins. Even though a non-FAIMS run yields maximal CSMs for a doubly charged precursor, the corresponding crosslinked proteins amount to similar levels such as a double CV FAIMS run using CVs = -34/-45 V or a triple CV FAIMS run using CVs = -35/-45/-55 V. Overall, the highest FAIMS setting using CVs = -60/-70 V resulted in the fewest numbers of identified CSMs and corresponding crosslinked proteins. Analyses for triple charged precursors illustrate a sharp increase in detection of triple charged precursors in CSMs for said high CVs = -60/-70 V (C)-D)) that collapse to baseline levels on the protein level, however. Overall, a non-FAIMS run seems to be optimal for both UV-crosslinked and DEB-crosslinked samples in identifying triply charged precursor ions. Most interestingly, any charge state exceeding [M+3H^{3+}] could not be detected by triple FAIMS runs. Quadruply charged precuror ions were best identified at a double CV FAIMS run at CVs = -55/-65 V for UV samples, and CVs = -40/-50 V for DEB-mediated crosslinking events (E)-F)). Also, quintuply charged precursor ions were best isolated and identified at higher charged states, preferentially at CVs = -55/-65 for UV-crosslinking and CVs = -60/-70 V for DEB-mediated crosslinking. These data suggest that higher CV settings also correlate with identifications of higher charged states even in the case of crosslinked peptide-RNA heteroconjugates.

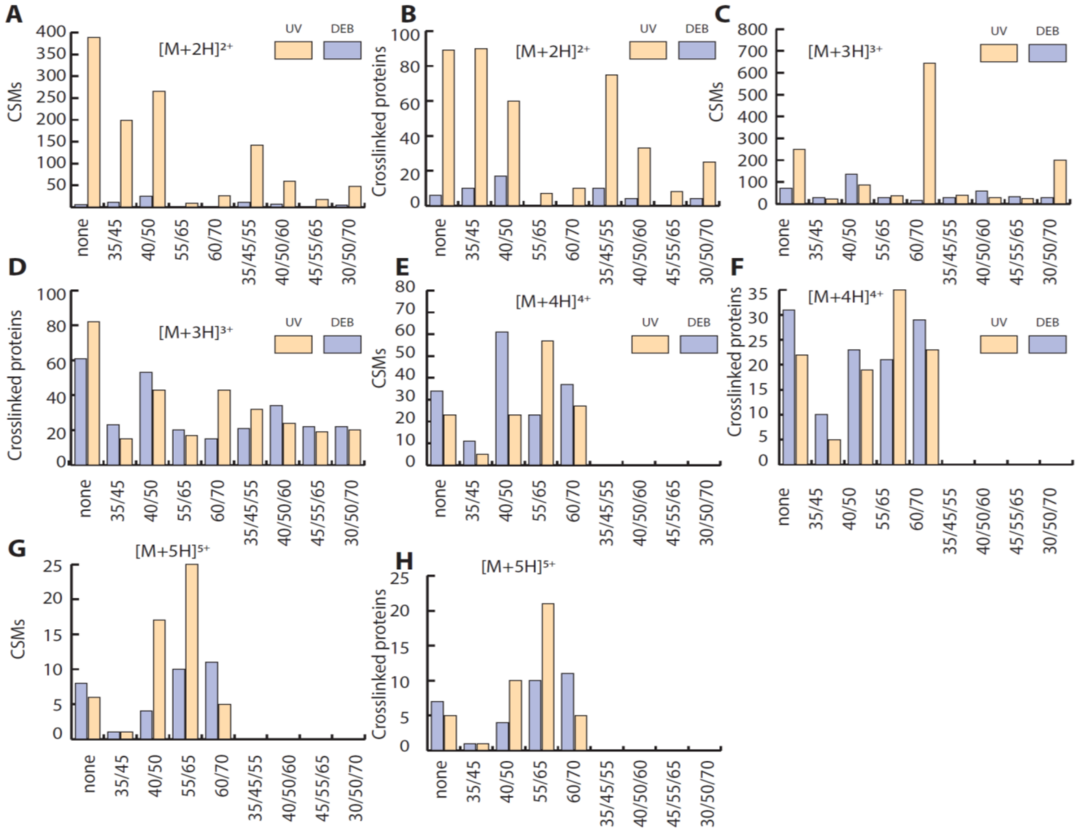


Figure 48: Analysis of the charge distribution from all FAIMS and non-FAIMS runs. A)-H) Number of CSMs and collapsed crosslinked proteins from differently charged crosslinked precursors. Most CSMs stem from doubly charged precursor ions (A), aggregating into the largest number of identified crosslinked proteins (B). Interestingly, UV-crosslinked peptide-RNA heteroconjugates resulting in a triply charged precursor ion were identified preferentially in a double CV run at CVs = -60/-70 V. DEB-crosslinked precursors were most often identified as quadruply charged precursor ions at a double FAIMS run with CVs = -40/-50 V that fell slightly short of the largest number of collapsed proteins from non-FAIMS runs (105 crosslinked proteins from non-FAIMS runs, 97 crosslinked proteins from double FAIMS run at CVs = -40/-50 V. Charge state [M+5H]$^{5+}$ was most often identified at CVs = -55/-65 V, also resulting in the greatest number of identified proteins from that double CV run. Interestingly, all triple CV runs failed to identify any CSM from precursors with a charge state higher than [M+3H]$^{3+}$.

Since there are different RNA nucleotides attached to the peptide, differences in adduct lengths may be interesting in regards to different FAIMS settings. To investigate this possibility, CSMs and crosslinked proteins from FAIMS and non-FAIMS runs were analyzed for their corresponding adduct length. Owing to stringent settings in NuXL, the maximal adduct length was set to two nucleotides, so differences may only revolve around a mononucleotide or dinucleotide being crosslinked to the peptide. Figure 49 A) and B) depict the total number of CSMs and corresponding crosslinked proteins identified with a mononucleotide adduct on the peptide mass. Apart from the -30/-50/-70 V triple fAIMS run, CVs = -60/-70 V from a double FAIMS run in addition to a -40/-50 V double run and the non-FAIMS measurement all seem to prefer monocleotide adducts, but those CV settings also yield the highest number of CSMs. Scrutinizing the collapsed lists of proteins, the non-FAIMS run featured mononucleotides slightly more often than other CV settings for both UV-crosslinked and chemically crosslinked samples. C) and D) depict similar distribution for dinucleotides added to the peoptide. Major increases in detection coincide with the aforementioned increases in mononucleotides and are likely due to an increase in CSM counts. Two distinctions are emerging however. First, DEB-crosslinked samples are more likely to carry dinuclotide adducts that are detected at a double CV FAIMS run with CVs = -40/-50 V. Second, an unusually high CSM count of 121 CSMs collapsing into 80 crosslinked proteins was observed for a triple CV FAIMS run using CVs = -35/-45/-55 V and UV-crosslinked samples.

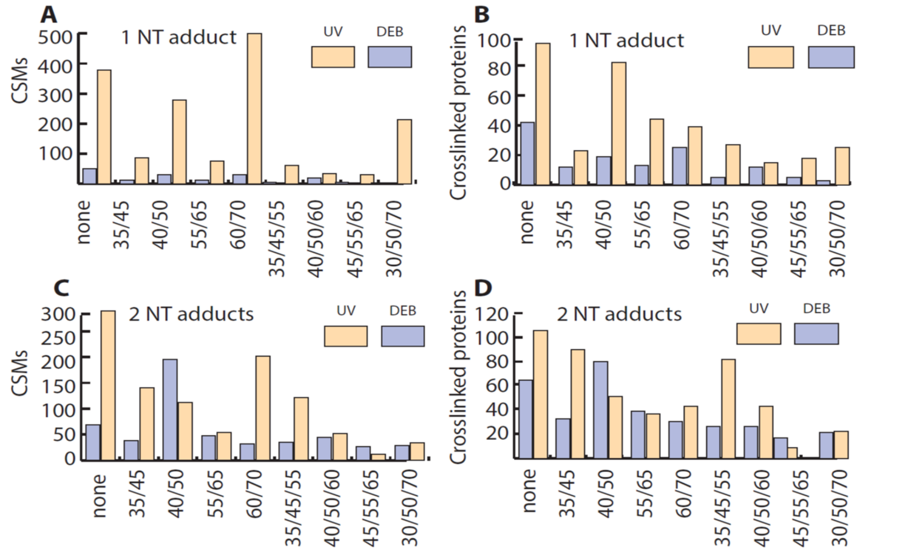


Figure 49: Comparison of the adduct length in FAIMS and non-FAIMS runs. A)-B) Number of CSMs and collapsed proteins according to the single adjunct nucleotide (NT) either chemically-crosslinked or UV-crosslinked to the peptide. While most single adduct masses corresponding to one nucleotide were identified through the double CV run at CV= -60/-70 V, only 88 proteins were mapped to those CSMs, falling slightly short of a non-FAIMS run that identified 103 crosslinked proteins. Also, DEB-crosslinked mononucleotides could best be identifed in a non-FAIMS run. **C)-D)** CSMs and aggregated proteins from peptides crosslinked to two nucleotides (NTs). Most UV-crosslinked dinucleotides were identified in non-FAIMS runs, also yielding the greatest number of crosslinked proteins. Chemically crosslinked dinucleotides could best be identified in a double CV run using CVs = -40/-50 V.

Lastly, the diffent RNA adduct masses on the peptide may be different depending on which CV settings were used to acquire FAIMS data. To investigate the possibility of a preferential setting for each RNA nucleotide, FAIMS and non-FAIMS runs were analyzed for the nucleotide identity of their adduct masses. Figure 50 illustrates both CSM and crosslinked protein level for each identifed nucleotide adduct identified in FAIMS and non-FAIMS runs. Most surprisingly, both CSM counts and crosslinked proteins counts for proposed adenine-containing nucleotides were identified at almost all FAIMS and non-FAIMS settings (A)-B)). In particular triple CV FAIMS run using CVs = -35/-45/-55 V resulted in a major increase in adenine-containing nucleotides UV-crosslinked to a peptide. Unfortunately, most CSMs fail to provide unamibigous evidence to locate the crosslinking amino acid and thus mainly rely on precursor identification, which oftentimes presented itself in MS2 spectra with little fragmentation and intact adduct masses (data not shown). Cytidine nucleotides are usually difficult to discern and their distribution across different CV settings basically reflects the increased overall CSM count of for UV-crosslinked samples compared to DEB-mediated crosslinking approaches and the resulting fluctuating numbers of crosslinked proteins. It seems that non-FAIMS measurements yield most cytosine-derived CSMs and crosslinked proteins with very little fluctuations outside the realm of systemic error and correlation to increases in overall CSMs for that sample (C)-D)). Two main distinctions can be made, however: DEB-mediated chemical crosslinking seems to prefer guanine-containing nucleotides best detected at double CV FAIMS run -40/-50 V (E)-F)). Interestingly, the total number of CSMs is higher for DEB-derived CSMs in this preliminary study than for UV-derived CSMs, but collapsed proteins result in tantamount numbers of G-crosslinked proteins in double CV FAIMS runs for both crosslinking approaches. Triple CV runs feature all more DEB-derived CSMs that aggregate into more crosslinked proteins identified that way than through UV-crosslinking. Lastly, UV-crosslinking showed an extreme preference for uracil-containing nucleotides both in on CSM and crosslinked protein level (G)-H)). Here, CSMs are almost always carrying uracil-nucleobases in the case of double CV FAIMS run with CVs = -60/-70 V, and non-FAIMS acquisitions.

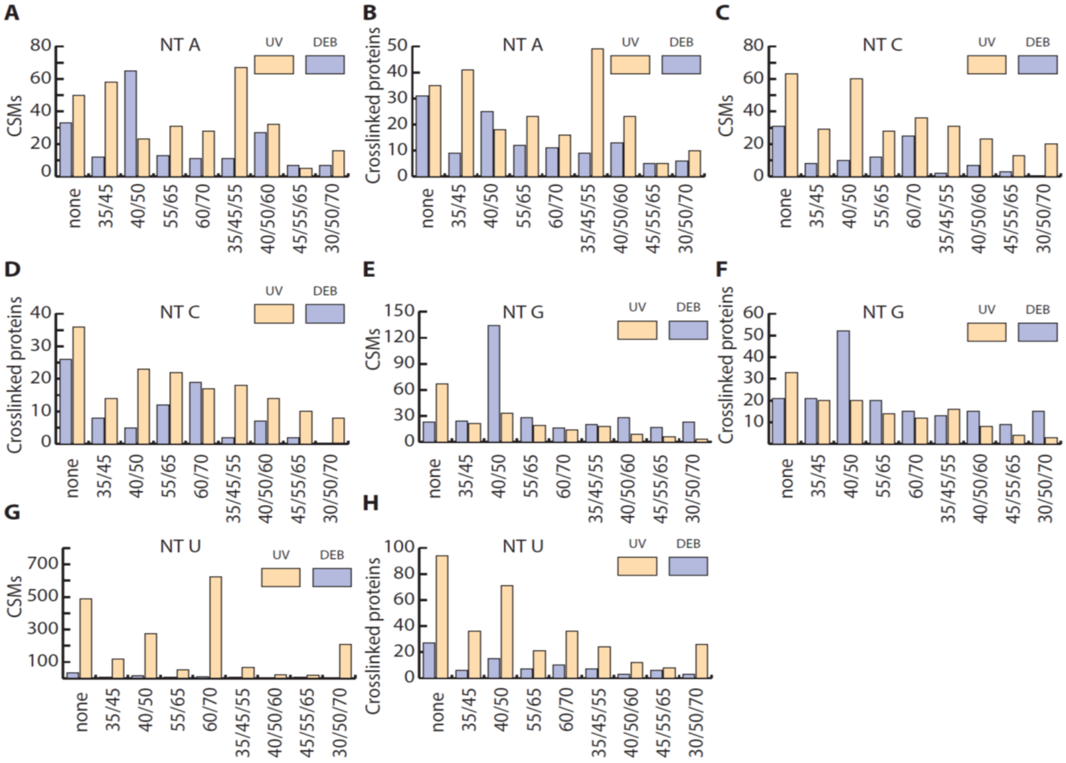


Figure 50: Comparison of the identity of the nucleotide adduct in FAIMS and non-FAIMS runs. A)-B) CSMs and aggregated proteins for adenine-containing nucleotides attached to the peptide via a covalent bond (UV-induced) or chemically by a spacer (DEB). Most surprisingly, NuXL reports a substantial number of potential A-crosslinks as illustrated. Manual inspections reveals, however, that most hits from NuXL fail to identify the crosslinking amino acid and are almost exclusively relying on an adenosine moiety being attached to the precursor ion. The triple CV run with Cvs = -35/-45/-55 V was found to select for this ions preferentially. **C)-D)** Cytdidine containing nucleotides attached to the peptide give a more homogenous distribution of CSMs and collapsed proteins with no-FAIMS runs identifying most crosslinks to cytosine nucleobases. Guanosine nucleobases are most often linked to the peptide via DEB, both on CSM level, as well as crosslinked protein level (**E)-F)**). Similarly, UV-crosslinking was yet again found to be highly preferential for uracil-containing nucleobases as despicted (**G)-H)**) This strong preference was found to be highest for CSMs at a double CV run with CVs = -60/-70 V, but aggregated crosslinked proteins peaked in non-FAIMS runs.

6.13.4 Major findings from combined FAIMS experiments

To summarize findings from DEB-crosslinked FAIMS experiments with complex cellular lysates, all 188 identified crosslinked proteins were combined into a STRING network depicted in figure 51 to illustrate how chemical crosslinking can provide meaningful information about *in vivo* interactions. The STRING network was computed with standard settings, except text mining associations between proteins were disabled, focusing on more physical evidence between the listed entries. The network was found to contain significantly more edges (relations between input proteins) than expected from random assignment against the reference background of *E. coli* proteins at a p-value of p = 1x10^{-16}. Most notably, 21 out of 56 ribosome-associated proteins were found to be present in significant numbers, forming a protein cluster that is enriched over the background at an FDR = 8.73x10^{-13}. This cluster of ribosomal proteins demonstrates that chemical crosslinking can be used effectively to to identify RNA binding ribosomal proteins in a similar fashion to UV-crosslinked proteins, as 12 out of 56 ribosomal proteins were identified by both chemical crosslinking and UV-crosslinking, making it the largest shared protein superfamily of the 63 identified shared crosslinked proteins (data not shown).

Additionally, structural similarities could also be detected on the protein domain level. Here, the OB-fold protein domain used in nucleic acid binding was found to be over-represented statistically at an FDR of 0.0017 and corresponding entries are denoted in yellow spheres. A good example for this would be the Polyribonucleotide nucleotidyltransferase Pnp that was identified to be crosslinked chemically in the K-homology domain known to bind RNA. Moreover, proteins of various RNA-binding functions including RNA-Polymerase (RNAP) and elongation factor G were identified to be chemically crosslinked at an FDR = 1x10^{-4} and are colored in green. Having identified subunits of RNAP to be crosslinked to RNA constitutes a prime example to validate chemical crosslinking coupled with mass spectromety in the context of biological annotation. In addition, other proteins with mRNA-binding properties such as Thr-tRNA ligase thrS and 30S ribosomal proteins S11, S6, and S4 were also identified and categorized as mRNA binding at an FDR of 0.0043. Identifying those proteins also further strengthens the validity of chemical crosslinking by providing more evidence of known RNA-interacting proteins that were indeed found to be crosslinked to RNA nucleotides. Interestingly, two other protein clusters were identified and revolve around chaparone DnaK, grpE, and groL. These chaperone proteins are possibly expressed in response to the DEB-treatment. More importantly, though, they are not known to be RNA-binding, but found to be either crosslinked to A or G nucleotides. The nucleotides do not necessarily stem from RNA in this case as nucleobases and nucleosides are also used in signaling and as energy sources by various enzymes. As such, the identified nucleobases adenine and guanine are possibly steming from ATP or GTP used as an energy source by those proteins. Both ATP and GTP are comprised of a nucleobase, namely adenine and guanine that could potentially react with a chemical crosslinker such as DEB. If nucleophilic amino acids were to be present in the nucleotide binding pocket of ATP or GTP binidn enzymes, DEB could possibly crosslink both protein and nucleotide-containing cofactor that would be detectable by mass spectrometry. A third protein cluster centers around the succinate-CoA ligase SucC protein that was found to be crosslinked to guanine nucleotides. Again, this guanosine nucleotide is possibly steming from GTP used in catalysis. Overall, this network demonstrates how chemical crosslinking can identify a multitude of proteins from various biological processes that still coalesce into a comprehensive nucleotide binding feature.

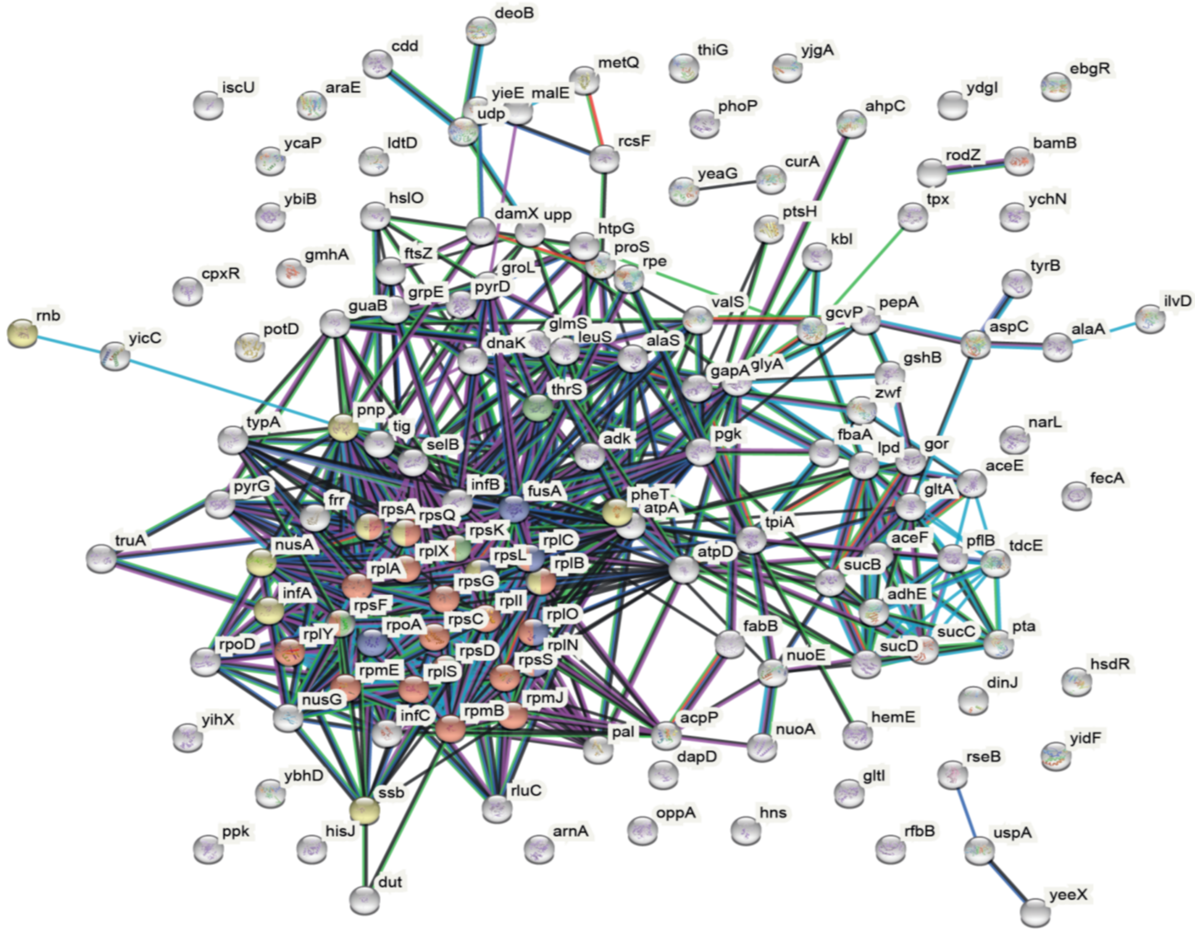


Figure 51: STRING networkf of chemically crosslinked proteins from FAIMS runs. Three protein clusters emerge from crosslinked input data (188 crosslinked proteins) that are all significantly enriched over the reference backgound of *E. coli* proteins. First, the ribosome-associated proteins colored in red cluster form a protein cluster in the lower left that is supported by ample evidence to be physically interacting and statistically enriched over the bacterial background (FDR = $8.73x10^{-13}$). Yellow sphere indicate proteins with an OB-fold protein domain used in nucleic acid binding that was statistically enriched at an FDR of 0.0017. Spheres in purple denote proteins of various RNA-binding functions including RNA-Polymerase and elongation factor G (FDR = $1x10_{-4}$. Moreover, a couple of green spheres indicate proteins with mRNA-binding properties such as Thr-tRNA ligase thrS and 30S ribosomal proteins S11, S6, and S4 that were also enriched at an FDR of 0.0043. The other two protein clusters revolve around chaparone DnaK, grpE, and groL that are likely expressed in response to the DEB-treatment, and where found to be either crosslinked to A or G nucleotides. It cannot be ruled out that those nucleotides are possibly steming from ATP or GTP used as an energy source by the enzyme. Similarly, the third protein cluster centers around the succinate-CoA ligase SucC protein on the right that was found to be crosslinked to guanine nucleotides also maybe steming from GTP used in catalysis. Overall, this network demonstrates how chemical crosslinking can identify a multitude of proteins from various biolgical processes that still coalesce into a comprehensive nucleotide binding feature.

As depicted in figure 51, there is a plethora of RNA-interacting proteins based on evidence provided in this work, as well as annotational input from consensus data bases. The aforementioned ribosomal protein cluster, comprising of both 30S ribosomal proteins and 50S ribosomal proteins, reveals 16 proteins that were found to be chemically crosslinked to RNA nucleotides via DEB. Even though almost all proteins were identified through one crosslinked peptide, multiple CSMs support the identified crosslinking site for said peptide. In some rare cases, however, multiple CSMs corroborated more than one crosslinked peptide, allowing multiple crosslinking sites to be identified for that ribosomal protein. The following examples illustrate how chemical crosslinking identifies crosslinking sites that fall within known RNA-interacting regions of ribosomal proteins, thus validating findings in their native biological context. A complete list of all identified crosslinking sites can be found in Appendix C, section 13.3.

Figure 52 illustrates some findings for ribosomal proteins with their native rRNAs. 16 ribosomal proteins could be identified to be RNA-interacting by chemical crosslinking coupled with mass spectrometry, with most proteins being identified through one crosslinked peptide, yet some were found to include more than one crosslinked peptide. Since the ribosome constitutes a mega complex, ribosomal proteins that were not identified to be crosslinked were omitted in figure 52, as well as the crosslinked amino acid being represented as boxed spheres to avoid unnecessary complexity in the figure. As depicted, crosslinking sites were found to be between histidine residues of 30S ribosomal protein S6, and 50S ribosomal proteins L3, and L1. Cysteines were identified in 50S ribosomal protein L31, L36, and 30S ribosomal protein S12. Additionally, two potential glutamate crosslinks were idientified in 50S ribosomal protein L2, as well as on in L19. Other crosslinking amino acids included threonine (50S ribosomal protein L2) and methionine (50S ribosomal protein L1). Overall, all identified crosslinking sites were located in close proximity to either 16S rRNA or 23S rRNA, documenting strong evidence to confirm crosslinking sites in a biological context.

Overall, 21 ribosomal proteins were identified to be crosslinked to RNA using extensive FAIMS analyses of chemically crosslinked *E. coli* cells and identified proteins stem from both components of the small 30S subunit and the large 50S subunit. Similarly, UV-crosslinking displayed ribosomal proteins to be the majority clustering protein group with 27 ribosomal proteins being identified by UV-crosslinking, of which 12 ribosomal proteins were identified by both approaches (data not shown). There did not appear to be a predilection for either 30S or 50S subunit in regards to the crosslinking method, displaying an more scrambled distribution of identified proteins from both subunits in both cases. Moreover, the total overlap of identified crosslinked proteins between UV-crosslinking and chemical crosslinking was found to be approximately 20% (data not shown). This finding illustrates even though there are substantial differences in identifying a particular subset of different RNA-interacting proteins such as the ribosomal protein family, there is still common ground between the two approaches, and the complementary aspect enlarges information gain tremendously.

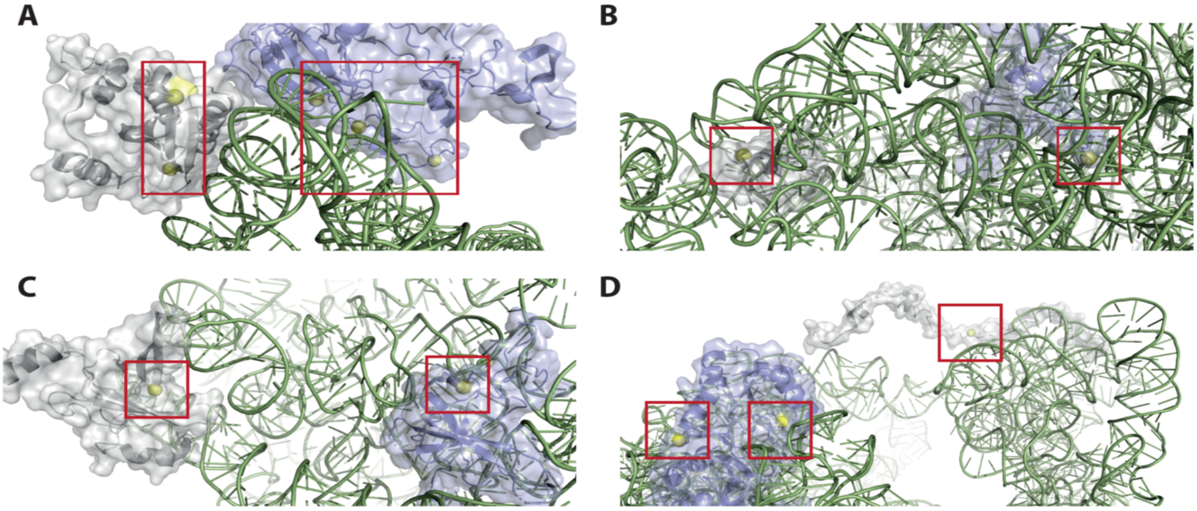


Figure 52: Selected DEB-mediated crosslinking sites of ribosomal proteins from FAIMS runs. **A)** 30S ribosomal protein S6 (RS6), 50S ribosomale protein L2 (RL2), and the 16S rRNA were found to be crosslinked at possibly 5 positions. RS6 is depicted in gray, the rRNA in green, and RL2 in blue. Two histidines within the same peptide from RS6 were identified to be crosslinked to the 16S rRNA, indicated by yellow spheres on the left. Similarly, a glutamic acid residue was found in RL2 to be crosslinked with rRNA, but the peptide contains two glutamates and an unambigious position could not be identified. The E residues, however, are only five residues apart, and all identified crosslinking sites are within close proximity to the rRNA. **B)** 50S ribosomal protein L36 (RL36 gray), 50S ribosomale protein L3 (RL3, blue), and the 23S rRNA were found to be crosslinked at two positions. A cysteine residue from RL36 found on the left is adjacent to the rRNA helix, similar to the crosslinked histidine from RL3 found on the right. Both crosslinking sites were identified with both uridine nucleotides and guanosine nucleotides. **C)** Crosslinking sites between 50S ribosomal protein L19 (RL19 gray), 30S ribosomale protein S12 (RS12, blue), and the 16S rRNA were identified at two positions. A glutamate residue from RL19 was found to be crosslinked to rRNA as shown on the left, similar to a cysteine residue from RS12 on the right. **D)** 50S ribosomal protein L31 (RL31 gray), 50S ribosomale protein L1 (RL1, blue), and the 16S rRNA were found to be crosslinked at three positions. Again, a cysteine residue from RL31 was chemically crosslinked to rRNA, depicted in the background right. Two different peptides from RL1, containing a crosslinked histidine and a crosslinked methionine each, pinpointed the RNA interface with 16S rRNA at the positions to the front left. Again, all crosslinking sites fall within a couple angstroms of the rRNAs. PDB model: 3J9Z [190]

In addition to the well-described ribosomal proteins depicted in figure 52 that were identified to be interacting with rRNAs, two additional examples further strengthened the applicability of chemical crosslinking coupled with mass spectrometry. Ribosome-associated elongation factor Tu2 (EFTu2) and tRNA pseudouridine synthase A (TruA) also featured many CSMs that could locate protein-RNA interfaces accurately and reliably. Figure 53 depicts findings from FAIMS-derived experiments. CSMs from EFTu2 contained three crosslinked peptides that indicate different RNA-protein interfaces. EFTu2 is part of a supercomplex involving two types of rRNAs (16S rRNA, and 23S rRNA) and different ribosomal proteins, two of which were included in figure 53 to demonstrate the complex. Histidines, a lysine, and cysteine were found to be the crosslinking amino acids in three different peptides. Most surprisingly, histidine/cysteine residue were identified to be crosslinked to guanosine nucleotides right at the GTP-binding site the complex involving Leu-tRNA. This raises the question if the identified nucleotide is stemming from the tRNA or the cofactor GTP which could not be distinguished by mass spectrometric analyses since the adduct mass corresponded to a guanosine nucleobase attached to the ribose (phosphate loss). The corresponding cysteine residue three amino acids apart also alluded to guanine-containing moieties, showing internal consistency. Besides additional histidine crosslinks, a lysine crosslink in close proximity to 16S rRNA was identified just two residues apart from a crosslinked histidine residue, pinpointing the RNA-protein interface to this region of EFTu2. Lastly, tRNA pseudouridine synthase A (TruA) was found to contain a histidine that was identified to be crosslinked to both guanine-containing nucleotides and uracil-containing nucleotides. Incidentally, the corresponding Leu-tRNA in the crystal structure shows that there are three nucleotides that are known to be modified by TruA, and fit perfectly with crosslinking data. U_{38}, G_{39}, and G_{40} were discribed to be converted into pseudouridine by TruA, and fit the crosslinked nucleotide profile.

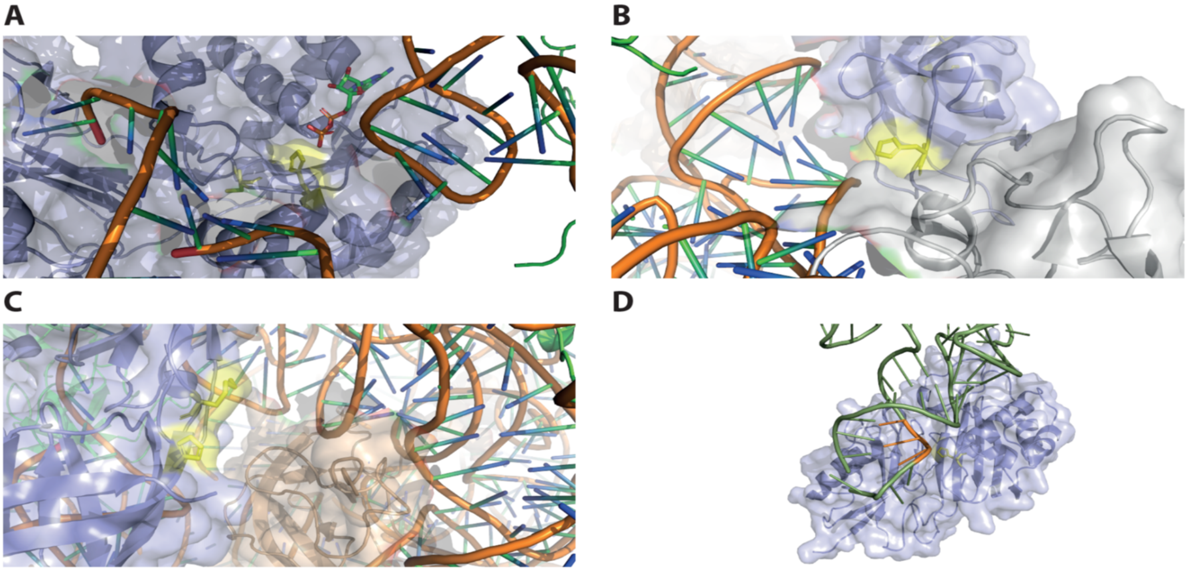


Figure 53: Selected DEB-mediated crosslinking sites of EFTu2 and TruA from FAIMS runs. **A)** Two crosslinking sites between the 23S rRNA and EFTu2 were identified in close proximity to the GTP-binding site. The histidine protrudes into the rRNA binding cavity and is also juxtaposed to the GTP molecule depicted in the center. The crosslinking amino acid was indeed found to be a guanosine nucleotide, positing that chemical crosslinking also identifies nucleotide-containing cofactors. Similar results were found for the cysteine residue located five amino acids downstream from the histidine residue. **B)** Another histidine crosslink between EFTu2 (blue) and Phe-tRNA in close proximity. The gray protein in front depicts 50S ribosomal proteine L11. **C)** Two crosslinking sites between a lysine residue (front) and histidine residue two amino acids downstream from EFTu2, interacting with 16S rRNA. The lysine residue is particularly close to the rRNA, making it very accessible to chemical crosslinking. 30S ribosomal protein S12 is depicted in orange up front. **D)** Lastly, tRNA pseudouridine synthase A was found to be crosslinked to both uridine nucleotides, as well as guanosine nucleotides found in Leu-tRNA. The histidine residue is adjacent to the three nucleotides U_{38}, G_{39}, and G_{40} of Leu-tRNA that are known to be modified by TruA. Again, all crosslinking sites fall within a couple angstroms of the RNAs, validating findings for chemical crosslinking coupled with mass spectrometry biologically. PDB codes: 4v69 (EFTu2), 2nr0 (TruA)[194, 195]

7 Recommendations for peptide-nucleotide crosslinking

Having analyzed FAIMS and non-FAIMS runs exhaustively, a few number of general rules become evident and may be formulated with optimistic caution as further replicating experiments need to be performed to ensure the validity of the rules:

1. Simple protein-RNA complexes with a few proteins generally yield the best results in terms of spectral quality and CSM counts, regardless of the crosslinking method.
2. Aromatic residues such as tyrosines are generally best detected by UV-crosslinking, whereas nucleophilic amino acids such as lysines, cysteins or serines generally prefer to be DEB-crosslinked.
3. If sample amounts are not an issue, applying both UV-crosslinking and DEB-mediated crosslinking to complex samples will generally increase identifications of crosslinked proteins by about 30%. 60:10:30% ratio unique UV, shared, and unique DEB proteins.
4. If one wishes to increase the number of CSMs to validate and confirm a crosslinking site numerously, non-FAIMS runs should be considered that also boost UV-crosslinked dinucleotides or cytosine-containing nucleotides.
5. Non-FAIMS runs also aid in detecting UV-crosslinked dinucleotides attached to the peptides, as well as slightly boosting identification of adenine and cytosine containing nucleotides when UV-crosslinked.
6. If charge states higher than 3+ are required, do not use triple FAIMS runs.
7. If a charge state of +5 is desired, a double CV run of -55/-65 V may be beneficial
8. Triple CV runs yield generally lower numbers of crosslinked proteins, but the highest number of crosslinked proteins for UV-crosslinked samples can be indentified at CVs = -35/-45/-55 V, for DEB-crosslinked samples at CVs = -40/-50/-60 V.
9. A double CV run of -40/-50 V also supports DEB-crosslinked dinucleotides
10. If the RNA contains numberous G nucleotides and is known to bind to the RBP using G-rich sequences, use DEB-crosslinking.
11. If the RNA contains numberous U nucleotides and is known to bind to the RBP using U-rich sequences, use UV-crosslinking.
12. Adenine-containing nucleotides crosslinked to a peptide are best identified through a triple CV run if UV-crosslinked (-35/-45/-55 V) or a double CV run if DEB-crosslinked which would also identify cytosine-containing crosslinks (-40/-50 V). The adenosine nucleotide oftentimes does not fragment properly and can only be identified on MS1-level, though.
13. When in doubt, a double CV FAIMS run with CVs = -40/-50 V returns very good results for both UV-crosslinked and DEB-crosslinked samples.

8 Chemical crosslinking of *Bacillus subtilis* cells *in vivo*

To extend DEB-mediated crosslinking *in vivo*, another bacterial organism was chosen to be crosslinked and analyzed in single no-FAIMS runs in a proof of principle fashion. This preliminary data was stringently analyzed with NuXL at an FDR set to 1%. Figure 54 illustrates findings from *B. subtilis.* The STRING network was not found to be that extensive compared to the network obtained from *E. coli*, featuring 24 crosslinked proteins, yet this initial experiment only contained a single-shot mass spectrometric non-FAIMS run. Still, the identified proteins fit previous finding quite well, as the vast majority of crosslinked proteins was identified to be ribosomal, aggregating into the major cluster. Besides that, tRNA binding proteins (Thr-tRNA ligase thrS, Val-tRNA ligase ValS, and Trp-tRNA ligase trpS) were found to be enriched over the background of *B. subtilis* at an FDR of 4.7x10^{-3}, illustrating a decisive preference in the crosslinked protein set. In fact, 50% of the identified proteins were annotated to be ribosomal, amounting to a significantly enriched set of proteins over the background at an FDR of 1.72x10^{-14}. Even when the remaining proteins were scrutinized for an overarching binding preference, nucleotide binding was found to be highly enriched at an FDR of 2.2x10^{-4}. Ultimately, tRNA processes and translation emerged as the two most significantly enriched cellular processes, determined in a Fischer exact test with $p < 0.001$. Following this notion of truly having primarily identified RNA-binding proteins, identified crosslinked hits were scrutinized.

Two example of ribosomal proteins from *B. subtilis* demonstrate that chemical crosslinking can be used quite effectively to identify RNA-binding proteins even in single run measurments. As depicted in figure 54 B), 50S ribosomal protein L28 was found to be crosslinked to 23S rRNA via a cysteine residue depicted in yellow. The residue is in direct proximity to the rRNA depicted in green, allowing for chemical crosslinking to occur. Additionally, 50S ribosomal protein L13 was found to be crosslinked to the same 23S rRNA via a histidine residue, depicted in C) that stacks on top of the RNA nucleobases in the model, perfectly explaining the identified crosslink. Moreover, 55% of the CSMs indicated that the crosslinking nucleotide was of a guanine nature, wheras only 28% of CSMs indicated a uracil-containing nucleotide to be crosslinked, meaning that more than half of the CSMs would likely have been lost or dramatically reduced had only UV-crosslinking been applied. Cytosine-containing nucleotides added up to 15% of the CSMs, while only 2% of the crosslinked nucleotides were annotated to be adenosine nucleotides. The strong prefernce for guanine nucleobases was proven once again in the case of chemical crosslinking using DEB.

Additionally, the inosine-5'-monophosphate dehydrogenase GuaB was identified to be crosslinked to a uridine nucleotide in both *E. coli* and *B. subtilis.* The enzyme catalyzes the conversion of inosine 5'-phosphate (IMP) to xanthosine 5'-phosphate (XMP), effectively adding a keto group to the purine ring. While NAD$^+$ is used as a cofactor for catalysis, involvements of pyrimidines such as uracil nucleobases have not been described yet. In *E. coli*, GuaB was found repeatedly in various FAIMS and non-FAIMS runs, bolstering its identification as a truly nucleic acid binding protein across different bacterial species. The protein is known to be involved in the *de novo* synthesis of guanine nucleotides as the rate-limiting step, and thus plays an important role in cellular growth and responses to changes in the environment [196]. It should be restated, however, that the data from *B. subtilis* serves as a solid starting point as a proof of concept study and lacks the proper number of replicates similar to the comprehensive analyses performed on *E. coli* samples, but identifying RNA-binding proteins in their described context validates findings.

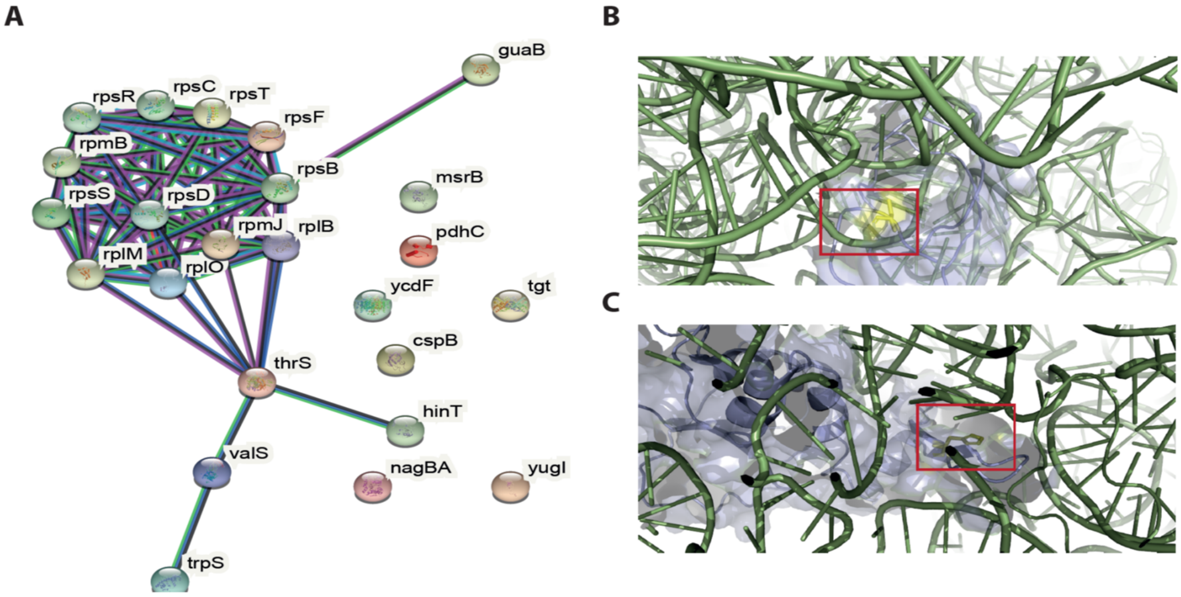


Figure 54: DEB-mediated crosslinking of *B. subtilis*. A) STRING network of identified crosslinked proteins. The network clusters around ribosomal proteins of both samll and large subunit. Additional proteins include Threonine-tRNA ligase 1 (thrS), Valine-tRNA ligase (ValS), and Trp-tRNA ligase (trpS), illustrating translational processes to be highly enriched in this set of crosslinked proteins. **B)** 50S ribosomal protein L28 crosslinked to 23S rRNA via a cysteine residue. The cysteine is in close proximity to the rRNA, allowing for chemical crosslinking to occur. Similarly, 50S ribosomal protein L13 was found to be crosslinked to 23S rRNA via a histidine residue that almost stacks on top of the rRNA nucleobase depicted in **C)**. These findings provide further evidence that chemical crosslinking coupled with mass spectrometry povides value insights into RNA-protein interactions crosslinked *in vivo*. PDB code: 6HTQ [197].

Lastly, DEB-crosslinking identified queuine tRNA-ribosyltransferase Tgt. Tgt was found to be crosslinked to U nucleotides exclusively via DEB at Cys-271. This finding coincides incidentally with experiments with Dnmt2 from *S. pombe* described above, where the G nucleobase is methylated by Dnmt2 at position-34. Here, Tgt catalyzes the frist reaction to exchange the nucleobase G found at the Wobble position-34 with the queuine precursor 7-aminomethyl-7-deazaguanine (PreQ1) [198]. The identified crosslinked peptide 265-GVDMFDCVL-276 has high sequence similarity to Tgt from *Ec. coli* with two amino aid substitutions: V-266 to I-260 and L-276 to M-267. The corresponding sequence in *E. coli* (259-GIDMFDCVM-267) also contains an acitve site Asp that initiates catalysis [198, 199].

9 Chemical crosslinking of human HeLA cells *in vivo*

To expand the applicability of chemical crosslinking coupled with mass spectrometry to eukaryotic cells, the same workflow was applied to HeLa cells, investigating chemical crosslinking of human cells *in vivo*. Similar to the preliminary study using *B. subtilis* crosslinking experiments of HeLa cells were conducted in a proof of concept fashion. Even though 202 CSMs were identified by NuXL, the aggregate number of crosslinked proteins was only found to be 28. While not ideal in terms of prolific output, identifying crosslinked proteins by mulitple CSMs strengthens validity in the identification. Also, one needs to remember that no FAIMS was used to select for different ion species in this preliminary study, as well as the same standard workflow having been applied to human cells, which may not be optimal. Since HeLa cells contain a lot of DNA, sample preparation was impeded by massive amounts of DNA making liquids very viscous and harder to work with. However, to show that chemical crosslinking coupled with mass spectrometry can be used to reliably identify protein-RNA crosslinking sites across different species using the outlined approach, the same workflow was faithfully applied.

Figure 55 A) depicts a STRING network constructed around the 28 proteins identified to be crosslinked chemically to RNA. Two small clusters could be identified, one being ribosomal proteins, and the other clustering around histone proteins. While the ribosomal protein cluster was fairly small compared to findings even from *B. subtilis*, it did speak to a coherent biological narrative of truly RNA-binding proteins being identified at an FDR value of 0.0115. The second cluster constitutes three histone proteins, inducing a significant enrichment of histone proteins (40 histones, or histone-associated proteins) over the genetic human background at an FDR of 5.9x10^{-4}. Interestingly, the histone proteins that were identified all fall within the linker histone protein family of histone H1.4 or H1.2. Investigating the ribosomal protein cluster identified in human HeLa cells, figure 55 B)-C) depicts how the identified crosslinking sites fit well with published models of the ribosomal complex with its cognate rRNA. The three examples strengthen the suggestion of a preference for nucleophilic amino acids such as cysteins and hisitidiens by DEB, as well as demonstrating how guanosine nucleotides are reliably detected using chemical crosslinking. The ribosomal proteins were found to also be highly enriched over the human genetic background at an FDR of 9.8x10^{-4}, albeit only 4 ribosomal proteins were identified out of the known 130 ribosomal and ribosome-associated proteins.

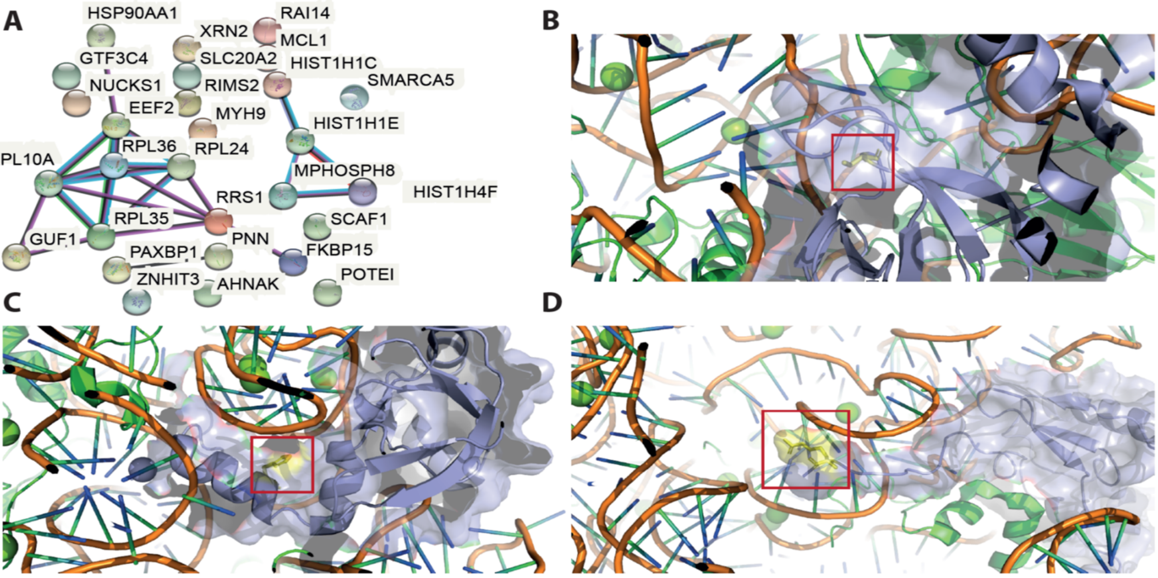


Figure 55: DEB-mediated crosslinking of HeLa cells A) STRING network of identified crosslinked proteins. Overall, 202 CSMs from NuXL output collapsed into 28 crosslinked human proteins with little clustering due to the small number of input proteins. Two small clusters could be identified, however. The bigger cluster on the left aggregates ribosomal proteins into a defined conglomerate of proteins and the small cluster on the right refers to histone proteins. **B)** 60S ribosomal protein L14 (blue) was identified to be chemically crosslinked to 28S rRNA via a cysteine depicted in yellow. The uracil nucleotide of 28S rRNA is in direct proximity to the crosslinked amino acid side chain. **C)** 40S ribosomal protein S15 (blue) was identified to be crosslinked to 18S rRNA via a histidine crosslink. The identified guanosine nucleotide of the adjacent RNA helix is in close proximity, validating the crosslinking site contextually **D)** 60S ribosomal protein L26 (blue) was identified to be crosslinked to 28S rRNA via another histidine crosslink. The histidine side chain is in direct proximity of the RNA helix, making contact with L26 at the designated region of the protein. Again, the identified crosslinking site validates the chemical crosslinking approach as data fits well with previously obtained data. PDB code: 5AJ0 [200].

10 Discussion

Chemical crosslinking coupled with mass spectrometry can be used to effectively identify protein-RNA interfaces at amino acid resolution. The first part of this thesis described necessary requirements for correctly identifying chemical protein-RNA crosslinks (i), leading to exemplifying chemical crossslinking using multiple protein-RNA complexes (ii). A direct comparison to canonical UV-crosslinking (iii) followed, and the later parts dealt with supervised machine learning approaches to classify spectral data from complex data sets (iv) before demonstrating the applicability and utility of the devised approach to different complex biological whole-cells system (v). Those five sections are discussed in regards to the feasibility, reliability, and utility of chemical crosslinking coupled with mass spectrometry.

10.1 Establishing a workflow to identify chemically crosslinked protein-RNA heteroconjugates

Testing the feasibility of chemical crosslinking of protein-RNA complexes using the single RNA-binding protein Hsh49 and radioactively labeled synthetic RNAs requires purification of sufficient amounts of Hsh49 protein. The purification methods described for Hsh49 alone and Hsh49 copurified with Cus1 all yield highly purified proteins. Omitting the heparin column as the third purification step increases protein yield, but suffers in protein purity, akin to loading the TEV-digested Hsh49:Cus1 complex onto a heparin column, omitting the second NiNTA step. In any case, a high salt wash with 1M NaCl removes most endogenous RNA and greatly improves the 260/280 nm ratio. Crosslinking Hsh49 alone with synthetic RNAs, comprised of a stretch of polynucleotides of the same nature, the crosslinkability of Hsh49 and synthetic RNA was demonstrated successfully. Surprisingly, both G and A nucleotides were detected by radioactive labeling of the polynuceltode RNA, which were previously rarely observed reliably using UV-crosslinking [46]. Since radioactive RNA can be detected in minuscule amounts using SDS-PAGE as described for all four nucleotides and all crosslinkers, a workflow needed to be developed that would accommodate higher amounts of protein and RNA to be detected by mass spectrometry.

The primary goal of establishing a workflow supporting the identification of spectra that unambiguously show protein-RNA heteroconjugates relies on two main pillars: sample preparation and analysis of spectral data. Having identified known and novel true crosslinking sites using the workflow for chemical crosslinking coupled with mass spectrometry described above, the application of affinity-based TiO_2 enrichment provides a robust strategy not only for UV-crosslinked samples, but also chemically crosslinked samples. Enriching UV-crosslinked peptides by TiO_2 has worked well for simple protein-RNA complexes, as well as complex lysates from UV-crosslinking experiments. [3, 5, 10, 16, 46]. This workflow remains relatively simple and true to its original UV-based origin, and also allowed for direct comparison between chemically crosslinked and UV-crosslinked samples. It should be noted however, that other methods to separate crosslinked nucleotides from pure protein do exist and have been used successfully. For example, basic reversed-phase chromatography can be used effectively to separate crosslinked proteins/peptides from linear peptides, as well as size exclusion chromatography, as previously shown by others for UV-crosslinked samples. [10]. The zero-length crosslinking event occurring during UV-irradiation is well described, but chemical crosslinking agents occupy a certain space and thus generate spacer lengths between the two crosslinking moieties.

10.1.1 The chemical crosslinkers 2IT, DEB, and NM have short spacer lengths

Unlike UV-crosslinking that forms a direct covalent bond between peptide and nucleotide, termed a zero-length crosslinking events, chemical crosslinkers occupy physical space. This spacing length is contingent on the chemical structure of the crosslinker and its three-dimensional structure, which in turn, depends on the bond lengths between the atoms of the crosslinker. 2IT has been used for decades in protein-protein crosslinking and has been studied extensively in the context of structural analyses, establishing a reliable spacer length of 8.1 Å [135, 201]. While both DEB and NM have been described to be crosslinking reagents, they have not been used in the context of structural biology to the same extent. Therefore, little information on their spacer length is available, but the distances can be estimated using bond lengths of the involved atoms.

A DEB-mediated cross-link between the N^7-position of guanosine and the terminal amino group of lysine is depicted in figure 56 with bonds between differently hybridized atoms also colored differently. The following bond lengths were taken from tabulated x-ray diffraction and neutron diffraction data found elsewhere [202]. The N^7 atom of the purine ring in guanosine is sp^3 hybridized that can form a single bond with a bond length of 1.485 Å to one carbon of DEB (magenta colored bond). All bonding atoms in DEB are sp^3 hybridized, but form bonds of different lengths due to different substitutions on different C-atoms. The following C-C bond (blue color) is 1.513 Å long, while the following C-C (red color) is slightly longer (1.524 Å). Another C-C bond of 1.513 Å forms the end of the DEB molecule that can form another short bond 1.469 Å to the sp^3 hybridized pyramidal N-atom of the terminal amino group found in lysine. Flattening all chemical bonds into a plane and adding their respective bond lengths results in a hypothetical length of 7.504 Å as an upper maximal distance from the N^7-position of guanine to the terminal amino group of lysine. Accounting for three-dimensional arrangements using MOLView, a minimal distance of 6.19 Å has been calculated [203].

When distances between DEB-cross-linked nucleotides and the C_α-atom of the peptide bond are taken into considerations, side chain compositions are important. For instance, the lysine residue mentioned above is comprised of sp^3-hybridized atoms with slightly different bond lengths colored in brown in figure 56: The first C-N bond connecting the pyramidal N-atom to the carbon backbone of the side chain is 1.469 Å long, while the three subsequent C-C bonds are 1.53 Å in length. The last C-C bond to the C_α -atom is 1.524 Å long, resulting in a fully extended and flattened chain length of 7.532 Å. MOLView calculations accounting for spatial arrangements yields a value of 6.25 Å [203]. The maximal distance between a nucleotide and the C_α-atom of a lysine side chain can hence be estimated to be 15.036 Å, whilst the minimal distance using MOLView computes to 12.16 Å.

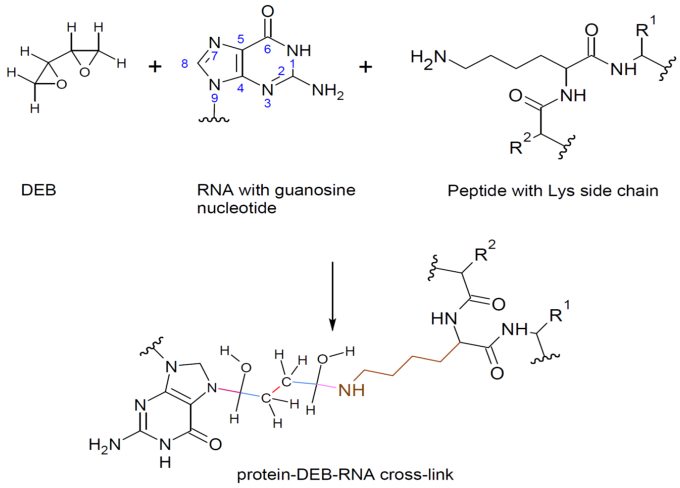

DEB
RNA with guanosine nucleotide
Peptide with Lys side chain
protein-DEB-RNA cross-link

Figure 56: Spacer length of a DEB-mediated cross-link. DEB spacer lengths can be estimated by bond lengths of the atoms involved in the cross-link. Here, DEB connects a guanine nucleotide at the N^7 to a lysine residue of a peptide. All connecting chemical bonds are formed via different sp^3-hybridized carbon atoms. Different colors indicate differently hybridized atoms. See text for details on differently hybridized atoms and resulting bond lengths. Note that bond lengths of C-C bonds depends on different substitutions on those carbons atoms. Flattening all chemical bonds into a plane and adding their respective bond lengths results in a hypothetical length of 7.504 Å as an upper maximal distance from the N^7-position of guanine to the terminal amino group of lysine. Accounting for three-dimensional arrangements using MOLView, a minimal distance of 6.19 Å has been calculated.

NM-mediated cross-links between the N^7 of guanosine and the terminal amino group of lysine are depicted in figure 57. Again, differently hybridized atoms and their respective bonds are colored differently and the bond lengths were taken from the same diffraction data [202]. Akin to the DEB-mediated cross-link, the N^7 atom of the purine ring in guanosine is sp^3 hybridized forming a single bond with a bond length of 1.485 Å to one carbon of NM (blue colored bond). All bonding atoms in NM are sp^3 hybridized, forming bonds of various lengths: the following C-C bond (red color) is 1.524 Å long, while the following C-C (magenta color) is slightly shorter (1.469 Å). Another C-C bond of 1.524 Å forms the end of the NM molecule, colored in red, that can form another bond 1.469 Åto the N-atom of the terminal amino group found in lysine. Planarizing all bonds again results in a hypothetical length of 7.471 Å as an upper maximal distance from the N^7-position of guanine to the terminal amino group of lysine. The minimal distance allowing for spatial considerations using MOLView has been calculated to be 7.29 Å [203]. Similar to considerations taken into account for maximal C_α-atom distances, NM forms both maximal distances computed from planarized bond lengths and minimal distances from MOLView calculations. In the case of lysine side chains mentioned above, the distance between NM-cross-linked nucleotides and the C_α -atom of the peptide bond spans over the bonds calculated above, as well as the aforementioned lysine side chain colored in brown in figure 57. The maximal distance between a nucleotide and the C_α-atom of a lysine side chain has been estimated to be 16.472 Å, whilst the minimal distance using MOLView computes to 13.50 Å.

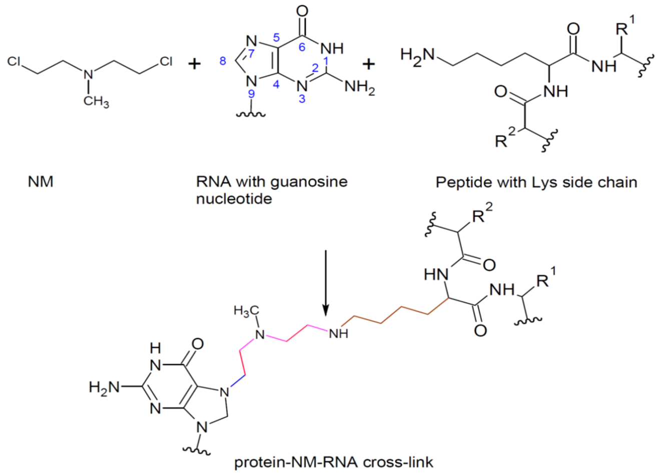

NM
RNA with guanosine nucleotide
Peptide with Lys side chain
protein-NM-RNA cross-link

Figure 57: Spacer length of a NM-mediated cross-link. Spacer lengths of NM-mediated cross-links can be estimated by bond the lengths of involved atoms. Akin to figure 56, NM connects a guanine nucleotide at the N^7 to a lysine residue of a peptide. All connecting chemical bonds are again formed via different sp^3-hybridized carbon atoms. Different colors indicate differently hybridized atoms and their corresponding bonds as explained in the text. Note that bond lengths of C-C bonds differ in lengths depending on the substitutions on those carbons atoms. Planarizing all bonds and adding their respective bond lengths results in a hypothetical length of 7.471 Å as an upper maximal distance from the N^7-position of guanine to the terminal amino group of lysine. Using MOLView to account for three-dimensional arrangements, a minimal distance of 7.29 Å has been calculated. A small difference between maximal and minimal distance indicates that the bonds are mostly planar.

10.1.2 Precursor stabilization for different chemically crosslinked nucleotide adducts

In contrast to UV-crosslinked peptide precursors that fragment relatively easily during CID, yielding a sufficient number of fragment ions carrying the RNA-shifted adduct, chemically crosslinked peptides behave more stably during CID. As alluded to in figure 17 D) and figure 18 B), both precursor masses corresponding to the peptide, crosslinker (DEB and NM. respectively), and nucleotide adduct remain largely intact, leaving a high intensity signal for the unfragmented precursor in that MS2 spectrum. While this does not necessarily pose a problem, if CID-fragmentation still results in enough fragment ions to locate the crosslinking amino acid, yet as depicted in figure 17 D) for example, scanty fragment ions provide little evidence to locate the crosslinking site on the peptide. DEB was also previously found to crosslink DNA at either G or A nucleotides in a very stable fashion that could be detected using mass spectrometry at high precursor intensities of largely unfragmented purine nucleotides linked by DEB [121]. This resistance to sufficient CID fragmentation of crosslinked dinucleotides may stem from the crosslinking reagent DEB, as stabilized purine nucleobases were frequently observed in this study. A similar case can be made for NM as the crosslinking reagent. For NM in particular, it has been described that NM induces adenine-containing DNA-DNA crosslinking lesions that are highly stable, so a potentially stabilizing effect even in the case of single RNA nucleotides crosslinked to protein may not be that surprising [204]. Hence, chemical crosslinking using either DEB or NM stabilizes the precursor at the nucleotide position, possibly by providing a larger backbone to dissipate the energy. To address this phenomenon, one would need to vary the energies applied during fragmentation and observe if the precursor fragments better at higher fragmentation energies.

Additionally, stabilized precursors were more often observed with the adenosine-containing nucleotide than guanosine-containing nucleotides, even though both nucleotides belong to the purine family. More specifically, insufficient fragmentation is also reflected in the localization score of NuXL, where a high localization score > 10 [12] indicates that the crosslinked amino acid could likely be identified accurately. Looking at all DEB-mediated crosslinking experiments using FAIMS, only 22% of all supposed adenosine-bearing adducts (26.7% of all total CSMs) could actually be localized to a specific amino acid, whereas 48% of all DEB-crosslinked guanosine-containg adducts (50.7%) could also be localized to a specific amino acid residue. This indicates that adenosine nucleotide crosslinks are twice as stable as guanosine nucleotides. There was only a limited number of cytosine-containing nucleotides or uridine-containing nucleotides serving as adduct masses onto the peptide from all DEB-crosslinked *E. coli* samples that were FAIMS-analyzed. 11.7% of all CSMs correspond to crosslinked C adducts with 90% of them localizing the crosslinked amino acid, and only 10.9% of all CSMs were U adducts, 87% of them with identified crosslinked amino acid. The specific nucleotide preference is discussed in more detail below, but the observed precursor stabilization is most apparent. A nucleotides, followed by G, and C/U being least stabilized. Albeit some nucleotides are more likely than others to stabilize the precursor, reliable identification of the crosslinked amino acid is still possible in most cases.

10.1.3 Reliable identification of XLinking sites

Having depicted a crosslinking mechanism for the chemical ligation of the preferred guanosine nucleotide and a nucleophilic side chain, the proposed adduct masses from tables 17 (2IT), 18 (DEB), and 19 (NM) lead to the successful identification of the crosslinked moeities to the identified peptides. As such, those three tables contain the proper m/z shifts that were found for crosslinked ion fragments in MS2 spectra such as spectra shown in figure 19 (2IT), figure 17 (DEB), and figure 18 (NM). Having set the ppm tolerance level on MS1 to a stringent 7 ppm, the precursor mass is detected very accurately, allowing for accurate determination if a crosslinked RNA moeity plus chemical crosslinker fit additively to the peoptide mass. This ppm tolerance level is even more stringent than the commonly used 10 ppm in UV-crosslinking [3, 5, 10, 46, 102]. Moreover, the level of deviation of observed ion fragments in MS2 spectra to their theoretical fragment masses was set to 20 ppm in RNP_{xl} and NuXL, assuring that crosslinked fragment ions are annotated reliably as crosslinked fragments, akin to the methods used previously [3, 5, 10, 46, 102]. This two-level approach ensures high levels of confidence in the annotations by RNP_{xl} when the additional rules for manual validation listed in section 5.3.23 are met. The new NuXL software improves upon RNP_{xl} by featuring a stable overall FDR control that was stringently set to 0.01, ensuring that proposed CSMs are truly showing a chemically crosslinked peptide to RNA nucleotides. Ultimately, validation was achieved by scrutinizing crosslinking sites and cross-checking their biological function with known data, or proposing a biologically functional narrative for such sites that aligns with previous findings.

10.2 Chemical crosslinking coupled with mass spectrometry identifies RNA-interacting sites in protein-RNA complexes

To test if chemical crosslinking coupled with mass spectrometry is suitable to identify protein-RNA interactions reliably and unambiguously, various simple protein-RNA complexes were crosslinked *in vitro* to identify base level parameters such as frequencies of expected adduct masses, nucleotide preferences and amino acid preferences. Knowing these, it raises confidence in the approach and workflow, so it can be applied to more complex systems such as whole cells confidently *in vivo*.

10.2.1 Hsh49 complexed with synthetic RNAs yields highly specific nucleotide crosslinking events

After successful identification of crosslinking events between Hsh49 and synthetic RNA oligonucleotides by radioactive SDS-PAGE, the same two-component system was used to identify crosslinked peptides from Hsh49 by crosslinking mass spectrometry. The developed workflow was successfully used to generate spectra that unambiguously showed crosslinking events between Hsh49 and the synthetic RNAs for all chemical crosslinkers and all four different nucleotides at amino acid resolution. One critical drawback about these findings lies within the artificial nature of the experiments using a single protein and an RNA that is only built from nucleotides of one nucleobase. The important finding is thus not necessarily the identified crosslinking regions that generally align well between the two chemical crosslinkers DEB and NM, but rather the fact that identification of chemically crosslinked nucleotides was possible in the first place. This is not trivial since the adduct masses and compositions are unknown and have to fit to the observed shifts, depending on the nucleotide present. Only utilizing one different nucleobase as the modifying nucleotide at a time reduces complexity of the adduct masses and therefore allows for their decisive composition to be deduced. Similar to UV-crosslinking, any adducts exceeding two nucleotides of RNA, increase the computational complexity and consequently the rate of false positives and while there are some good examples for trinucleotides to be attached to a peptide following UV-crosslinking, chemical crosslinking usually fails to produce spectra with convincing trinucleotides attached to the peptide. Increasing the nucleotide length from two to three in $\mathrm{RNP}_{\mathrm{xl}}$ for example, results in about 5 times more potential CSMs, even though the number of possible combinations only doubles from 10 possible combinations for two nucleotides attached to the peptide to 20 combinations for three nucleotides. That increase in potential CSMs does not translate into actual true positives though, as >99% of those potential triple nucleotides are actually verifiable by the standards set above, meaning that the increase is due to random noise being fitted by an increase in potential combinations.

Analyzing obtained crosslinking sites from Hsh49 being crosslinked to synthetic RNAs, the wide distribution of crosslinking sites in both RRM1, RRM2 and the flexible linker region connecting the two domains, suggests a rather unhampered mode of crosslinking Hsh49 alone. While both RRM domains are canonical in structure, it was previously shown that RRM1 was mainly involved in RNA binding when complexed with the Cus1 protein [144, 205]. Hsh49 alone displayed rather promiscuous crosslinking potential to artificial RNAs, especially in RRM2 as depicted above. Structurally, RRM domains feature highly conserved aromatic residues of the anti-parallel β-sheets, mediating nucleotide contact, but the conserved aromatic residues in RRM1 of Hsh49

(Tyr-12, Tyr-52 and Phe-54) were not identified to be crosslinked either chemically or by UV-irradiation [144]. UV-crosslinking failed to yield any crosslinking sites in RRM1, and chemical crosslinking pointed at the same region (38-IKYP-41) N-terminal of the conserved aromatic residues. Crosslinking sites in RRM2 however, are closer to the conserved aromatic residues and generally greater in numbers. Here, the two conserved solvent-exposed aromatic residues, Phe-111 and Tyr-152 in are close to the chemically crosslinking sites of K-113 (NM), S-118 (DEB) and K-147 (DEB, NM). [144]. 2IT-crosslinking identified only crosslinking sites to lysine due to its mechanism of action that were distant to the conserved amino acids known to involved in RNA-binding and displays limited utility in only identifying lysine-uracil crosslinks. Interestingly, the third aromatic residue commonly found in RRM domains is replaced by Cys-150 in RRM2 of Hsh49. [144]. This Cys-150 revealed central importance in RNA-binding, as it was identified to be crosslinked by both UV-crosslinking and chemical crosslinking using DEB and NM. Therefore, one can conclude that both RRM domains bind RNA *in vitro* when Hsh49 is complexed with synthetic RNAs. Following the surprisingly high quantity of crosslinks detected in RRM2, Hsh49 was complexed with its cognate U2 RNA, as well as the Cus1 protein that is thought to confer RNA binding specificity to RRM1, to test if preferences in RRM domains are reversible under more native-like conditions.

Lastly, two preferences could be observed using *in vitro* crosslinking of protein-RNA complexes that fit previous findings from others. First, the exclusivity of 2IT to lysine amino acids that are chemically converted to thiols that react with uracil-containing nucleotides in a canonical UV-crosslinking fashion was already observed in 1995 by Urlaub *et al.*. [7]. Research by Tretyakova *et al.* using DEB alluded to a potential preference to guanonine nucleobases in DNA that could be confirmed for RNA using the synthetic RNA comprised of G and U dinucleotides [121]. This comparison is only relative to uridine and cytidine nucleotides, though, as adenine-containing dinucleotides RNAs were not investigated. Moreoer, it shows that the strong predilection for uracil nucleobases by canonical UV-crosslinking can be overcome by using chemical crosslinking. Chemical crosslinkers feature a G preference for DEB, and a more equally spaced distribution between U and G nucleoides for NM. The observed preference of DEB for guanosine nucleotides identified by mass spectrometry was also identified by others in DNA, as well as NM exhibiting a preference for guanosine-cysteine crosslinks in DNA [206]. Even though canonical UV-crosslinking strongly prefered uridine nucleotides in this study, a small portion of crosslinked nucleotides could be attributed to G or C which are also frequently observed in UV-crosslinking of DNA [10]. While these findings mirror results obtained for DNA crosslinking by others, absolute mass spectrometric proof for RNA samples as elaborated upon has been missing. Nonetheless, those preferences were obtained from an artificial system and may not reflect conditions met in nature. To circumvent this problem Hsh49 was complexed with Cus1 and its cognate U2 RNA in a very similar fashion to experiments performed by Van Roon *et al.* to investigate if preferences in RRM domain, amino acids, and nucleotides hold up to the previous findings.

10.2.2 Hsh49:Cus1 complexed with U2 RNAs

Analyses of the crosslinking sites obtained from Hsh49:Cus1 protein complex binding to either U2_47 RNA or U2_90 RNA revealed significantly more crosslinking sites within the RRM1 domain, probably attributing to Cus1 binding and providing an additional protein interface used in RNA binding in RRM1 [146]. Essential amino acid residues Tyr-52 or Phe-54 were not identified by UV-crosslinking that is generally thought to induce crosslinking events between tyrosines and RNA at minimal distances. However, Phe-57 was identified to be crosslinked instead in UV-crosslinked samples, targeting the crucial region of RRM1 involved in RNA binding. Both UV-crosslinked samples and NM-crosslinked complexes exhibit fewer crosslinking sites with the longer U2_90 RNA, raising the question if the second and third stemloop of the U2 RNA alter the binding behavior of the protein complex to the RNA. It is described that the shorter, one stemloop containing U2_47 RNA is sufficient for proper complex binding, also supported by the evidence provided above [146]. Additionally, the 5'-end of the RNA present in both RNAs used is thought to enhance binding of the complex to the RNAs in RRM1, a finding that might explain why the crosslinking sites in RRM1 generally did not vanish upon using the longer RNA [146].

Overall, though chemical crosslinking still identified most crosslinks within the RRM2 of Hsh49 in the case of DEB-mediated crosslinking, while almost all RRM2 crosslinks vanished upon Cus1 binding in the case of NM-crosslinking. Grouping NM-mediated crosslinks with UV-induced crosslinking sites is extended when scrutinizing the Cus1 crosslinks in the C-terminal region of the protein that is targeted by both approaches. Unfortunately, the protein sequence beyond the second α-helix of Cus1 is intrinisically disordered and could not be crystallized, allowing to posit its involvement in RNA-binding as shown here. Interactions within that region of the Cus1 proteins may be transient and highly dynamic, evading strong interacting effects that would have been detected by flourescence anisotopy experiments performed by Roon *et al.* [144]. Additionally, Arg-290 conferring a described increase in RNA-binding of Cus1 was not identified in the experiments described above, even though Arg-290 constitutes the N-terminus [146]. A possible explanation could be the mere size of the truncated protein, now being drastically shortened and thus more flexible to evade RNA contact. Lastly, DEB-derived crosslinking sites defy any logic imposed by UV-crosslinking sites or NM-crosslinking sites. Here, RRM2 displays multiple crosslinking events, especially for the three-stem loop containing U2_90 RNA that almost form a line across the planar anti-parallel β-sheets, evoking a sense of visual evidence where the RNA runs across the domain. Unfortunately, it is not described that RRM2 is responsible for RNA-binding, in fact its binding was found to be so transient not to be quantified by binding coefficients in flouresence anisotopy experiments alone [146]. Still, its sequence in RRM2 is close to consensus as described above, probably infering RNA-biniding activity in combination with RRM1 (and Cus1). Flouresence anisotropy data presented in this study indicates that Hsh49 alone can bind to poly-G_5U with a K_d of 2.4 µM. Binding may be largely driven by RRM1 in the absence of Cus1, but the data described above suggest that under the given conditions chemical crosslinking with DEB captures transient stages of RNA-binding of RRM2 that aggregate into the picture shown in figure 24.

Lastly, the two different approaches yielded different crosslinking amino acids as described. UV-crosslinking generally crosslinks most often through either aromatic amino acid residues such as Tyr, Phe, and Trp, that were also identified among the top3 crosslinking amino acids described here, or lysine residues [5, 10, 16]. Lysine crosslinks were strongly favored by DEB, attributing 91% of localizable crosslinks to be lysine induced, endorsing the idea of DEB favoring nucleophilic amino acid residues. Its nucleophilicity was already described for nucleophiles such as the N-7 position of purines found in DNA-interlinks, as well as bound to lysine of histone proteins [207] NM, however, exhibits a preference for histidines, followed by lysine and proline. Nucleophilicity was already described for NM, but proline seems to be incongruous to the nucleophilic narrative. Some DEB crosslinks were also identified to be proline-bearing, raising the question if there is a radical based mechanism that would allow for proline to be crosslinked via DEB or NM since DEB is naturally epoxic, and NM undergoes cyclopropane cyclization as an intermediate that is exhibiting high ring tension and high nucleophilicity due to the quaterrnary ammonium ion that is transiently formed, allowing it to undergo further reactions in aqueous solvents extremely quickly [208]. In any case, identified chemical crosslinking sites appear to be quite different from UV-identified sites which can be explained by the different preferences and chemistry described above, combining into an orthogonal crosslinking approach.

10.2.3 Modeling of Dnmt2 can be refined by additional chemical crosslinking positions and leads to an alternative conformation

Building upon this complementary approach of identifying different crosslinking sites, a previously UV-crosslinked complex obtained from the Ficner lab was crosslinked chemically using DEB and NM. As shown above, crosslinking sites differ dramatically in their numbers. While the original model was build to accommodate four crosslinking sites to uridine nucleotides (Lys-91, Trp-221, His-223, and Cys-303) from UV-irradiation a prodigious number of new crosslinking sites emerged, altering the original model by Johansson *et al.* in which the handle of the tRNA was facing the C-terminus and centered around the crosslinked Cys-303, which is proximal to a positively charged pocket formed by Arg-371, Lys-295, and Lys-367 [5]. However, the newly identified crosslinked regions in the co-factor binding domains, as well as nucleotide binding domains, spanning amino acids Lys-55 to C-142, cannot be explained by the original model by Johansson *et al.* since the handle of the tRNA faces away from the crosslinked region. In collaboration with Piotr Neumann from the Ficner lab, the new model was created to fit the new crosslinking sites. As described above, the new model build from ROSETTA was even initial favored when only the UV-crosslinking sites were known, but dismissed to accommodate the Lys-91 crosslinking site. This crosslinked site was retained as fitting the updated model well. Rotating the handle of the tRNA brings tRNA$^{\mathrm{Asp}}$ close to the novel crosslinking sites of basic amino acids such as His-110, His-56, His-50, and Lys-55 that would interact well with the negatively charged phosphate-ribose-backbone of the tRNA. Even so, the model does not accommodate input crosslinking data perfectly, as Trp-221 and His-223, fitting the original mode, are now located farther away from the tRNA, as well as Cys-72 and Cys-142 still not being well accommodated by the model. Discrepancies may still arise as RNA-binding generally describes a dynamic process with RNAs being able to change conformations for example. Similar to the reasoning behind discrepancies for the original model, there is the possibility that the active loop of the protein changes conformation upon substrate binding, reducing distances of crosslinking sites to the tRNA [5]. This induced fit is well described for many enzymes, including DNA polymerase, that also bind nucleotides [209]. Collectively speaking, the numerous crosslinking sites obtained from chemical crosslinking using DEB and NM could aid in redesigning a novel model to fit novel data and harmonize initial modeling efforts based on rigid-body modeling performed by ROSETTA.

10.2.4 Crosslinking sites from the NELF complex bound to TAR-RNA

Protein-protein crosslinking data using BS^3 revealed extensive crosslinking events between the four subunits of the NELF complex. This finding proves the assumption that the NELF complex is fully assembled and subunits are interacting with each other. Additionally, crosslinking sites between the subunits correlate well with findings reported by Vos *et al.* earlier, ensuring proper sample quality to identify crosslinking amino acids between the NELF complex and TAR-RNA [150]. The NELF complex is of crucial importance in promoter-proximal pausing by RNA polymerase II [151]. NELF forms a protein complex with the protein DRB sensitivity factor (DSIF) and RNA polymerase II that stabilizes the paused polymerase in the promoter-proximal region [151]. While the NELF complex exerts many functions crucial to pausing RNA pol II such as binding the polymerase funnel, or binding of the anti-pausing factor TFIIS, in addition to possessing two tentacle regions that can contact DSIF and exiting RNA [151]. The latter factor was addressed by chemical crosslinking of the NELF complex with its cognate RNA as shown above, proving that the tentacle regions make contact with the RNA at various points. The tentacle regions were defined by Vos *et al.* as the residues 189-528 for the 'NELF-A tentacle' and the residues 139-363 'NELF-E tentacle' [151]. As such, the tentacle regions are fairly long and highly disordered evading successful crystallization. The NELF-E tentacle mounts into a C-terminal RRM domain and is thought to extend to exiting RNA [151]. Previously performed UV-crosslinking experiments by Vos *et al.* already alluded to both tentacle regions to be involved in RNA binding, but UV-crosslinking restricted the crosslinking amino acids in this case to lysines [151]. Even though UV-crosslinking was also performed as described above, UV-induced crosslinking sites did not match sites obtained from the initial crosslinking experiments. Both sets of crosslinking sites fall within the unstructured region absent in the mode. This discrepancy is probably due to the different method of enrichment, as well as mass spectrometric data acquisition that was more extensive in the original UV-trail. Because of that, chemical crosslinking data provided in this study is referenced against data published by Vos *et al.*[151].

Chemical crosslinking coupled with mass spectrometry identified crosslinked amino acid residues in all four NELF subunits, some of which were identified through multiple crosslinking sites such as His-167 of NELF-C, but crosslinking sites in NELF-B and NELF-C may only be artifacts of weak *in vitro* binding to RNA as described by others [148]. The vast majority of crosslinking sites from chemical crosslinking experiments described above fall mainly within both tentacle regions of NELF-A and NELF-E, which are discussed here. While the original UV-experiments indicated nine lysine residues in the tentacle region of NELF-A to be crosslinked, here we also identify nine crosslinking sites, eight of which are different than the original lysine crosslinks [151]. Specifically, Lys-207 was identified in both trails, but chemical crosslinking extends the range of detectable crosslinked amino acids in this case to also include, Ser, His, and Arg besides Lys. His-198 was also identified by all three modes of crosslinking, indicating that complementary crosslinking methods can aggregate into a single amino acid. Similarly, chemical crosslinking sites from NELF-E includes amino acid residues such as Cys, His, and Ser besides Lys again. Interestingly, the only cysteine found in the RRM of NELF-E was identified to be crosslinked using DEB that may be bound by nascent RNA to help recruit NELF to pause sites by binding to RNA hairpin structures amply found near pause sites, while not necessarily required for pausing [151]. These data add to the previously identified crosslinking sites and paint a more vivid picture of multiple amino acid residues of both tentacle regions to be heavily involved in RNA binding, which is also supported by fluorescence anisotropy finding as explained above. Identifying Cys-300 within the NELF-E RRM and its posited function to guide NELF to pause sites upon binding nascent RNA may explain why RNA binding is impaired so heavily in NELF-E and even more so in both NELF-E/A tentacle deletion mutants. In that regard, chemical crosslinking coupled with mass spectrometry could provide additional information on crosslinking sites within the NELF complex that broadens understanding of their involvement in pausing at promoter-proximal regions.

10.3 Supervised machine learning approaches can aid in identifying CSMs from large data sets

The majority of data for this thesis was acquired at a time where the RNP$_{\mathrm{xl}}$ software was mainly used to identify CSMs steming from crosslinked proteins. As described above, manual validation can be quite a laborious process, even for a single set of experiments. Analyzing RNP$_{\mathrm{xl}}$ output features to find correlative variables did not provide a way to alleviate this problem by identifying a variable that could be used as filter for chemically crosslinked samples. As depicted in figures 34 and figure 35, the original RNP$_{\mathrm{xl}}$ score works well when filtering UV-crosslinked entries with a cut-off value of 0.15, immensely reducing the number of false positive entries. The same score for DEB-crosslinked samples fails to provide a cut-off value that could be used, probably due to the original RNP$_{\mathrm{xl}}$ algorithm being trained on UV-data only, which looks different from DEB-derived data. The best option for DEB-derived potential CSMs to be classified by a strict cut-off value would have been the total matched ion current score (MIC score) that would have led to a considerable loss in CSMs by at least 25%. Such drastic losses did not seem reasonable, as the overall number of CSMs is generally lower compared to UV-samples.

To address the problem of not finding a metric that can be used to distinguish TPs from TNs in subsequent supervised machine learning approaches, additional features were derived from RNP$_{\mathrm{xl}}$ output data. Most importantly, the length of the crosslinked peptide was extracted numerically and added as a potential feature that proved to be very valuable for algorithms building decision trees, as most crosslinked peptides had been generally short peptides in the past. Additional features such as pI, mw, or percentages of the peptide being charged/polar/non-polar were thought to be of greater importance than decision algorithms would later find them to be. This is probably due to the overall composition of the peptide being considered. Patches of polar or charged amino acids that generally like to crosslink chemically lose their mathematical weight (a high polarity score) over longer peptide distances of non-polar amino acids as the polarity index is dependent on the number of amino acids present in a peptide (the summed polarity weight is divided by the overall peptide length). Also as an example, an agglomerate of positively charged amino acids, predisposed for RNA binding, would not have been detected in any case, as trypsin would have cleaved the peptide at those positions. Even so, the novel NuXL algorithm computes peptide lengths, reporting them in the output table, possibly indicating that the algorithm utilizes this features when performing the search. Due to the lack of handling hundreds of thousands of spectra reasonably and time-efficiently from chemical crosslinking experiments, supervised machine learning approaches were sought out to guide the curation process. k-nearest neighbor, C5.0, and RIPPER were chosen to classify unknown spectral data that were augmented by additionally computed features.

10.3.1 The k-NN algorithm builds the most conservative classifying model to evaluate spectral data

k-NN is one of the simplest, yet also most effective, algorithms that does not make assumptions about the underlying data distribution. It is used frequently in classifying medical data, for suggesting a next song based on previously heard songs, and even in metabolomics to imputate missing values for metabolites [51, 52]. The k-NN algorithm treats the features as coordinates in a multidimensional feature space [52]. As the output data from $\mathrm{RNP}_{\mathrm{xl}}$ consists of 39 numeric features, the feature space would be an n-dimensional vector with n = 39. The classifying variables TPs and TNs, corresponding to true CSMs and False Negatives, cluster in the feauture space and a Euclidean distance can be calculated between an unclassified data point within the feature space and k of its nearest neighbors [51]. Surprisingly, k-NN classified unknown $\mathrm{RNP}_{\mathrm{xl}}$ FAIMS output data most conservatively, only returning a couple hundred of potentially true CSMs, whereas both C5.0 and RIPPER classified the vast majority of entries to be true CSMs. UV-data was classified particularly well using k-NN, performing with an overall predictive value of 0.95. The high predictive value can probably be attributed to an overall $\mathrm{RNP}_{\mathrm{xl}}$ score that could have been used to filter the data with strict cut-off value. In relations to that, the poor overall score for DEB-crosslinked samples could not be compensated in any dimension to values exceeding an overall predictive value of 0.84. This is probably due to the fact that the best discerning feature from correlating statistics was the MIC score, that could have been used as a cut-off value with a loss of 25% in CSMs.

The two decion-based algorithms C5.0 and RIPPER performed well on the test data set, amount to OPVs of 0.94 and 0.92, respectively. Nonetheless, both algorithms performed poorly on unseen FAIMS data, overestimating the number of TP CSMs greatly. A possible explanation could be that FAIMS data is substantially different to the algorithm, as it affects probably all features of RNP$_{\mathrm{xl}}$ in a way that is different from the the non-FAIMS data feeding into the training data set and testing data set, even though the acquisition methods and gradient settings were exactly the same. As described, even penalizing FPs did not diminish the number of potential CSMs significantly in C5.0 and both models were thus seen as unfit to classify novel data. Two rules used in building the models by C5.0 and RIPPER indicate that crosslinked peptides generally elude over the first 27 min of the gradient and contain fewer than 4 hydrophobic amino acids.

The combination of lazy learning algorithm (k-NN) and two decision-based algorithms should have ideally covered the problem from two different angles and resolved classification mishaps either by one algorithm or the other. Unfortunately, all three algorithms performed well on the test data set, but the two decision-based algorithms returned obscure results for unseen, unclassified data. k-NN seemed to be impervious to subtle changes in underlying data, as manual inspection of 100 k-NN classified spectra yielded a 79% success rate, which is slightly lower than the reported overall predictive value of 84% from the testing data set for DEB-crosslinked samples, but still reasonable to identify the majority of the most abundant proteins with true positive CSMs.

10.3.2 Drawbacks on k-NN and comparison to NuXL

Apart from the recent development of NuXL which now features strict FDR control as a potent filter to obtain true CSMs quickly, k-NN does inherently suffer from a pertinent drawback to this situation: k-NN does not actually provide a model that would explain input variables and relations to the class outcome [51]. It also requires choosing the proper number of neighbors (k) to function optimally, for which different numbers of k's were tested on the training data set to minimize the number of false positives. k-NN is considered a lazy learning algorithm because no abstraction occurs, allowing the algorithm to build a model that would detect relational patterns across the input features other than Euclidean distances that are technically lengths of vectors and not modeled relationships between variables [51]. Thus k-NN relies heavily on the training data set that is effectively stored as working memory to compare new incoming data points to their respective positions within the feature space. Additionally, k-NN did not perform as well as NuXl that is supposed to have a strict FDR control, set to 1% FDR to greatly reduce false positives, while k-NN classifies output data with 15-20% FDR.

KNN still overestimated the number of TP CSMs, especially for DEB-crosslinked samples. Here, the positive predictive value of 0.8 leads to an inflation of supposedly TP CSMs, by carrying a 20% error rate into classifying a TP that cannot be compensated by the negative predictive value of 0.87. An OPV 0.84 for DEB crosslinked samples does reflect poorly when thousands of spectra are to be classified and therefore leads to grossly overestimated CSMs following k-NN classification. Even UV-crosslinked samples are overestimated by 5%, which does not sound like a lot, but accumulates into unrealistically high values when the total number of spectra to be classified is also very high. Knowing that the error rate was higher for k-NN, supposedly crosslinked proteins were ranked according to the frequency of supposedly crosslinked CSMs. CSMs were aggregated into supposedly crosslinked proteins and top-10 protein lists were generated that were identical in about 76% of the cases between k-NN and NuXL, demonstrating that k-NN classifications can be used to identify the most abundantly crosslinked proteins successfully. However, once NuXL was made available, the complex data sets were searched again and spectral data were reclassified using NuXL to eliminate the 14-19% of potential error, ensuring minimal error and high confidence in the results.

10.4 Determining optimal crosslinking conditions for *in vivo* crosslinking of *E. coli* cells

Having proven that chemical crosslinking using DEB reliably identifies protein-RNA interacting sites in protein-RNA complexes reconstituted *in vitro*, treating live *E. coli cells* with DEB to identify crosslinking sites *in vivo* requires the crosslinking reagent to be taken up by the bacterial cells. As shown in figure 30, DEB is taken up by the cell even at 5 mM and incubation times of at least 60 min for visible "smearing" on an SDS-PAGE gel, that indicates crosslinking events that shift protein bands to higher molecular weights. However, a longer incubation time with a toxic crosslinking reagent may also prolong toxic exposure and thus alter the biochemical processes of live cells to respond to the chemical stressor. Shortening incubation times requires increasing DEB concentrations that seemed to have reached a turning point at 50 mM DEB. Here, even a 2 min incubation period shows some smearing on SDS-PAGE gels as depicted. Higher concentrations than 50 mM could correspond to a toxic overload that effectively represents a failed attempt of the bacterium to response to the stressor and were chosen because of that. Similarly, a concentration of 25 mM and 20 min incubation period was not chosen because 20 minutes is the minimal time required for *E. coli* to undergo binary fission and stressing the cell for an entire cell cycle may have resulted in a prolonged stress response that would have primarily been about repairing the DNA-damage DEB is known to induce [210]. Consequently, a concentration of 50 mM and an incubation time of 10 min was chosen to hopefully induce some tolerable stress that would crosslink RNA-binding proteins in their native conformation.

However, crosslinked proteins based on SDS-PAGE gels do not distinguish if the cells were still alive at the given times and DEB concentrations. In *E. coli* the SOS-response is induced after DNA-damage by the activation of the LexA repressor and recA [211]. The protease LexA binds to multiple promoter on various SOS-genes in unstressed cells, repressing gene transcription of SOS genes [211]. DNA damage activates RecA to exert proteolytic activity on LexA, initiating and enhancing auto-proteolysis of LexA to free the promoters of SOS-genes [211]. Transcribed genes, are involved in specific DNA repari mechanisms such as nucleotide excision repair (NER), tolerance of DNA damage, and delaying progression to the cell cycle [211]. The SOS-response thus activated mainly DNA-binding proteins, not specifically RNA-binding proteins other than proteins involved in transcription of SOS-genes. Consequently, a concentration of 50 mM and an incubation time of 10 min was chosen to hopefully induce some tolerable stress that would crosslink RNA-binding proteins in their native conformation.

The growth curve of *E. coli* at crosslinking concentrations of 50 mM DEB depicted above shows an incisive bacteriostatic effect that is eventually overcome at the 90 min mark post DEB exposure; meaning that it takes *E. coli* at least four normal cell cycles that are halted to repair the damage induced by DEB, as OD_{600} values remain stationary over that period, before slowly increasing again into exponential growth. Most importantly, 50 mM does not seem to be bacteriocidal so that identified protein-RNA crosslinking sites are more likely to reflect a cellular state that is "stressed" and not "dead". As such, identified crosslinking sites are likely to reflect conformations close to "native", limiting the possibility of detecting artefacts of the the treatment.

10.5 Piloting *in vivo* crosslinking of *E. coli* cells

The pilot experiments already provided reasonable data to conclude that chemical crosslinking of *E. coli* cells *in vivo* using DEB is a feasible approach to identify RNA-binding proteins. Subfractionated S30 and S100 samples yielded different numbers of total CSMs obtained by manual validation. UV-crosslinking clearly returns more CSMs, oftentimes the same crosslinked peptide being fragmented multiple times. The cold-shock proteins Csps were identified very often. Looking at the S30 UV-crosslinked sample for instance, the high number of 535 CSMs contains 231 CSMs that correspond to subunits of the cold-shock protein complex (124 CSMs corresponding to SCPC, 55 CSMs corresponding to CspA, and the remaining 52 CSMs corresponding to peptides that are shared between the Csp proteins). Identifying Csp proteins especially in UV-treated samples is not surprising, as UV-crosslinking is performed at 4 °C. The major cold-shock protein found in *E. coli* is CspA, binding RNA weakly at various positions to prevent RNA secondary structures formed at low temperatures [212]. CspA thus acts as a RNA chaperone allowing translation of various mRNAs at low temperatures. [212]. While RNA binding is low in sequence specificity and affinity, Phadtare *et al.* could identify an AT/AU-rich RNA consensus sequence of ssRNAs with the sequence AAAUUU for all four cold-shock proteins [212]. Speaking to the UV-bias, favoring the detection of uracil-containing nucleotides, exclusively finding uracil nucleobases attached to Csp peptides is of no surprise. Interestingly, cold-shock proteins were exclusively found in UV-treated samples, meaning that even though cold-shock proteins were originally discovered in response to low temperatures, their associated functions expand from a mere temperature response to a more generalized stress response that does not seem to be activated much upon DEB treatment. While no Csp proteins were identified in the pilot experiment, a FAIMS double CV run using CVs = -35/-45 V did identify a shared peptide between CspE and CspC with the sequence WFNESK to be crosslinked to a Uridine via DEB by a single CSM. Identifying only one CSM may sound meager, but if SCPs are mainly activated at low temperatures with some accessory functions at higher temperatures, identifying only one Csp in a crosslinked sample at 37 °C may not be surprising. Unfortunately, the corsslinking site could not be determined unambiguously, yet identifying this peptide may indicate Csp involvement in the response to DEB treatment.

Inspecting CSMs from DEB-crosslinked samples that are lower in numbers compared to UV-derived CSMs, their aggregate crosslinked protein count resembles that of UV-crosslinked proteins, probably by missing Csp proteins that were not majorly induced upon crosslinking. The overlap between crosslinked proteins is small (4%), identifying ribsomale proteins such as 50S ribosomale protein L2, 50S ribosomale protein L20, ranscription antitermination protein NusB, and Elongation factor G. 30S ribosomal proteins were identified in the DEB subsets of crosslinked proteins that is not shared with UV-identified proteins in this pilot experiment. Investigating the crosslinking site from 30S ribosomal protein S7, its proximity to the 16S rRNA is evident and binding to 16S rRNA has been described through crosslinking experiments by Urlaub *et al.* at Lys-75 years ago [7]. Structurally, S7 is an essential protein of the head domain of the small ribosomal subunit, found close to the decoding center of the, where it contacts the mRNA [7]. In contrast to the original Lys-75 identified previously, DEB-crosslinking identifies the peptide 132-GTAVKK-137, crosslinked at Lys-136 to a cytidine nucleotide, which fits the ribosmal mode very closely.

50S ribosomal protein L2 was identified to be crosslinked by both approaches with DEB crosslinking pinpointing the crosslinking amino acid to His-53 from the crosslinked peptide 48-HIGGGHK-54. This crosslinking site is was previously unknown, with UV-derived findings locating Lys-67 by 2IT-crosslinking or Met-200 by UV-crosslinking from data acquired by Urlaub *et al.* years ago [7]. Still, the novel crosslinking site fits the crystal structure very well with the histidine side chain facing the 23S rRNA that is in close proximity. L2 constitutes an integral protein of the 70S ribosome, required for ribosome assembly and tRNA binding, that was not found to be crosslinked in this pilot experiment, but in later experiments using FAIMS [7].

Lastly, transcription antitermination protein NusB was found to be crosslinked via DEB. NusB plays an important role in transcription antitermination and rRNA transcription at ribosomal RNA operons. [191]. To achieve this, NusB binds to the boxA antiterminator sequence of the operons with enhanced affinity towards them if 30S ribosomal protein S10 is present [191]. Here, S10 and NusB form a protein complex as part of the processive rRNA transcription and antitermination complex (rrnTAC) [191]. DEB-crosslinking identifies Ser-113 present in the peptide 107-SFGAEDSHK-115 to be crosslinked to RNA. A slightly longer peptide was previously identified to be UV-crosslinked, where the peptide 113-SFGAEDSHKFVNGVLDK-129 was crosslinked to a UAC, probably at positions 7-9 of the λBoxA RNA used in their experiment [213]. This time, however, the crosslinking nucleotide was found to be a guanosine nucleotide that would be found at position 2 of the rrn BoxA with the sequence 5'-UGCUCUUAAACA-3' [191]. Together, these findings suggest that both approaches identify RNA-binding proteins in a mostly complementary fashion and that DEB-mediated crosslinking identifies sites that elaborate comprehensively on previous findings.

10.5.1 HCD targeted approach identifies RNA-binding proteins from *E. coli* cell lysates

Data obtained from analyzing DEB-crosslinked *E. coli* cell lysates using the MI-centric approach yielded numerous crosslinked proteins (208 in total from both UV-crosslinking and DEB-mediated crosslinking) with 30 proteins being identified by both approaches. Differences between UV-crosslinked samples and DEB-crosslinked samples appear right at the beginning, with a substantially larger number of CSMs being identified by UV-crosslinking in both S30 whole cell and ribosome depleted S100 fractionated samples. Aggregated proteins share about 10% overlap between the two different modes of crosslinking, routinely observed in this thesis, with 103 uniquely UV-crosslinked proteins, and 75 uniquely DEB-crosslinked proteins. The following two paragraphs highlight the biological relevancy of the two examples for each mode of crosslinking as described above, unequivocally proving the merit of utilising a MI-centric approach in crosslinking coupled with mass spectrometry to identify RNA-binding proteins.

Three crosslinked residues from 50S ribosomal protein L11 that were identified are in close proximity to the adjacent 23S rRNA. The peptide 114-AADMTGADIEAMTR-127 was found to be crosslinked at the three neighboring DMT residues Asp-116, Meth-117, Thr-118 at the end of an extended α-helix (Asp-116) that forms a turn before continuing as a second α-helix from position 120 onward. In the structure, 50S ribosomal protein L10 is also binding the 23S rRNA and forms with L11 the base of the ribosomal stalk important in binding translational GTPases that allow translation to occur at high speeds [214]. However, the crosslinking sites are accessible to the 23S rRNA as the crosslinking region lies within the two α-helices, fitting perfectly into the lower the nucleobases of dsRNA that exert a slightly tilted opened conformation in the mode. The second example for UV-identified crosslinking sites from the MI-centric approach involves 30S ribosomal protein S7 that was again identified. Crosslinked amino acids are located in close proximity to the adjacent RNA and are stemming from two different peptides. Peptide 150-AFAHYR-155 was found to be crosslinked at the two neighboring H-153 and Y-154 residues and peptide 79-VGGSTYQVPVEVR-92 was found to be crosslinked at the Tyr-85 that. Even though Glu-tRNA is making contact with S7, the contact site for the tRNA is on a different side of the protein and all identified crosslinks are located on surfaces touching the 16S rRNA, possibly serving as structural attaching points within the ribonucleoprotein.

50S ribosomal protein L1 was identified to be crosslinked to the 23S rRNA at three positions in close proximity to rRNA. The peptide 106-GEMNFDVVIASPDAMR-121 was found to be crosslinked at the neighboring EMN residues to G and U nucleotides. The crosslinked position on the 23S rRNA is close to the 3'-end, where L1 is part of the ribosomal stalk [194, 215]. This stalk is described to be movable by approximately 20 Å when deacylated tRNA moves in and out of the E site of the large subunit [215]. While the model used to investigate the exact positions of crosslinking sites does include Glu-tRNA binding to L1, the crosslinking site again faces the 23S rRNA on the dorsal side of the large subunit, probably anchoring the protein to its rRNA component. The slightly bent RNA helix in that region seems to fit the protein structure very well, suggesting that the region exhibits a more structural function of the ribonucleoprotein. Another DEB-mediated crosslinked protein identified by the MI-centric approach was 30S ribosomal protein S3. S3 was found to be crosslinked to a guanosine nucleotide of 16S rRNA via DEB at position His-176. The histidine-containing peptide 174-VPLHTLR-179 lies C-terminally from two crosslinking sites previously identified: Lys-44 and Lys-88 [7]. The identified crosslinking site seems to be centered around an region of the protein circumscribed by 16S rRNA, possibly anchoring it to the 16S rRNA. Interestingly, the opposite site of S3 is known to be involved in mRNA unwinding by assembling into a processivity clamp with ribosomal protein S4 that form a tunnel for the mRNA to pass through to the ribosome, acting as an mRNA helicase [67]. While S4 is also directly bound to the 16S rRNA, as indicated by crosslinking sites involving the peptide 34-IEQAPGQHGAR-44 at His-41 (not shown in figure 39, no crosslinking sites away from 16S rRNA could be identified in either S3 or S4 to hint at mRNA being unwound by the ribosome. The crosslinking sites support the idea of an anchoring point within the 16S rRNA for both S3 and S4 on the same medial proximal site of the 30S subunit.

10.5.2 FAIMS poses a potent strategy to identify RNA-crosslinked proteins in *E. coli*

Applications of ion mobility mass spectrometry to enhance peptide coverage has been around for years. FAIMS has been employed in a variety of different experiments to enhance peptide coverage by increasing the number of unique peptides, and thus, their respective protein [34, 216]. Applications of FAIMS to crosslinking mass spectrometry have been limited to protein-protein crosslinking up to date, but FAIMS separation has been proven to contribute greatly in detecting crosslinked peptides from protein-protein crosslinked complexes [193]. While peptide coverage is most crucial in most FAIMS applications, crosslinked peptide-RNA heteroconjugates display a different behavior across LC-separation and mass spectrometric detection, warranting a closer inspection at different FAIMS settings that play a diagnostic role in detecting peptide-RNA crosslinks. Most notably, detection of nucleotide marker ions in MS2 scans can indicate the presence of a CSM and therefore deserve proper investigation. Unfortunately, a contaminating immonium ion from Tyr is close to the detectable mass of marker ion adenine, making FAIMS separation of the two ion species desirable. Additionally, linear peptides are not of interest and should be avoided if possible, specifying FAIMS settings to reduce their detection. Current literature cites CVs of -60V as an ideal candidate for maximizing the number of unique peptides from a single CV run, albeit the number of corresponding proteins fairs equally well with lower CVs down to -45V according to others [35].

When investigating different runs for the presence of diagnostic marker ions for andenine and guanosine at various CVs, only the double FAIMS run at CVs of -40/-50V yielded well-separated XICs for the two marker ions compared to the Tyr immonium ion. Higher CVs cumulating in -60V and -70V narrow the elution window to guanine marker ions to the first 35 min of the gradient, coinciding with marker ions from adenine that extend up to 45 min. More importantly, the IM-Tyr is not separated from the two marker ions and coelutes within the same regions, indicating poor FAIMS-seperation. Elution of crosslinked peptides at the beginning of the gradient speaks to the hydrophilicity of the conjugates eluting easily due to weak retention to the C18 resion. Interestingly, appearance of the second adenine peak around 70 min in

CVs = -40V/-50V indicate a second ion species that is substantially more hydrophobic than the RNA-containing heteroconjugates eluting at the beginning, but also carries some adenine-containing moeity. It can be speculated that this peak may derive from ATP-bound peptides that are generally longer and thus more hydrophobic, but further investigations have to confirm that in replicate experiments that are lacking here.

Investigating at which CVs the number of PSMs is most reduced required identifying which PTMs are most prevalent in *E. coli* derived peptides, and placing methionine oxidation at the top of the list is not uncommon [35, 193]. Other studies have reported N-terminal acetylation of peptides, which necessitated including it as a variable modification to increase coverage, but the data presented here does not identify N-terminal acetylation to be of major concern [193, 216]. Also, semitryptic peptides, resulting from missed cleavages, are common in crosslinked samples and seem to align in prevalence in comparison to protein-protein crosslinking observed by others [193]. There is, however, a discrepancy in regards to the number of linear peptides emerging from the enriched samples presented in this thesis. Others have reported the number of unique peptides to be maximal at high CVs of -60V or -70V, whereas the here, the number of linear peptides was highest in double and triple runs with CVs using -40V [216]. This is likely be due to the enrichment process, as others did not enrich for certain peptides using TiO_2, but it is intriguing that the enrichment process must exert such a prominent effect as others have reported a drop in linear peptide identifications of about 29% at CV = -40V [35]. Analogously, the acclaimed CVs of -60V/-70V most conducive to linear peptide identifications by others were found to be reversed in the data presented here, as both double and tripe runs using that pair of CVs returned the lowest number of linear peptides [35]. However, the increase in linear peptide identification by including an additional CV in a triple FAIMS run is congruent with data reported by others and speaks to highly specific separability of ions using FAIMS [35].

Comparing the different double CV FAIMS runs horizontally between UV-irradiation and DEB-treatment, an overlap of 10% repeatedly emerges from the data, indicating high levels of complementary, where both approaches add the vast majority of identified crosslinks from unique identifications if combined. Interestingly, UV-irradiation is always outperforming DEB in terms of absolute numbers of CSMs. However, aggregated crosslinked proteins fall within a comparable range, but UV-treated samples yield slightly better yields of identified crosslinked proteins. No-FAIMS runs trumps any individual FAIMS double CV run in terms of absolute numbers of CSMs and aggregated crosslinked proteins. Since FAIMS is used as a filtering device, it comes at no surprise that individual FAIMS runs, even at double CVs, cannot perform better than a single-shot, unfiltered run. However, combining multiple FAIMS runs alleviates this problem and even surpasses no-FAIMS runs in crosslinking depth. Similar to reports from other researches, investigating linear peptides in proteomic experiments striving to enhance proteome coverage in different organisms, a pair of CVs = -60V/-70V UV results in a high number of identifications [35]. Here, the sharp peak in CSMs observed at a double CV FAIMS run at -60V/-70V contains multiple CSMs for individual cold-shock proteins (CSPs), that aggregate into four cold-shock proteins. The highest count of DEB CSMs at -40V/-50V even surpasses the no-FAIMs run in this case, but lack of replicate measurements mean a higher degree of uncertainty in making this comparison, when all other double FAIMS runs return lower numbers of CSMs compared to the no-FAIMS run. On the protein level, the double CV FAIMS run using CVs = -40/-50V is best for both UV and DEB, as aggregated crosslinked proteins are both highest. In all double CV FAIMS runs, the S30 sample generally yielded more CSMs and crosslinked proteins than S100 samples, indicating that subfractionation of whole-cell lysates (S30) into a more ribosome-depleted sample (S100) works adequately well in reducing sample complexity, yet some ribosomal proteins are still identified from S100 samples.

CVs that differ slightly in counter voltages were identified to share more aggregated proteins than distant CVs for both S30 and S100 samples. Effects can be attributed to the principle of the FAIMS technique, applying different voltages that result in different electric fields. A larger difference in applied voltages results in two electric fields that are more different than fields generated by more similar voltages. The ion filtering effect is consequently greatest when vastly different CVs are applied, as also noted by others [34, 35]. Doubly charged precursors are highest at no FAIMS or low FAIMS settings, confirming studies with both linear peptides and protein-protein crosslinked peptides, peaking at CV = -40V for doubly charged precursors [35, 193]. Triply charged precursors peak at CVs = -60V/-70V, which are the highest CVs investigated in this study. Again, both linear peptides and protein-protein crosslinked peptides peak at high CVs for triply charged precursors (CV = -80V) as observed by others [216, 217]. Interestingly, most crosslinked proteins were identified at -40V/-50V from precursors of charge state +4, corroborated by multiple CSMs for DEB-crosslinked samples. These findings support crosslinked identifications from protein-protein crosslinked using the chemical BS3, analyzed by similar FAIMS settings [217].

For UV-crosslinked samples, slightly higher CVs of -55V/-65V seemed to be more optimal to derived crosslinked proteins from quadruply or pentuply charged precursors, derived from copious CSMs. Slightly higher CVs also favor higher charged precursor ions reported by others in both protein-protein crosslinked samples, as well as linear peptides [216, 217]. Uniquely to nucleotide crosslinking, however, the NT adduct length also peaks at -40V/-50V in both CSMs and crosslinked proteins with G/U nucleotides peaking at -40V/-50V that has not been reported previously. Additionally, U nucleotides feature a sharp peak at -60/-70V crosslinked to higher charged precursor ions. Most importantly, identified crosslinked proteins from FAIMS runs encompasses all crosslinked proteins identified from non-FAIMS runs, displaying both sensitivity and reproducibility of the approach to conventional single-shot DDA MS2.

Investigating the triple CV FAIMS runs, a general consensus of UV outperforming DEB on CSMs and crosslinked protein level may reflect the search algorithm's predilection for UV-derived cross-linked spectra, as there were more UV-derived CSMs feeding into the training dataset of the NuXL algorithm than there were DEB-mediated CSMs. While more crosslinked proteins from the S30 samples, reflecting the whole-cell lysate, than the ribosome-depleted S100 sample is consistent with biological reasoning, identifying generally more unique crosslinked proteins with either approach (overlap only 5%), seems puzzling. Such stark discrepancy was not observed in the double CV runs and may be solely due to sampling size since the triple CV run using CVs = -40V/-50V/-60V amagamates a 17% overlap, resulting from a higher overlap in the S30 UV and S100 DEB samples. Overall, this single-shot preliminary pilot experiment lacks sufficient replicates to definitely conclude that triple CV runs result in a lower overlap than double CV runs, albeit it would make sense: having a third CV applied to the precursor ions results in greater filtering effects than only having two CVs, selecting more uniquely for certain ions, as demonstrated by others [216].

In accordance with previous findings for protein-protein crosslinking, the crosslinked protein distribution across the triple FAIMS runs is fairly homogenous, with CVs -30V/-50V/-70V peaking in CSMs because of the high -70V contribution [216]. A greater collapse into crosslinked proteins was also observed in protein-protein crosslinked peptides analyzed by FAIMS [216]. Interestingly, the triple CV runs using CVs = -35V/-45V/-55V seems to be most unique in its identifications, even though the counter voltages are fairly close to each other, resulting a more inclusive filtering effect of similarly behaving precursor ions by FAIMS compared to radically different CVs that would select for ions that behave more differently in the electric field. The strong preference of double charged ions at lower CVs from UV-crosslinked samples could not only be identified in the double CV runs mentioned above, but also by others [216]. Analogous to finding from protein-protein crosslinking experiments from others, the aggregate number of crosslinked proteins is lower, indicating multiple CSMs per crosslinked protein [216]. Interestingly, the highest detectable precursor ions from triple FAIMS runs is triply charged and highest in numbers of CSMs for -30V/-50V/-70V. Different from studies using protein-protein crosslinked samples, nucleotide crosslinked heteroconjugates were not identified from precursor ions of charge states +4 or +5 in any triple CV run [216]. In regards to the nature of the attached nucleotide, two NTs added are more easily identified at lower CVs -35V/-45V/-55V that also seems to be a penchant for adenine-containing nucleotides. Since a triple run comprising of CVs = -40V/-50V/-60V was generally best for all nucleotides combined on both CSM and protein level, the narrowed distribution around -50V seems to be particularly suited to DEB-crosslinked nucleotides that was also identified for double CV runs to be best suited.

In conclusion, FAIMS can greatly enhance identification rates of protein-nucleotide heteroconjugates across different settings with seeming specificity to certain aspects such as nucleotide adducts and crosslinker used. However, since this study is preliminary and lacking sufficient replicates, finding have to be considered carefully and re-evaluated with additional data. Still, comparisons to findings from others in regards to protein-protein crosslinking discussed similarities that allude to a genuine effect described in this study, allowing for a generalized recommendation of using FAIMS in either double of triple runs centered around a CV of -50V with the chemical crosslinking reagent DEB. Additionally, higher charged precursors above +3 were absent in all triple runs (n = 4), so some careful judgement on not using triple CV runs when higher charged peptides-nucleotide precursors are expected.

10.5.3 Chemical crosslinking sites identified in FAIMs analyses from *E. coli*

The STRING network illustrated above depicts how the identified crosslinked proteins are known to interact with each other, but it is more difficult to decipher biological processes from a STRING network. Identified biological processes are depicted in figure 58 and show how diverse processes featuring protein-nucleotide interactions can be detected using chemical crosslinking coupled with mass spectrometry. Apart from the obvious RNA-featuring processes such as tRNA synthesis, mRNA processing, Transcription, and ribosome assembly, other more metabolic processes emerge from crosslinked proteins. Interestingly, two functionally cell wall associated proteins were identified to be crosslinked to nucleotides that allude to the toxicity of DEB, as well as crosslinker import into the bacterial cell.

Cell wall protein MurA was crosslinked at Met-366 (359-LSGAQVMATDLR-371) to a uracil nucleotide. MurA adds enolpyruvyl to UDP-N-acetylglucosamine, forming UDP-N-acetyl-3-O-(1-carboxyvinyl)-α-D-glucosamine as the first step in peptidoglycan synthesis of the bacterial cell wall inside the cytoplasm [218]. Although, the UDP-binding site is described to end at amino acid 327 N-terminal of the identified crosslinking site, the same peptide with the same crosslinked amino acid was identified in two different samples independently, alluding to a potential transient interaction that was captured via DEB-crosslinking. Intriguingly, the epoxide antibiotic fosfomycin carrying one epoxide group instead of two found in DEB is described to inhibit MurA, leading to the bactericidal effect of fosfomycin [218]. In addition to that, the peptide 319-TNLQEAQPLGNGK-231 from outer membrane protein F (OmpF) was identified to be crosslinked, but no crosslinking site could be determined. Even so, having identified a crosslink within this protein could potentially illustrate how DEB enters the bacterial cell, as fosfomycin is also taken up by the cell via the OmpF protein [219]. The peptide in question lies on the extracellular side of the transmembrane protein that forms a channel through the plasma membrane [219]. Both OmpF and MurA demonstrate how epoxides potentially enter the bacterial cell and then act to inhibit the same nucleotide binding enzyme without the epoxide groups immediately being hydrolized and losing activity.

As noted above, even DEB treatment elicits a probable cold-shock response, albeit substantially fewer cold-shock proteins were identified in DEB-treated samples. This is surprising since cold-shock proteins act as RNA chaperones, stabilizing mRNAs in conformations most stable for active translation in cold adaptation, but involvement in optimal growth under benign conditions have also been reported [212, 220]. It may be that DEB-treatment signals a noxious environment that also warrants CSP involvement in stabilizing mRNAs that would also be favorably translated during cold adaptation. In contrast, UV-treatment performed at 4 °C for 10 min results in many CSMs from various different CSPs to be identified, roughly reflecting the increased number of CSP proteins being expresses at low temperatures. Other studies have found that CSPA in particular can make up 10% of the total protein synthesis during cold adaptation, explaining its prominent CSM count in UV-treated samples [66]. Moreover, DNA-dependant RNA-polymerase was also identified by crosslinking sites within the four subunits RPOA, RPOB, RPOC, and RPOD, illustrating that transcriptional processes can also be mapped by chemical crosslinking coupled with mass spectrometry.

Perhaps one of the most studied non-canonical RNA-binding moonlighting enzyme, glyceraldehyde-3-phosphate dehydrogenase (GAPDH), could also be identified in this study using DEB-crosslinking coupled with mass spectrometry. Originally identified as a glycolytic enzyme as part of the Embden–Meyerhof–Parnas (EMP) pathway, GAPDH converts glyceraldehyde-3-phosphate to glycerate-1,3-bisphosphate. As a moonlighting enyzme, though, GAPDH is known to be binding AU-rich regions without preferring a known RNA-binding motif, involved in translocation of tRNA and regulation of cellular mRNA stability and translation [221]. Its RNA-binding properties could be confirmed both by UV-crosslinking in this study as well as others, and corroborated using chemical crosslinking, all directing at similar regions of the protein [221]. Another katabolic enyzme identified through its cofactor binding is succinyl-CoA synthetase that is used in substrate level phosphorylation of ADP to ATP via the hydrolysis of succinyl-CoA to succinate [222]. Other metabolic enzymes of the glyoxylate pathway, the pentose phosphate pathway, as well as the electron transport chain are most probably identified through their nucleotide-containing cofactors such as FAD ad ATP, and not through previously unknown moonlighting functions involving RNA-binding. The metabolic enzymes in question include aceE, aceF, lpd, dapD, gltA, sucB, sucC, sucD, pflB, aspC, adhE, gapA, zwf, pta, glyA, fbaA, pgk, tdcE, rpe, yihX, and tpiA of which GAPDH (gapA) is best described in literature to exert alternative RNA-centric functions.

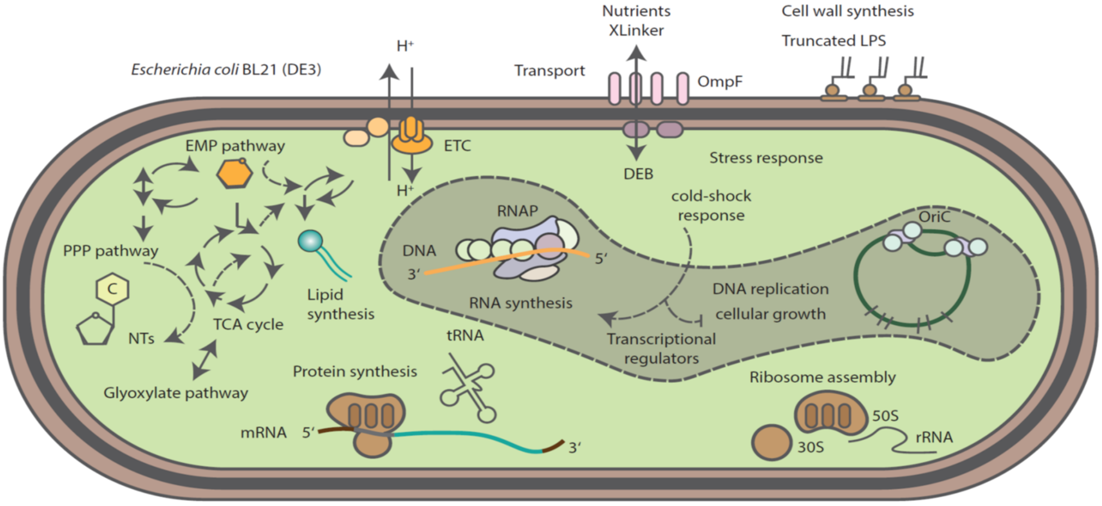


Figure 58: Cellular processes identified by chemical crosslinking. Different biological processes identified through chemical crosslinking of nucleotides to proteins via DEB. *E. coli* BL21DE3 cells are probably importing DEB via OmpF, crossing the plasma membrane and then inhibiting peptidoglycan synthesis. Cells growing in complete LB medium exert an increased catabolic metabolism, supplying energy for cellular development and growth through the EMP (Embden–Meyerhof–Parnas pathway), citric acid cycle (TCA) and pentose phosphate pathway (PPP). Anabolic processes such as elevated protein synthesis is demonstrated by multiple crosslinking sites identified in ribosomal proteins, as well as accessory proteins required for translation. Additionally, DNA-dependant RNA-polymerase illustrates the process of transcription that is also reflected in the crosslinking sites.

At last, the ribosomal proteins described above constitute the basis of solid evidence supporting RNA-binding *in vivo*. Crosslinking sites from ribosomal proteins S6, L36, L19 and L39 all cluster around regions that are in direct vicinity of the rRNAs in models from crystal structures or from cryo-electron microscopy as outlined above, validating the approach biologically. Active protein biosynthesis involving increased ribosomal translation makes sense under the given conditions growing the bacterial cells in complete LB-medium, being rich in carbohydrates and amino acids from yeast extracts and casein peptones that would support increased catabolism of glycogenic sources in addition to increased anabolism of proteins. Overall, the largest group of crosslinked proteins, 21 RNA-crosslinked ribosomal proteins speak of a bias towards ribosomal proteins that may reflect the benevolent growth conditions. Elongation factor EFTu2 and tRNA pseudouridine synthase A serve as two additional examples of canonical RNA-binding proteins identified by this approach as described above and can also indirectly allude to cellular growth as increased translation by EFtu2 and tRNA charging is required for protein biosynthesis.

10.5.4 Extending chemical crosslinking coupled with mass spectrometry to identify RNA-binding proteins in *B. subtilis*

Following extensive analyses of crosslinking sites using various methods in *E. coli*, the same workflow was applied to *B. subtilis* to prove that RNA-binding proteins can also be identified by chemical crosslinking *in vivo*. The STRING network shows the main cluster around the ribosomal proteins of both large and small subunit. 50S ribosomal proteins L28, L13, L36, L2, and L15 were identified to be crosslinked via DEB in addition to 30S ribosomal proteins S2, S3, S4, S6, S18, and S19. From that cluster, the crosslinking site of 50S ribosomal protein L28 being crosslinked at Cys-5 of the peptide 5-CVITGK-10 via DEB to an adenonsine nucleotide serves as prime example of Adenosine crosslinks that fit the biological model closely. Additionally, 50S ribosomal protein L13 crosslinked at His-81 to either G or U nucleotides found in the peptide 78-HTQhPGGLK-86, fits the narrative of chemical crosslinking identifying rRNA interacting proteins reliably as shown. Both L13 and L28 are part of the large ribosomal subunit, binding rRNA, but interacting amino acids or rRNA binding regions of L13 and L28 have not been identified in previous studies yet [223, 224].

Interestingly, three tRNA ligases from the second STRING cluster were identified to be chemically crosslinked via DEB, showing that not only rRNA is identified by chemical crosslinking, but also tRNA. While most functional data is inferred from phylogenetic homologs identified in *E. coli* and computational calculations based on similarity, the basic biological function of catalyzing the attachment of Trp, Thr, and Val by Tryptophan-tRNA ligase (TrpS), Threonine-tRNA ligase I (ThrS), and Valine-tRNA ligase (ValS), respectively, has been shown by others [225–227]. In TrpS, the peptide 183-IMSLNDPLK-191 was found to be crosslinked at Asp-198 via DEB to U nucleotides. The peptide falls within a few amino acids short of the characterized KMSKS-region of TrpS, ranging from amino acids 193-197, that is initiates tRNA binding of the protein [228]. Allowing for a little more room spatially to accommodate the crosslinker as well as dynamic flexibility of both tRAN and protein, the identified crosslinking site extending the KMSKS-region makes sense. Additionally, Thrs was found to be crosslinked at either S-371 or S-375 in the crosslinked peptide 368-YEMSGALSGLQR-379 to U nucleotides, indicating which regions of the protein are interacting with the tRNA in this case where a crystal structure of the protein with corresponding tRNA is missing. The catalytically active region of ThrS, ranging from amino acids 245 – 542, was not identified in this study. In the active site the 3'AMP residue of threonyl-tRNA is charged with threonine to give threonyl-adenylyl-tRNA under concurrent ATP hydrolysis [226]. Consequently, there are RNA residues in direct proximity to amino acids that could be crosslinked, but failing to identify crosslinks may be due to the lack of replicates and/or analytical depth that could be increased by additional fractionation of the sample [226]. Lastly, ValS was found to be crosslinked at either at T-618 or T-623 via DEB to C nucleotidies (crosslinked peptide: 615-LNETIEHVTQLADR-628). ValS also exhibits a KMSKS-region, ranging from amino acids 525-529, that was not identified in this study as the crossliked peptide is located C-terminally of the KMSKS-region [225].

Inosine-5'-monophosphate dehydrogenase guaB was found to be crosslinked at Cys-308 (crosslinked peptide: 299-VGIGPGSICTTR-311) via DEB to G nucleotides. GuaB catalyzes the first step in the *de novo* synthesis of guanine nucleotides by converting inosine 5'-phosphate (IMP) to xanthosine 5'-phosphate (XMP). [229]. More importantly, DEB crosslinking identified the catalytic cysteine side chain that is bound to the intermediate during catalysis [229]. Unfortunately, there is no crystal structure available for *B. subtilis* guaB, nor its homolog IMDH found in *E. coli* that was also found to be chemically crosslinked using various measuring methods. However, an IMDH crystal structure from *Mycobacterium thermoresistibile* bound to purine-riboside-5'-monophosphate could be used instead [230]. The sequence coverage of the active site is nearly identical to the sequence found in *B. subtilis* with the crosslinked peptide from *B. subtilis* 299-VGIGPGSICTTR-311 being altered in one single amino acid substitution from I-301 to V-301 in *M. thermoresistibile* (193-VGVGPGSICTTR-204) [230]. Since the sequences share high levels of similarity, the crosslinking site was mapped as depicted in figure 59. The Cys-201, corresponding to Cys-308 in *B. subtilis*, contacts the purine-riboside-5'-phosphate, and IMDH displays an wide active center that is opened enough for other nucleotides to diffuse into [230]. The authors of that structure claim to have modeled IMDH with the inhibitor CPR, 6-Chloropurine-riboside-5'-monophosphate, but the 6'-Position lacks the chlorine atom in the model [230]. Still, the structure suggests that there is enough space for a purine nucleotide to fit the active site, which would accommodate identified DEB-crosslinked guanosine nucleotides at the catalytic cysteine in both *E. coli* and *B. subtilis.* Since guanosine monophosphate ($C_{10}H_{13}N_5O_5$) only differs by -NH in comparison to the product xanthosine monophosphate ($C_{10}H_{12}N_4O_5$) accidentally fitting a GMP into the active site would be possible. Adding to a possible promiscuity in nucleotides entering the active site, uridine monophosphate was also found to be crosslinked via DEB to

IMDH in both *E. coli* and *B. subtilis.* Adduct masses comprised of intact G and U nucleotides without any neutral losses that bring a GMP closer to XMP (only missing a proton). This begs the question of if IMDH does bind RNA (transiently) or if DEB-crosslinking detects other nucleotdie-containing cofactors in this instance similar to the GTP cofactor identified in EFTu2 described above for *E. coli.* Others have reported that substitutions of the purine ring to include 6-Chloropurine-riboside-5'-monophospate, 6-thio-IMP, 2'-Deoxy-IMP and arabinose-IMP do not interfere with active site binding [231]. Additionally, it was reported by Hestrom *et al.* that even dinucleotides can fit the active center of IMDH such as they report acetylpyridine adenine dinucleotide, thionicotinamide adenine dinucleotide, 3-pyridinealdehyde adenine dinucleotide, nicotinamide hypoxanthine dinucleotide and nicotinamide guanine dinucleotide to be sufficiently binding to the active site of IMDH [231]. These findings by others shift the focus of transient RNA-binding of IMDH away to an enzyme with broad binding specificity to any purine-containing compound that happens to be detectable by DEB-mediated crosslinking coupled with mass spectrometry.

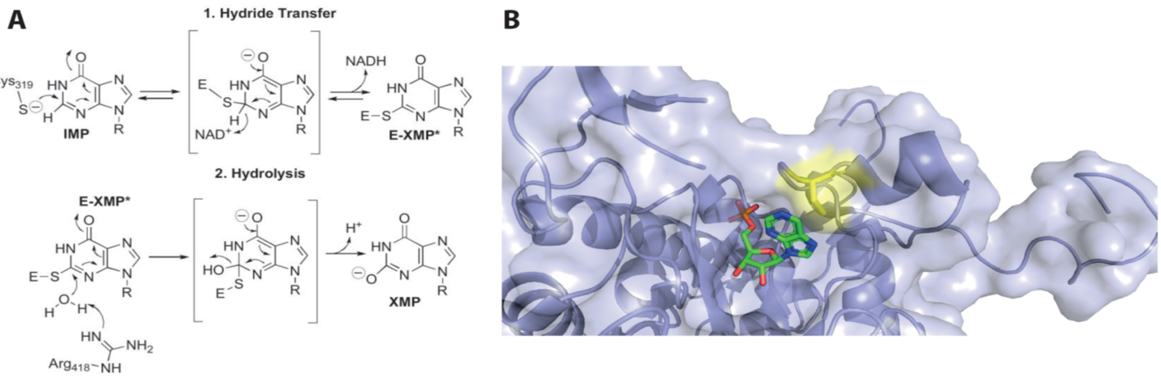


Figure 59: DEB-crosslinked cysteine in the active center of IMDH from *M. thermoresistibile.* A) Reaction mechanism of the IMDH-catalyzed conversion of inosine monophosphate to xanthosine monophosphate involving Cys-119 in *Tritrichomonas foetus* that corresponds to Cys-308 in *B. subtilis* and to Cys-201 in *M. thermoresistibile*, attacking the purine ring nucleophilically in the first step of the reaction. Figure taken from Hedstrom *et al.*[231]. **B)** Crystal structure of the IMDH protein from *M. thermoresistibile* binding purine-riboside-5'monophosphate via Cys-201. The active site of IMDH is wide and open, providing enough space to accommodate various nucleotide species to enter. DEB-mediated crosslining of IMDH found in *B. subtilis* identifies both G and U nucleotides to be crosslinked to the catalytic Cysteine-308. PDB code: 6mjy[230]

DEB-crosslinking identified queuine tRNA-ribosyltransferase Tgt to be crosslinked to U nucleotides via DEB at Cys-271. The corresponding peptide was identified to be 265-GVDMFDCVLPTR-276 is part of the homology-inferred active site with Asp-270 acting as the nucleophile attacking the guanine-34 to be exchanged by the queuine precursor 7-aminomethyl-7-deazaguanine (PreQ1)[199]. Inferences are made from homology to the Tgt homolog found in *E. coli*, where Asp-89 initiates catalysis. The nucleophilic Asp side chain attacks the C1' of nucleotide G-34 (Wobble base position) to displace the nucleobase, forming a covalent enzyme-RNA intermediate [198, 199]. Subsequently, the incoming PreQ1 is deprotonated, allowing a nucleophilic attack on the C1' of the ribose to form 7-aminomethyl-7-deazaguanine [198]. To fully modify preQ1 to become the nucleoside queine Q, two additional enzymatic reactions follow, resulting in the fully modified queine at the Wobble postion-34 [198]. DEB-crosslinking seems to detect the Cys next to the active site Asp that launches the first nucleophilic attack, but the cysteine was not found crosslinked to G nucleotides, but U nucleotides. This is not surprising given that the queuine-modified Wobble-bases are found within GUN anticodon sequences that contain a U nucleobase in position-35, leading to the hypothesis that DEB-mediated crosslinking identifies the nucleophilic amino acid cysteine being in direct contact with the uracil nucleotide of the GUN-anticodon [198].

An overall lowered number of yielded crosslinked proteins compared to *E. coli*, however, warrants a closer inspection of the sample processing step. Since SDS-PAGE analyses of crosslinked proteins from *B. subtilis* similar to figure 30 for *E. coli* reveals smearing of the bands at the given crosslinking concentration of 50 mM DEB akin to findings from *E. coli* (data not shown), it stands to reason that *B. subtilis* either i) does not tolerate DEB treatment as well as *E. coli* or ii) the sample preparation is not optimal for Gram-positive samples. The first hypothesis could easily be tested by generating a growth kinetic for *B. subtilis* similar to the one shown in figure 31. The latter hypothesis would require tinkering with lysis conditions and sample clean-up that would need to be evaluated sequentially. Additionally, one could increase depth of analysis by fractionating the sample, commonly by basic RP-LC. Since the data provided stems from preliminary experiments, generating replicate measurements may alleviate the problem of lowered identifications and should be addressed first when moving from preliminary data to full investigation. Nonetheless, data discussed here show that chemical crosslinking coupled with mass spectrometry can be successfully applied to identify RNA-binding proteins crosslinked *in vivo* from *B. subtilis*, extending its detectable range from Gram-negative bacterial cells to Gram-positive bacterial cells. A lowered identification of crosslinked proteins compared to *E. coli* indicates that improvements have to be made in sample preparation, though.

10.5.5 Chemical crosslinking of human epithelial cervical cancer cells (HeLa cells)

Crosslinking human epithelial cervical cancer cells (HeLa) chemically *in vivo* resulted in a small number of crosslinked proteins. The same workflow that was originally developed to accommodate protein-RNA complexes crosslinked *in vitro* was applied to crosslinked HeLa cells; excessive amounts of DNA being released upon cell lysis impeded subsequent sample clean-up steps, probably resulting in a lowered number of identified crosslinked proteins. Still, a small STRING network could be constructed around ribosomal proteins of both the small 40S subunit and the large 60S subunit, as well as three histone proteins clustering together. The three identified histone proteins H1.4, H2A1B, and H1.2 were all crosslinked to dinucleotides via a guansine nucleotide. Specifically, linker histone H1.4 was identified to be crosslinked to a GG dinucleotide via DEB at Lys-140, identified in the peptide 140-KATGAATPK-148. Another linker histone (H1.2) was found to be chemically crosslinked at T-165 or T-167 to a GG dinucleotide, and the core histone protein H2A1B that is part of the nucleosome core particle was found to be chemically crosslinked at His-83 to a CG dinucleotide via the guanosine nucleotide. As expected, crosslinking amino acids are nucleophilic in nature and the guanosine-bias from DEB affects crosslinked nucletides heavily.

While histone proteins are known to compact DNA inside the nucleus by forming nucleosomes, binding RNA by the (linker) histones seems out of the ordinary at first glance. While there is great consensus around nucleosomes mainly shifting their positions during DNA transcription to allow for DNA-dependent RNA-Polymerase to actively transcribe certain genes, complete dissociation of nucleosomes from the DNA is primarily limited to instances involving replication [232]. Interestingly, RNA binding of histone proteins, including H1.4 is described at very low levels by Baltz *et al.* as oligo(dT) capture of mRNA-binding proteins upon UV-crosslinking also enriches for histone proteins at low levels that were confirmed mass spectrometrically in an embryonic kidney cell line [233]. One could reason that polyploid HeLa cells that undergo rapid and unrgulated cell growth require the DNA-replication machinery to be constantly active, replicating DNA and transcribing genes to support tumor cell growth; DNA/remodeling must also occur simultaneously. HeLa cells contain 70-90 chromosomes, compacted into an enlarged nucleus that is constantly signaling the cells to undergo cell division by replicating and transcribing DNA, making crowding effect inside the karyoplasm quite plausible [159]. Thus, it stands to reason that histone proteins are probably one of the most abundant proteins classes, if not the most abundant protein class in cancerous cells carrying 70-90 chromosomes that require DNA compaction. In such a crowded environment, transient interacting between histone proteins and RNAs may be more than improbable and would explain the low levels of identified crosslinked histone proteins by Baltz *et al*, as well as serving as an initial hypothesis to explain histone-RNA crosslinks in cancer cells. Analyzing non-cancerous cells via oligo(dT) capture and/or DEB-mediated chemical crosslinking should decrease levels of identified histone proteins to (near) zero if an enhanced replication and transcription machinery is the reason for crowding-induced histone-RNA crosslinks.

The main cluster of ribosomal proteins L14, L26, and S15, however, is in accordance with previous knowledge on ribosomal proteins, being rRNA binding proteins. 60S ribosomal protein L14 was found to be crosslinked at Cys-54 to a GU dinucleotide via the guanosine (crosslinked peptide: 54-CMQLTDFILK-63). L14 is known to be involved in the assembly of the large subunit, exerting its main corresponding function in protein translation and binds the 28S rRNA that was also confirmed mass spectrometrically by the the identified crosslinking site [234]. Similarly, 60S ribosomal protein L26 involved in 60S large subunit assembly of the human ribosome was found to be crosslinked at His-18 to a GG dinucleotide (crosslinked peptide: 18-HFNAPSHIR-26) Interestingly, L26 is known to be binding to the 5' untranslated region (UTR) of p53 mRNA, increasing p53 translation post DNA damage [235]. A DNA-damage response post DEB-exposure is very likely to occur as DEB is known to induce intra-DNA crosslinks, as well as inter-crollinks to associated proteins [236]. The identified crosslinking site is closest to the 28S rRNA, but the crosslinked peptide falls within a fairly elongated and possibly flexible region on the surface of the ribosome to which 5'-UTRs of mRNAs could potentially bind. Lastly, 40S ribosomal protein S15 was identified to be chemically crosslinked at R-130 to GU via the G nucleotide. R-130 of the corresponding crosslinked peptide 128-HGRPGIGATHSSR-140 is positioned in a way that it juxtaposed the adjacent RNA-helix by facing both base-paring RNA-nucleobases.

Overall, chemical crosslinking of HeLa cells *in vivo* reveals that DEB can be used successfully to identify the most abundant nucleotide-binding proteins in human cancer cells, extending its applicability from protein-complexes and bacterial cells to eukaryotic cells. However, sample preparation needs to be optimized for future experiments, possibly including DNAse treatments to reduce the viscosity of the lysate that impeded proper sample clean-up in this instance. Additionally, fractionation of crosslinked samples probably increases analytical depth compared to this preliminary single-shot mass spectrometric run.

11 Future research

Having proven that DEB-mediated chemically crosslinking coupled with mass spectrometry can be used effectively to identify peptide-nucleotide heteroconjugates both *in vitro* and *in vivo* robustly, reliably, and unambiguously, future research might aim at expanding identifications of crosslinked proteins in different *in vivo* systems. To increase depth of analysis, sample fractionation using basic RP for example can be investigated systematically to identify best sample preparation methods to maximize benefits of chemical crosslinking coupled with mass spectrometry. Other orthogonal purification and enrichment methods besides TiO_2 are available such as SEC, silica-based chromatography, RBP or RNA pull-down methods and warrant proper investigation to increase sample output. Additionally, nucleotide-specific, but mainly metabolic enzymes using ATP, NAD, GTP, FAD, FMN, and coenzyme A that are inevitably co-identified if the nucleotide or nucleotide analog breaks off the cofactor can be studied in tandem to verify active centers or allosteric regulatory centers of different proteins. As such, cofactors in general can be crosslinked and investigated by mass spectrometry. It would be interesting to see, for example, if the cofactor GTP cross-link from EFtu2 in *E. coli* can be found in *in vitro* complexes or if Inosine-5'-monophosphate dehydrogenase from either *E. coli* or *B. subtilis* does indeed feature crosslinking sites to the NAD cofactor via adenince or if the inosine can be detected if the modifications are changed to include other nucleotides such as inosine or hypoxanthine. Additionally, some enzymes such as the pyruvate dehydrogenase complex utilize vitamins such as thiamine pyrophosphate (vitamine B1) as cofactors that are also comprised of a purine ring that could resemble a nucleotide given neutral losses of water and ammonia. Even though the ppm differences do not match perfectly, future studies could try to identify if DEB-crosslinking does indeed capture vitamin cofactors in stable states bound to enzymes.

Moreover, this study includes preliminary data using FAIMS to enhance identifying crosslinked proteins, screening for settings most optimal to the approach. While some intriguing findings seem to hint at crosslinker specificity, nucleotide composition, nucleotide length, as well as precursor charges in regards to CV settings, additional replicates are required to formulate generalized findings conclusively. The FAIMS specific settings can also be fine-tuned and expanded to higher CVs beyond -70V to investigate maximal crosslinked proteins output, as well as attempting to include a fourth CV on a long gradient. Overall, FAIMS coupled with the method described in this thesis can greatly increase the number of identified crosslinking sites in different organisms or even tissues. Since Nuxl can be adapted quite easily to DNA, future research can be aimed at investigating different cancer lines that are continuously entering the cell cycle and exhibit increased nucleotide usage in cellular metabolism, replication, transcription and translation. To do that, though, sample preparation would need to be adjusted to eukaryotic cells, but essentially any protein-nucleotide interaction can be analyzed using DEB-mediated crosslinking coupled with mass spectrometry.

Lastly, nitrogen mustard was not evaluated *in vivo* in this study due to difficulties in delivery from the producing company. Once NM can be delivered in sufficient amounts, crosslinking *E. coli* or *B. subtilis in vivo* should not be a problem since adduct masses for NM-crosslinking can be found in this thesis, supported by ample evidence from *in vitro* studies. Potential differences between the crosslinkers such as the broadened nucleotide spectrum for NM could also be evaluated if they are resurfacing *in vivo*, akin to the heavy G nucleotide bias of DEB that was also shown prominently in both *E. coli* and *B. subtilis*. Additionally, recent work in protein-nucleic crosslinking using formaldehyde by Baszo *et al.*, manuscript in preparation, could be performed in tandem to compare identified crosslinking sites. Investigating new systems using the chemical crosslinkers DEB and NM in combination to canonical UV-crosslinking would undoubtedly increase the amount of identified crosslinked proteins across all nucleotides since both approaches complement each other well, painting a more comprehensive picture of the protein-RNA landscape under investigation.

12 Literature cited

1. Lössl, P., Waterbeemd, M. & Heck, A. J. The diverse and expanding role of mass spectrometry in structural and molecular biology. *The EMBO Journal* **35,** 2634–2657. ISSN: 0261-4189 (Dec. 2016).
2. Rostislavleva, K. *et al.* Structure and flexibility of the endosomal Vps34 complex reveals the basis of its function on membranes. *Science* **350.** ISSN: 10959203. doi:10.1126/science.aac7365. https://pubmed.ncbi.nlm.nih.gov/26450213/ (Oct. 2015).
3. Özcan, A. *et al.* Type IV CRISPR RNA processing and effector complex formation in Aromatoleum aromaticum. *Nature Microbiology* **4,** 89–96. ISSN: 20585276 (Jan. 2019).
4. Marcoux, J. *et al.* Mass spectrometry reveals synergistic effects of nucleotides, lipids, and drugs binding to a multidrug resistance efflux pump. *Proceedings of the National Academy of Sciences of the United States of America* **110,** 9704–9709. ISSN: 00278424 (June 2013).
5. Johannsson, S. *et al.* Structural insights into the stimulation of S. pombe Dnmt2 catalytic efficiency by the tRNA nucleoside queuosine. *Scientific Reports* **8,** 8880. ISSN: 20452322 (June 2018).
6. Bui, K. H. *et al.* XIntegrated structural analysis of the human nuclear pore complex scaffold. *Cell* **155.** ISSN: 10974172. doi:10.1016/j.cell.2013.10.055. http://dx.doi.org/10.1016/j.cell.2013.10.055 (Dec. 2013).
7. Urlaub, H., Kruft, V., Bischof, O., Müller, E. C. & Wittmann-Liebold, B. Protein-rRNA binding features and their structural and functional implications in ribosomes as determined by cross-linking studies. *The EMBO journal* **14,** 4578–88. ISSN: 0261-4189 (Sept. 1995).
8. Thiede, B., Urlaub, H., Neubauer, H., Grelle, G. & Wittmann-Liebold, B. Precise determination of RNA-protein contact sites in the 50 S ribosomal subunit of Escherichia coli. *Biochemical Journal* **334,** 39–42. ISSN: 02646021 (Aug. 1998).
9. Schuller, J. M., Beck, F., Lössl, P., Heck, A. J. & Förster, F. Nucleotide-dependent conformational changes of the AAA+ ATPase p97 revisited. *FEBS Letters* **590,** 595–604. ISSN: 18733468 (Mar. 2016).
10. Stützer, A. *et al.* Analysis of protein-DNA interactions in chromatin by UV induced cross-linking and mass spectrometry. *Nature Communications* **11.** ISSN: 20411723. doi:10.1038/s41467-020-19047-7. https://pubmed.ncbi.nlm.nih.gov/33067435/ (Dec. 2020).
11. D'Arcy, S. *et al.* Chaperone Nap1 Shields Histone Surfaces Used in a Nucleosome and Can Put H2A-H2B in an Unconventional Tetrameric Form. *Molecular Cell* **51,** 662–677. ISSN: 10972765 (Sept. 2013).
12. Linden, A. *et al.* A Cross-linking Mass Spectrometry Approach Defines Protein Interactions in Yeast Mitochondria. *Molecular & cellular proteomics : MCP* **19,** 1161–1178. ISSN: 15359484 (July 2020).
13. Uetrecht, C., Barbu, I. M., Shoemaker, G. K., Van Duijn, E. & Heck, A. J. Interrogating viral capsid assembly with ion mobility-mass spectrometry. *Nature Chemistry* **3,** 126–132. ISSN: 17554330 (Feb. 2011).
14. Silbern, I. *et al.* Protein phosphorylation in depolarized synaptosomes: Dissecting primary effects of calcium from synaptic vesicle cycling. *Molecular and Cellular Proteomics* **20.** ISSN: 15359484. doi:10.1016/J.MCPRO.2021.100061. https://pubmed.ncbi.nlm.nih.gov/33582301/ (2021).

15. Shi, Y. *et al.* A strategy for dissecting the architectures of native macromolecular assemblies. *Nature Methods* **12,** 1135–1138. ISSN: 15487105 (Dec. 2015).

16. Sonnleitner, E. & Bläsi, U. Regulation of Hfq by the RNA CrcZ in Pseudomonas aeruginosa Carbon Catabolite Repression. *PLoS Genetics* **10.** ISSN: 15537404. doi:`10.1371/journal.pgen.1004440` (2014).

17. Urnavicius, L. *et al.* The structure of the dynactin complex and its interaction with dynein. doi:`10.1126/science.aaa4080`.

18. Giansanti, P., Tsiatsiani, L., Low, T. Y. & Heck, A. J. R. Six alternative proteases for mass spectrometry-based proteomics beyond trypsin. *Nat. Protocols* **11,** 993–1006. ISSN: 1754-2189 (2016).

19. Coon, J. J., Syka, J. E. P., Shabanowitz, J. & Hunt, D. F. *Tandem Mass Spectrometry for Peptide and Protein Sequence Analysis* tech. rep. 4 (2005).

20. Huang, M. Z., Yuan, C. H., Cheng, S. C., Cho, Y. T. & Shiea, J. Ambient ionization mass spectrometry. *Annual Review of Analytical Chemistry* **3,** 43–65. ISSN: 19361327 (July 2010).

21. Purves, R. W. & Guevremont, R. Electrospray ionization high-field asymmetric waveform ion mobility spectrometry-mass spectrometry. *Analytical Chemistry* **71,** 2346–2357. ISSN: 00032700 (1999).

22. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. *Electrospray ionization for mass spectrometry of large biomolecules* 1989. doi:`10.1126/science.2675315`. `https://pubmed.ncbi.nlm.nih.gov/2675315/`.

23. Zhang, Y., Fonslow, B. R., Shan, B., Baek, M. C. & Yates, J. R. *Protein analysis by shotgun/bottom-up proteomics* Apr. 2013. doi:`10.1021/cr3003533`.

24. Manes, N. P. & Nita-Lazar, A. Application of targeted mass spectrometry in bottom-up proteomics for systems biology research. *Journal of Proteomics* **189,** 75–90. ISSN: 18767737 (Oct. 2018).

25. Karr, U., Simonian, M. & Whitelegge, J. P. Integral membrane proteins: bottom-up, top-down and structural proteomics. *Expert Review of Proteomics* **00,** 14789450.2017.1359545. ISSN: 1478-9450 (2017).

26. Giansanti, P., Tsiatsiani, L., Low, T. Y. & Heck, A. J. Six alternative proteases for mass spectrometry-based proteomics beyond trypsin. *Nature Protocols* **11,** 993–1006. ISSN: 17502799 (May 2016).

27. Domon, B. & Aebersold, R. Mass spectrometry and protein analysis. *Science* **312,** 212–217. ISSN: 1095-9203 (2006).

28. Winger, B. E., Light-Wahl, K. J., Ogorzalek Loo, R. R., Udseth, H. R. & Smith, R. D. Observation and implications of high mass-to-charge ratio ions from electrospray ionization mass spectrometry. *Journal of the American Society for Mass Spectrometry* **4,** 536–545. ISSN: 18791123 (1993).

29. Charles, L., Pépin, D., Gonnet, F. & Tabet, J. C. Effects of liquid phase composition on salt cluster formation in positive ion mode electrospray mass spectrometry: Implications for clustering mechanism in electrospray. *Journal of the American Society for Mass Spectrometry* **12,** 1077–1084. ISSN: 10440305 (2001).

30. Haag, A. M. in *Advances in Experimental Medicine and Biology* 157–169 (Springer New York LLC, 2016). doi:10.1007/978-3-319-41448-5{_}7. https://link.springer.com/chapter/10.1007/978-3-319-41448-5_7.

31. Zubarev, R. A. & Makarov, A. Orbitrap mass spectrometry. *Analytical Chemistry* **85,** 5288–5296. ISSN: 00032700 (June 2013).

32. Orbitrap Exploris 480 Mass Spectrometers - DE. //www.thermofisher.com/de/de/home/industrial/mass-spectrometry/liquid-chromatography-mass-spectrometry-lc-ms/lc-ms-systems/orbitrap-lc-ms/orbitrap-exploris-mass-spectrometers/orbitrap-exploris-480-mass-spectrometers.html.

33. Kanu, A. B., Dwivedi, P., Tam, M., Matz, L. & Hill, H. H. *Ion mobility-mass spectrometry* Jan. 2008. doi:10.1002/jms.1383. https://pubmed.ncbi.nlm.nih.gov/18200615/.

34. Purves, R. W., Prasad, S., Belford, M., Vandenberg, A. & Dunyach, J. J. Optimization of a New Aerodynamic Cylindrical FAIMS Device for Small Molecule Analysis. *Journal of the American Society for Mass Spectrometry* **28,** 525–538. ISSN: 18791123 (Mar. 2017).

35. Hebert, A. S. *et al.* Comprehensive Single-Shot Proteomics with FAIMS on a Hybrid Orbitrap Mass Spectrometer. *Analytical Chemistry* **90,** 9529–9537. ISSN: 15206882 (Aug. 2018).

36. Bekker-Jensen, D. B. *et al.* A compact quadrupole-orbitrap mass spectrometer with FAIMS interface improves proteome coverage in short LC gradients. *Molecular and Cellular Proteomics* **19,** 716–729. ISSN: 15359484 (2020).

37. Roepstorff, P. & Fohlman, J. *Letter to the editors* 1984. doi:10.1002/bms.1200111109. https://pubmed.ncbi.nlm.nih.gov/6525415/%20https://pubmed.ncbi.nlm.nih.gov/6525415/?dopt=Abstract.

38. Biemann, K. Peptides and proteins: Overview and strategy. *Methods in Enzymology* **193,** 351–360. ISSN: 15577988 (Jan. 1990).

39. Domon, B. & Costello, C. E. Structure Elucidation of Glycosphingolipids and Gangliosides Using High-Performance Tandem Mass Spectrometry. *Biochemistry* **27,** 1534–1543. ISSN: 15204995 (Mar. 1988).

40. *Peptide Identification by Tandem Mass Spectrometry with Alternate Fragmentation Modes* | Elsevier Enhanced Reader* https://reader.elsevier.com/reader/sd/pii/S1535947620325706.

41. Noble, W. S. Mass spectrometrists should search only for peptides they care about. *Nature Methods* **12,** 605–608. ISSN: 1548-7091 (June 2015).

42. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry&ndash;based shotgun proteomics. *Nature Protocols* **11.** doi:10.1038/nprot.2016.136. https://thermo.flexnetoperations.com/control/ (2016).

43. Urgen Cox, J. *et al.* Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *J. Proteome Res* **10.** doi:10.1021/pr101065j. www.biochem.mpg.de/mann/tools/. (2011).

44. Lundgren, D. H., Han, D. K. & Eng, J. K. Protein identification using TurboSEQUEST. *Current protocols in bioinformatics* **Chapter 13,** Unit 13.3. ISSN: 1934-340X (July 2005).

45. Fan, S. B. *et al.* Using pLink to analyze cross-linked peptides. *Current Protocols in Bioinformatics* **2015,** 1–8. ISSN: 1934340X (2015).

46. Kramer, K. *et al.* Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins. *Nature Methods* **11,** 1064–1070 (2014).

47. Chen, C., Hou, J., Tanner, J. J. & Cheng, J. Molecular Sciences Bioinformatics Methods for Mass Spectrometry-Based Proteomics Data Analysis. doi:10.3390/ijms21082873. www.mdpi.com/journal/ijms.

48. Spraggins, J. M. *et al.* Next-generation technologies for spatial proteomics: Integrating ultra-high speed MALDI-TOF and high mass resolution MALDI FTICR imaging mass spectrometry for protein analysis. *Proteomics* **16,** 1678–1689. ISSN: 16159861 (2016).

49. Kotsiantis, S. B. *Supervised Machine Learning: A Review of Classification Techniques* tech. rep. (2007), 249–268.

50. Mantock, J. M. A Nonparametric Two-Dimensional Display for Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-4,** 427–436. ISSN: 01628828 (1982).

51. Lee, J. Y. & Styczynski, M. P. NS-kNN: a modified k-nearest neighbors approach for imputing metabolomics data. *Metabolomics* **14.** ISSN: 15733890. doi:10.1007/s11306-018-1451-8. https://pubmed.ncbi.nlm.nih.gov/30830437/ (Dec. 2018).

52. Bania, R. K. & Halder, A. R-Ensembler: A greedy rough set based ensemble attribute selection algorithm with kNN imputation for classification of medical data. *Computer Methods and Programs in Biomedicine* **184.** ISSN: 18727565. doi:10.1016/j.cmpb.2019.105122. https://pubmed.ncbi.nlm.nih.gov/31622857/ (Feb. 2020).

53. Chen, L. *et al.* Investigating the gene expression profiles of cells in seven embryonic stages with machine learning algorithms. *Genomics* **112,** 2524–2534. ISSN: 10898646 (May 2020).

54. Hentze, M. W., Castello, A., Schwarzl, T. & Preiss, T. *A brave new world of RNA-binding proteins* May 2018. doi:10.1038/nrm.2017.130. https://pubmed.ncbi.nlm.nih.gov/29339797/.

55. Esteller, M. Non-coding RNAs in human disease. doi:10.1038/nrg3074. www.nature.com/reviews/genetics (2011).

56. Nussbacher, J. K., Tabet, R., Yeo, G. W. & Lagier-Tourenne, C. *Disruption of RNA Metabolism in Neurological Diseases and Emerging Therapeutic Interventions* Apr. 2019. doi:10.1016/j.neuron.2019.03.014. https://pubmed.ncbi.nlm.nih.gov/30998900/.

57. Akhter, R. Circular RNA and Alzheimer's Disease. doi:10.1007/978-981-13-1426-1{_}19. https://doi.org/10.1007/978-981-13-1426-1_19.

58. Hentze, M. W., Castello, A., Schwarzl, T. & Preiss, T. *A brave new world of RNA-binding proteins* May 2018. doi:10.1038/nrm.2017.130.

59. Beckmann, B. M. *et al.* The RNA-binding proteomes from yeast to man harbour conserved enigmRBPs. *Nature Communications* **6,** 10127. ISSN: 2041-1723 (Dec. 2015).

60. Linder, P. *Dead-box proteins: A family affair - Active and passive players in RNP-remodeling* Sept. 2006. doi:10.1093/nar/gkl468. https://pubmed.ncbi.nlm.nih.gov/16936318/.

61. Wahl, M. C., Will, C. L. & Lührmann, R. *The Spliceosome: Design Principles of a Dynamic RNP Machine* Feb. 2009. doi:10.1016/j.cell.2009.02.009. https://pubmed.ncbi.nlm.nih.gov/19239890/.

62. Corley, M., Burns, M. C. & Yeo, G. W. *How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms* Apr. 2020. doi:10.1016/j.molcel.2020.03.011. https://pubmed.ncbi.nlm.nih.gov/32243832/.

63. De Graeve, F. & Bessé, F. *Neuronal RNP granules: From physiological to pathological assemblies* June 2018. doi:10.1515/hsz-2018-0141. https://pubmed.ncbi.nlm.nih.gov/29641413/.

64. Wheeler, J. R., Matheny, T., Jain, S., Abrisch, R. & Parker, R. Distinct stages in stress granule assembly and disassembly. *eLife* **5.** ISSN: 2050084X. doi:10.7554/eLife.18413. https://pubmed.ncbi.nlm.nih.gov/27602576/ (Sept. 2016).

65. Woodson, S. A., Panja, S. & Santiago-Frangos, A. Proteins That Chaperone RNA Regulation. *Microbiology Spectrum* **6.** ISSN: 2165-0497. doi:10.1128/microbiolspec.rwr-0026-2018. https://pubmed.ncbi.nlm.nih.gov/30051798/ (July 2018).

66. Rennella, E. *et al.* RNA binding and chaperone activity of the E. coli cold-shock protein CspA. *Nucleic Acids Research* **45,** 4255–4268. ISSN: 13624962 (Apr. 2017).

67. Takyar, S., Hickerson, R. P. & Noller, H. F. mRNA helicase activity of the ribosome. *Cell* **120,** 49–58. ISSN: 00928674 (Jan. 2005).

68. Tian, S. & Das, R. *RNA structure through multidimensional chemical mapping* Jan. 2016. doi:10.1017/S0033583516000020. https://pubmed.ncbi.nlm.nih.gov/27266715/.

69. Mitchell, D., Assmann, S. M. & Bevilacqua, P. C. *Probing RNA structure in vivo* Dec. 2019. doi:10.1016/j.sbi.2019.07.008. https://pubmed.ncbi.nlm.nih.gov/31521910/.

70. Daubner, G. M., Cléry, A. & Allain, F. H. T. RRM-RNA recognition: NMR or crystallography...and new findings. *Current Opinion in Structural Biology* **23,** 100–108. ISSN: 0959440X (2013).

71. Simpson, P. J. *et al.* Structure and RNA interactions of the N-terminal RRM domains of PTB. *Structure* **12,** 1631–1643. ISSN: 09692126 (Sept. 2004).

72. Oberstrass, F. C. *et al.* Structural biology - Structure of PTB bound to RNA: Specific binding and implications for splicing regulation. *Science* **309,** 2054–2057. ISSN: 00368075 (Sept. 2005).

73. Schindelin, H., Marahiel, M. A. & Heinemann, U. Universal nucleic acid-binding domain revealed by crystal structure of the B. subtilis major cold-shock protein. *Nature* **364,** 164–168. ISSN: 00280836 (1993).

74. Salah, P. *et al.* Probing the relationship between gram-negative and gram-positive S1 proteins by sequence analysis. *Nucleic Acids Research* **37,** 5578–5588. ISSN: 03051048 (July 2009).

75. Lewis, H. A. *et al.* Crystal structures of Nova-1 and Nova-2 K-homology RNA-binding domains. *Structure* **7,** 191–203. ISSN: 09692126 (Feb. 1999).

76. Burdisso, P. *et al.* Second double-stranded RNA binding domain of dicer-like ribonuclease 1: Structural and biochemical characterization. *Biochemistry* **51,** 10159–10166. ISSN: 00062960 (Dec. 2012).

77. Pavletich, N. P. & Pabo, C. O. Zinc finger-DNA recognition: Crystal structure of a Zif268-DNA complex at 2.1 Å. *Science* **252,** 809–817. ISSN: 00368075 (May 1991).

78. Jones, P. G. & Inouye, M. *The cold-shock response — a hot topic* 1994. doi:10.1111/j.1365-2958.1994.tb00359.x. https://pubmed.ncbi.nlm.nih.gov/8022259/.

79. Mihailovich, M., Militti, C., Gabaldón, T. & Gebauer, F. *Eukaryotic cold shock domain proteins: Highly versatile regulators of gene expression* 2010. doi:10.1002/bies.200900122.

80. Brandi, A., Giangrossi, M., Giuliodori, A. M. & Falconi, M. An interplay among FIS, H-NS, and guanosine tetraphosphate modulates transcription of the Escherichia coli cspA gene under physiological growth conditions. *Frontiers in Molecular Biosciences* **3.** ISSN: 2296889X. doi:`10.3389/fmolb.2016.00019`. `https://pubmed.ncbi.nlm.nih.gov/27252944/` (May 2016).

81. Brandi, A., Pon, C. L. & Gualerzi, C. O. Interaction of the main cold shock protein CS7.4 (CspA) of Escherichia coli with the promoter region of hns. *Biochimie* **76,** 1090–1098. ISSN: 61831638 (1994).

82. Briercheck, D. M. *et al.* 1H, 15N and 13C resonance assignments and secondary structure determination of the RNA-binding domain of E. coli rho protein. *Journal of Biomolecular NMR* **8,** 429–444. ISSN: 09252738 (1996).

83. Deriusheva, E. I., Machulin, A. V., Selivanova, O. M. & Serdiuk, I. N. [Family of ribosomal proteins S1 contains unique conservative domain]. *Molekuliarnaia biologiia* **44,** 728–734. ISSN: 00268984 (July 2010).

84. Matus-Ortega, M. E. *et al.* The KH and S1 domains of Escherichia coli polynucleotide phosphorylase are necessary for autoregulation and growth at low temperature. *Biochimica et Biophysica Acta - Gene Structure and Expression* **1769,** 194–203. ISSN: 01674781 (Mar. 2007).

85. Bycroft, M., Hubbard, T. J., Proctor, M., Freund, S. M. & Murzin, A. G. The solution structure of the S1 RNA binding domain: A member of an ancient nucleic acid-binding fold. *Cell* **88,** 235–242. ISSN: 00928674 (Jan. 1997).

86. Guo, X. *et al.* Structural Basis for NusA Stabilized Transcriptional Pausing. *Molecular Cell* **69,** 816–827. ISSN: 10974164 (Mar. 2018).

87. Wang, Z. *et al.* *The emerging roles of hnRNPK* Mar. 2020. doi:`10.1002/jcp.29186`. `https://pubmed.ncbi.nlm.nih.gov/31538344/`.

88. Nazarov, I. B., Bakhmet, E. I. & Tomilin, A. N. *KH-Domain Poly(C)-Binding Proteins as Versatile Regulators of Multiple Biological Processes* Mar. 2019. doi:`10.1134/S0006297919030039`. `https://pubmed.ncbi.nlm.nih.gov/31221059/`.

89. Wang, B.-D. & Lee, N. Aberrant RNA Splicing in Cancer and Drug Resistance. *Cancers* **10,** 458. ISSN: 2072-6694 (Nov. 2018).

90. Cohen, P. *Protein phosphorylation and hormone action* 1988. doi:`10.1098/rspb.1988.0040`. `https://pubmed.ncbi.nlm.nih.gov/2905457/`.

91. Musco, G. *et al.* *The solution structure of the first KH domain of FMR1, the protein responsible for the fragile X syndrome.* 1997. doi:`10.1038/nsb0997-712`. `https://pubmed.ncbi.nlm.nih.gov/9302998/`.

92. Oddone, A. *et al.* Structural and biochemical characterization of the yeast exosome component Rrp40. *EMBO Reports* **8,** 63–69. ISSN: 1469221X (Jan. 2007).

93. St Johnston, D., Brown, N. H., Gall, J. G. & Jantsch, M. A conserved double-stranded RNA-binding domain. *Proceedings of the National Academy of Sciences of the United States of America* **89,** 10979–10983. ISSN: 00278424 (1992).

94. Manche, L., Green, S. R., Schmedt, C. & Mathews, M. B. Interactions between double-stranded RNA regulators and the protein kinase DAI. *Molecular and Cellular Biology* **12,** 5238–5248. ISSN: 0270-7306 (Nov. 1992).

95. Kim, U., Wang, Y., Sanford, T., Zeng, Y. & Nishikura, K. Molecular cloning of cDNA for double-stranded RNA adenosine deaminase, a candidate enzyme for nuclear RNA editing. *Proceedings of the National Academy of Sciences of the United States of America* **91,** 11457–11461. ISSN: 00278424 (Nov. 1994).

96. Laity, J. H., Lee, B. M. & Wright, P. E. *Zinc finger proteins: New insights into structural and functional diversity* Feb. 2001. doi:10.1016/S0959-440X(00)00167-6. https://pubmed.ncbi.nlm.nih.gov/11179890/.

97. Finerty, P. J. & Bass, B. L. A Xenopus zinc finger protein that specifically binds dsRNA and RNA-DNA hybrids. *Journal of Molecular Biology* **271,** 195–208. ISSN: 00222836 (Aug. 1997).

98. Miki, T. *et al.* A conditional proteomics approach to identify proteins involved in zinc homeostasis. *Nature methods* **13,** 1–10. ISSN: 1548-7105 (2016).

99. Wolfe, S. A., Nekludova, L. & Pabo, C. O. *DNA recognition by Cys2His2 zinc finger proteins* 2000. doi:10.1146/annurev.biophys.29.1.183. https://pubmed.ncbi.nlm.nih.gov/10940247/.

100. *Beyond CLIP: advances and opportunities to measure RBP–RNA and RNA–RNA interactions* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6582316/.

101. Hafner, M. *et al.* CLIP and complementary methods. *Nature Reviews Methods Primers* **1,** 1–23 (Dec. 2021).

102. Sonnleitner, E. *et al.* Interplay between the catabolite repression control protein Crc, Hfq and RNA in Hfq-dependent translational regulation in Pseudomonas aeruginosa. *Nucleic Acids Research* **46,** 1470–1485. ISSN: 13624962 (Feb. 2018).

103. Stützer, A. *et al.* Analysis of protein-DNA interactions in chromatin by UV induced cross-linking and mass spectrometry. *Nature Communications* **11.** ISSN: 20411723. doi:10.1038/s41467-020-19047-7. https://pubmed.ncbi.nlm.nih.gov/33067435/ (Dec. 2020).

104. Van Esveld, S. L. & Spelbrink, J. N. in *Methods in Molecular Biology* 147–158 (Humana Press Inc., 2021). doi:10.1007/978-1-0716-0834-0{_}12. https://pubmed.ncbi.nlm.nih.gov/33230772/.

105. Milek, M. & Landthaler, M. in *Methods in Molecular Biology* 405–417 (Humana Press Inc., 2018). doi:10.1007/978-1-4939-7213-5{_}27. https://pubmed.ncbi.nlm.nih.gov/29130213/.

106. Castello, A. *et al.* System-wide identification of RNA-binding proteins by interactome capture. *Nature Protocols* **8,** 491–500. ISSN: 17542189 (Feb. 2013).

107. Perez-Perri, J. I. *et al.* Discovery of RNA-binding proteins and characterization of their dynamic responses by enhanced RNA interactome capture. *Nature Communications* **9.** ISSN: 20411723. doi:10.1038/s41467-018-06557-8. https://pubmed.ncbi.nlm.nih.gov/30352994/ (Dec. 2018).

108. Perez-Perri, J. I. *et al.* Global analysis of RNA-binding protein dynamics by comparative and enhanced RNA interactome capture. *Nature Protocols* **16,** 27–60. ISSN: 17502799 (Jan. 2021).

109. Perez-Perri, J. I. *et al.* Discovery of RNA-binding proteins and characterization of their dynamic responses by enhanced RNA interactome capture. *Nature Communications* **9,** 1–13. ISSN: 20411723 (Dec. 2018).

110. Castello, A. *et al.* Identification of RNA-binding domains of RNA-binding proteins in cultured cells on a system-wide scale with RBDmap. *Nature Protocols* **12,** 2447–2464. ISSN: 17502799 (Dec. 2017).

111. Shchepachev, V. *et al.* Defining the RNA interactome by total RNA -associated protein purification. *Molecular Systems Biology* **15.** ISSN: 1744-4292. doi:10.15252/msb.20188689. https://pubmed.ncbi.nlm.nih.gov/30962360/ (Apr. 2019).

112. Shchepachev, V. *et al.* Defining the <scp>RNA</scp> interactome by total <scp>RNA</scp> -associated protein purification. *Molecular Systems Biology* **15.** ISSN: 1744-4292. doi:10.15252/msb.20188689. https://onlinelibrary.wiley.com/doi/abs/10.15252/msb.20188689 (Apr. 2019).

113. Dorn, G. *et al.* Structural modeling of protein-RNA complexes using crosslinking of segmentally isotope-labeled RNA and MS/MS. *Nature Methods* **14,** 487–490. ISSN: 15487105 (Apr. 2017).

114. Dorn, G. *et al.* Structural modeling of protein–RNA complexes using crosslinking of segmentally isotope-labeled RNA and MS/MS. *Nature Methods* **14,** 487–490. ISSN: 1548-7091 (Apr. 2017).

115. Chen, Z. A. & Rappsilber, J. Quantitative cross-linking/mass spectrometry to elucidate structural changes in proteins and their complexes. *Nature Protocols* **14,** 171–201. ISSN: 17502799 (Jan. 2019).

116. Bartolec, T. K. *et al.* Crosslinking mass spectrometry analysis of the yeast nucleus reveals extensive protein-protein interactions not detected by systematic two-hybrid or affinity purification-mass spectrometry. *Analytical Chemistry,* acs.analchem.9b03975. ISSN: 0003-2700 (Dec. 2019).

117. Glover, B. P., Pritchard, A. E. & McHenry, C. S. τ binds and organizes Escherichia coli replication proteins through distinct domains: Domain III, shared by γ and τ, oligomerizes DnaX. *Journal of Biological Chemistry* **276,** 35842–35846. ISSN: 00219258 (Sept. 2001).

118. Lin, P.-C. & Xu, R.-M. Structure and assembly of the SF3a splicing factor complex of U2 snRNP. *The EMBO journal* **31,** 1579–90. ISSN: 1460-2075 (Mar. 2012).

119. National Toxicology Program. Diepoxybutane. *Report on carcinogens : carcinogen profiles* **12,** 152–3. ISSN: 1551-8280 (2011).

120. Sköld, S. E. RNA-protein complexes identified by crosslinking of polysomes. *Biochimie* **63,** 53–60. ISSN: 61831638 (Jan. 1981).

121. Park, S., Hodge, J., Anderson, C. & Tretyakova, N. Guanine-Adenine DNA Cross-Linking by 1,2,3,4-Diepoxybutane: Potential Basis for Biological Activity. *Chemical Research in Toxicology* **17,** 1638–1651. ISSN: 0893-228X (Dec. 2004).

122. Sköld, S. E. RNA-protein complexes identified by crosslinking of polysomes. *Biochimie* **63,** 53–60. ISSN: 61831638 (Jan. 1981).

123. Sköld, S. E. Chemical crosslinking of elongation factor G to the 23S RNA in 70S ribosomes from Escherichia coli. *Nucleic acids research* **11,** 4923–32. ISSN: 0305-1048 (July 1983).

124. Gherezghiher, T. B., Ming, X., Villalta, P. W., Campbell, C. & Tretyakova, N. Y. 1,2,3,4-Diepoxybutane-induced DNA-protein cross-linking in human fibrosarcoma (HT1080) cells. *Journal of Proteome Research* **12,** 2151–2164. ISSN: 15353893 (2013).

125. Zhang, X.-Y. & Elfarra, A. A. Characterization of 1,2,3,4-Diepoxybutane-2'-Deoxyguanosine Cross-Linking Products Formed at Physiological and Nonphysiological Conditions. *Chemical Research in Toxicology* **19,** 547–555. ISSN: 0893-228X (Apr. 2006).

126. Kallama, S. & Hemminki, K. Stabilities of 7-alkylguanosines and 7-deoxyguanosines formed by phosphoramide mustard and nitrogen mustard. *Chemico-Biological Interactions* **57,** 85–96. ISSN: 00092797 (Jan. 1986).

127. Erin, D., Michaelson, R., Xun, M. & Daniel, C. Mechlorethamine-Induced DNA-PRotein Cross-Linking in Human Fibrosarcoma Cells. *Journal of Proteome Research* **10,** 612–626. ISSN: 1535-3893 (2012).

128. Polavarapu, A., Stillabower, J. A., Stubblefield, S. G. W., Taylor, W. M. & Baik, M.-H. The Mechanism of Guanine Alkylation by Nitrogen Mustards: A Computational Study. *The Journal of Organic Chemistry* **77,** 5914–5921. ISSN: 0022-3263 (July 2012).

129. Loeber, R. L. *et al.* Proteomic Aanalysis of DNA-protein cross-linking by antitumor nitrogen mustards. *Chemical Research in Toxicology* **22,** 1151–1162. ISSN: 0893228X (June 2009).

130. Balcome, S. *et al.* Adenine-containing DNA-DNA cross-links of antitumor nitrogen mustards. *Chemical Research in Toxicology* **17,** 950–962. ISSN: 0893228X (July 2004).

131. Chatterjee, N. & Walker, G. C. *Mechanisms of DNA damage, repair, and mutagenesis* June 2017. doi:10.1002/em.22087. https://pubmed.ncbi.nlm.nih.gov/28485537/.

132. Masta, A., Gray, P. J. & Phillips, D. R. Nitrogen mustard inhibits transcription and translation in a cell free system. *Nucleic Acids Research* **23,** 3508–3515. ISSN: 03051048 (Sept. 1995).

133. Datta, B. & Weiner, A. M. Cross-linking of U1 snRNA using nitrogen mustard. Evidence for higher order structure. *The Journal of biological chemistry* **267,** 4503–7. ISSN: 0021-9258 (Mar. 1992).

134. Traut, R. R. *et al.* Methyl 4-mercaptobutyrimidate as a cleavable cross-linking reagent and its application to the Escherichia coli 30S ribosome. *Biochemistry* **12,** 3266–3273. ISSN: 0006-2960 (1973).

135. Kenny, J. W., Lambert, J. M. & Traut, R. R. Cross-Linking of Ribosomes Using 2-Iminothiolane (Methyl 4-Mercaptobutyrimidate) and Identification of Cross-Linked Proteins by Diagonal Polyacrylamide/Sodium Dodecyl Sulfate Gel Electrophoresis. *Methods in Enzymology* **59,** 534–550. ISSN: 15577988 (Jan. 1979).

136. Leitner, A. *Phosphopeptide enrichment using metal oxide affinity chromatography* Feb. 2010. doi:10.1016/j.trac.2009.08.007.

137. Carrascal, M., Ovelleiro, D., Casas, V., Gay, M. & Abian, J. Phosphorylation analysis of primary human T lymphocytes using sequential IMAC and titanium oxide enrichment. *Journal of Proteome Research* **7,** 5167–5176. ISSN: 15353893 (Dec. 2008).

138. Steigenberger, B., Pieters, R. J., Heck, A. J. R. & Scheltema, R. A. PhoX: An IMAC-Enrichable Cross-Linking Reagent. *ACS Central Science* **5,** 1514–1522. ISSN: 2374-7943 (Sept. 2019).

139. Buncherd, H. *et al.* Isolation of cross-linked peptides by diagonal strong cation exchange chromatography for protein complex topology studies by peptide fragment fingerprinting from large sequence databases. *Journal of Chromatography A* **1348,** 34–46. ISSN: 18733778 (June 2014).

140. Asencio, C., Chatterjee, A. & Hentze, M. W. Silica-based solid-phase extraction of cross-linked nucleic acid–bound proteins. *Life Science Alliance* **1.** ISSN: 25751077. doi:10.26508/lsa.201800088 (2018).

141. Zhao, X. *et al.* A simple methodology for RNA isolation from bacteria by integration of formamide extraction and chitosan-modified silica purification. *Analytical and Bioanalytical Chemistry.* ISSN: 1618-2642. doi:10.1007/s00216-021-03644-6. https://link.springer.com/10.1007/s00216-021-03644-6 (Sept. 2021).

142. Vijayalakshmy, K. *et al.* A novel combination of silane-coated silica colloid with hybrid RNA extraction protocol and RNA enrichment for downstream applications of spermatozoal RNA. *Andrologia* **50.** ISSN: 14390272. doi:10.1111/and.13030 (Aug. 2018).

143. Nierves, L. & Lange, P. F. Detectability of Biotin Tags by LC-MS/MS. *Journal of Proteome Research* **20,** 3002–3008. ISSN: 15353907 (May 2021).

144. Van Roon, A. M. M. *et al.* Crystal structure of U2 snRNP SF3b components: Hsh49p in complex with Cus1p-binding domain. *RNA* **23,** 968–981. ISSN: 14699001 (June 2017).

145. Pauling, M. H., McPheeters, D. S. & Ares, M. Functional Cus1p Is Found with Hsh155p in a Multiprotein Splicing Factor Associated with U2 snRNA. *Molecular and Cellular Biology* **20,** 2176–2185. ISSN: 0270-7306 (Mar. 2000).

146. Roon, A.-m. M. V. *et al.* with Cus1p binding domain. *Rna* (2017).

147. Jeltsch, A. *et al. Mechanism and biological role of Dnmt2 in Nucleic Acid Methylation* Sept. 2017. doi:10.1080/15476286.2016.1191737. https://pubmed.ncbi.nlm.nih.gov/27232191/.

148. Natarajan, M. *et al.* Negative elongation factor (NELF) coordinates RNA polymerase II pausing, premature termination, and chromatin remodeling to regulate HIV transcription. *Journal of Biological Chemistry* **288,** 25995–26003. ISSN: 00219258 (Sept. 2013).

149. Rawat, P. *et al.* Stress-induced nuclear condensation of NELF drives transcriptional downregulation. *Molecular Cell* **81,** 1013–1026. ISSN: 10974164 (Mar. 2021).

150. Vos, S. M. *et al.* Architecture and RNA binding of the human negative elongation factor. *eLife* **5.** ISSN: 2050-084X. doi:10.7554/eLife.14981. http://www.ncbi.nlm.nih.gov/pubmed/27282391%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4940160 (June 2016).

151. Vos, S. M., Farnung, L., Urlaub, H. & Cramer, P. Structure of paused transcription complex Pol II–DSIF–NELF. *Nature* **560,** 601–606. ISSN: 14764687 (Aug. 2018).

152. Zerbs, S., Giuliani, S. & Collart, F. in *Methods in Enzymology* 117–131 (Academic Press Inc., 2014). ISBN: 9780124200708. doi:10.1016/B978-0-12-420070-8.00011-8. https://pubmed.ncbi.nlm.nih.gov/24423272/.

153. Sadiq, S. M., Hazen, T. H., Rasko, D. A. & Eppinger, M. EHEC Genomics: Past, Present, and Future. *Microbiology Spectrum* **2.** ISSN: 2165-0497. doi:10.1128/microbiolspec.ehec-0020-2013. https://pubmed.ncbi.nlm.nih.gov/26104203/ (Aug. 2014).

154. Kim, S. *et al.* Genomic and transcriptomic landscape of Escherichia coli BL21(DE3). *Nucleic Acids Research* **45,** 5285–5293. ISSN: 13624962 (May 2017).

155. Errington, J. & Van Der Aart, L. T. Microbe Profile: Bacillus subtilis: model organism for cellular development, and industrial workhorse. *Microbiology* **166,** 425–427 (2020).

156. Aguilar, C., Vlamakis, H., Losick, R. & Kolter, R. *Thinking about Bacillus subtilis as a multicellular organism* Dec. 2007. doi:10.1016/j.mib.2007.09.006. https://pubmed.ncbi.nlm.nih.gov/17977783/.

157. Caulier, S. *et al. Overview of the antimicrobial compounds produced by members of the Bacillus subtilis group* 2019. doi:10.3389/fmicb.2019.00302.

158. Stein, T. *Bacillus subtilis antibiotics: Structures, syntheses and specific functions* May 2005. doi:10.1111/j.1365-2958.2005.04587.x. https://pubmed.ncbi.nlm.nih.gov/15853875/.

159. Heng, H. H. *HeLa genome versus donor's genome* Sept. 2013. doi:10.1038/501167d. https://www.nature.com/articles/501167d.

160. Lucey, B. P., Nelson-Rees, W. A. & Hutchins, G. M. *HeLa Cells and Cell Culture Contamination-Lucey et al 1463* tech. rep. (2009).

161. Landry, J. *et al.* The genomic and transcriptomic landscape of a HeLa cell line. doi:10.1534/g3.113.005777.

162. Bradford, M. M. *A Rapid and Sensitive Method for the Quantitation of Microgram Quantities of Protein Utilizing the Principle of Protein-Dye Binding* tech. rep. (1976), 248–254.

163. Smith, P. K. *et al.* Measurement of protein using bicinchoninic acid. *Analytical Biochemistry* **150,** 76–85. ISSN: 10960309 (Oct. 1985).

164. Kretschmer, J. *et al.* The m6A reader protein YTHDC2 interacts with the small ribosomal subunit and the 5'-3' exoribonuclease XRN1. *RNA* **24,** 1339–1350. ISSN: 14699001 (2018).

165. Valpadashi, A. *et al.* Defining the architecture of the human TIM22 complex by chemical crosslinking. *FEBS Letters* **595,** 157–168. ISSN: 18733468 (Jan. 2021).

166. Chen, Z.-L. *et al.* A high-speed search engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides. doi:10.1038/s41467-019-11337-z. https://doi.org/10.1038/s41467-019-11337-z.

167. Rodnina, M. V. & Wintermeyer, W. GTP consumption of elongation factor Tu during translation of heteropolymeric mRNAs. *Proceedings of the National Academy of Sciences of the United States of America* **92,** 1945–1949. ISSN: 00278424 (Mar. 1995).

168. Adusumilli, R. & Mallick, P. in *Methods in Molecular Biology* 339–368 (Humana Press Inc., Jan. 2017). doi:10.1007/978-1-4939-6747-6{_}23. http://link.springer.com/10.1007/978-1-4939-6747-6_23.

169. Holman, J. D., Tabb, D. L. & Mallick, P. Employing ProteoWizard to convert raw mass spectrometry data. *Current Protocols in Bioinformatics.* ISSN: 1934340X. doi:10.1002/0471250953.bi1324s46 (2014).

170. Wickham, H. *et al.* Welcome to the Tidyverse. *Journal of Open Source Software* **4,** 1686. ISSN: 2475-9066 (Nov. 2019).

171. Xie, Y. Helper Functions to Install and Maintain TeX Live, and Compile LaTeX Documents [R package tinytex version 0.30]. https://CRAN.R-project.org/package=tinytex (Mar. 2021).

172. Zhu, H. Construct Complex Table with 'kable' and Pipe Syntax [R package kableExtra version 1.3.4]. https://CRAN.R-project.org/package=kableExtra (Feb. 2021).

173. Calculate Indices and Theoretical Physicochemical Properties of Protein Sequences [R package Peptides version 2.4.3]. https://CRAN.R-project.org/package=Peptides (Nov. 2020).

174. Kuhn, M. & Kuhn, M. The caret Package. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.150.2466 (2009).

175. Kassambara, A. 'ggplot2' Based Publication Ready Plots [R package ggpubr version 0.4.0]. https://CRAN.R-project.org/package=ggpubr (June 2020).

176. Visualization of a Correlation Matrix [R package corrplot version 0.84]. https://CRAN.R-project.org/package=corrplot (Oct. 2017).

177. Read Excel Files [R package readxl version 1.3.1]. https://CRAN.R-project.org/package=readxl (Mar. 2019).

178. An R 'Pandoc' Writer [R package pander version 0.6.3]. https://CRAN.R-project.org/package=pander (Nov. 2018).

179. Tools to Make Developing R Packages Easier [R package devtools version 2.3.2]. https://CRAN.R-project.org/package=devtools (Sept. 2020).

180. *CRAN - Package Hmisc* https://cran.r-project.org/web/packages/Hmisc/index.html.

181. Neuwirth, E. ColorBrewer Palettes [R package RColorBrewer version 1.1-2]. https://CRAN.R-project.org/package=RColorBrewer (Dec. 2014).

182. Xie, Y. A General-Purpose Package for Dynamic Report Generation in R [R package knitr version 1.31]. https://CRAN.R-project.org/package=knitr (Jan. 2021).

183. A Forward-Pipe Operator for R [R package magrittr version 2.0.1]. https://CRAN.R-project.org/package=magrittr (Nov. 2020).

184. Kassambara, A. Visualization of a Correlation Matrix using 'ggplot2' [R package ggcorrplot version 0.1.3]. https://CRAN.R-project.org/package=ggcorrplot (May 2019).

185. Auguie, B. Miscellaneous Functions for "Grid" Graphics [R package gridExtra version 2.3]. https://CRAN.R-project.org/package=gridExtra (Sept. 2017).

186. Companion to Applied Regression [R package car version 3.0-10]. https://CRAN.R-project.org/package=car (Sept. 2020).

187. Dynamic Documents for R [R package rmarkdown version 2.7]. https://CRAN.R-project.org/package=rmarkdown (Feb. 2021).

188. Xiao, N., Cao, D. S., Zhu, M. F. & Xu, Q. S. *Protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences* in *Bioinformatics* **31** (Oxford University Press, June 2015), 1857–1859. doi:10.1093/bioinformatics/btv042. https://pubmed.ncbi.nlm.nih.gov/25619996/.

189. Hornik, K., Buchta, C. & Zeileis, A. Open-source machine learning: R meets Weka. *Computational Statistics* **24,** 225–232. ISSN: 16139658 (May 2009).

190. Li, W. *et al.* Activation of GTP hydrolysis in mRNA-tRNA translocation by elongation factor G. *Science Advances* **1.** ISSN: 23752548. doi:10.1126/sciadv.1500169 (2015).

191. Luo, X. *et al.* Structural and Functional Analysis of the E. coli NusB-S10 Transcription Antitermination Complex. *Molecular Cell* **32,** 791–802. ISSN: 10972765 (Dec. 2008).

192. Komar, J. *et al.* Membrane protein insertion and assembly by the bacterial holo-translocon SecYEG-SecDF-YajC-YidC. doi:10.1042/BCJ20160545 (2016).

193. Schnirch, L. *et al.* Expanding the Depth and Sensitivity of Cross-Link Identification by Differential Ion Mobility Using High-Field Asymmetric Waveform Ion Mobility Spectrometry. *Analytical Chemistry* **92,** 10495–10503. ISSN: 15206882 (Aug. 2020).

194. Villa, E. *et al.* Ribosome-induced changes in elongation factor Tu conformation control GTP hydrolysis. *Proceedings of the National Academy of Sciences of the United States of America* **106,** 1063–1068. ISSN: 00278424 (Jan. 2009).

195. Hur, S. & Stroud, R. M. How U38, 39, and 40 of Many tRNAs Become the Targets for Pseudouridylation by TruA. *Molecular Cell* **26,** 189–203. ISSN: 10972765 (Apr. 2007).

196. McLean, J. E. *et al.* Inosine 5'-monophosphate dehydrogenase binds nucleic acids in vitro and in vivo. *Biochemical Journal* **379,** 243–251. ISSN: 02646021 (Apr. 2004).

197. Pausch, P. *et al.* Structural Basis for Regulation of the Opposing (p)ppGpp Synthetase and Hydrolase within the Stringent Response Orchestrator Rel. *Cell Reports* **32,** 108157. ISSN: 22111247 (Sept. 2020).

198. Kittendorf, J. D., Barcomb, L. M., Nonekowski, S. T. & Garcia, G. A. tRNA-guanine transglycosylase from Escherichia coli: Molecular mechanism and role of aspartate 89. *Biochemistry* **40,** 14123–14133. ISSN: 00062960 (Nov. 2001).

199. Reader, J. S., Metzgar, D., Schimmel, P. & De Crécy-Lagard, V. Identification of Four Genes Necessary for Biosynthesis of the Modified Nucleoside Queuosine. *Journal of Biological Chemistry* **279,** 6280–6285. ISSN: 00219258 (Feb. 2004).

200. Behrmann, E. *et al.* Structural snapshots of actively translating human ribosomes. *Cell* **161,** 845–857. ISSN: 10974172 (May 2015).

201. Walleczek, J., Redl, B., Stoffler-Meilicke, M. & Stoffler, G. Protein-protein cross-linking of the 50 S ribosomal subunit of Escherichia coli using 2-iminothiolane. Identification of cross-links by immunoblotting techniques. *Journal of Biological Chemistry* **264,** 4231–4237. ISSN: 00219258 (1989).

202. Allen, F. H. *et al.* Tables of bond lengths determined by x-ray and neutron diffraction. Part 1. Bond lengths in organic compounds. *Journal of the Chemical Society, Perkin Transactions 2,* S1–S19. ISSN: 1472779X (Jan. 1987).

203. Smith, T. J. MOLView: A program for analyzing and displaying atomic structures on the Macintosh personal computer. *Journal of Molecular Graphics* **13,** 122–125. ISSN: 02637855 (1995).

204. Michaelson-Richie, E. D. *et al.* Mechlorethamine-induced DNA-protein cross-linking in human fibrosarcoma (HT1080) cells. *Journal of Proteome Research* **10,** 2785–2796. ISSN: 15353893 (June 2011).

205. Muto, Y. & Yokoyama, S. *Structural insight into RNA recognition motifs: Versatile molecular Lego building blocks for biological systems* Mar. 2012. doi:10.1002/wrna.1107. https://pubmed.ncbi.nlm.nih.gov/22278943/.

206. Groehler, A., Degner, A. & Tretyakova, N. Y. *Mass Spectrometry-Based Tools to Characterize DNA–Protein Cross-Linking by Bis-Electrophiles* Sept. 2017. doi:10.1111/bcpt.12751.

207. Sayou, C. *et al.* RNA Binding by Histone Methyltransferases Set1 and Set2. *Molecular and Cellular Biology* **37.** ISSN: 0270-7306. doi:10.1128/mcb.00165-17. https://pubmed.ncbi.nlm.nih.gov/28483910/ (July 2017).

208. Karmakar, S., Purkait, K., Chatterjee, S. & Mukherjee, A. Anticancer activity of a cis-dichloridoplatinum(ii) complex of a chelating nitrogen mustard: Insight into unusual guanine binding mode and low deactivation by glutathione. *Dalton Transactions* **45,** 3599–3615. ISSN: 14779234 (Feb. 2016).

209. Johnson, K. A. *Role of induced fit in enzyme specificity: A molecular forward/reverse switch* Sept. 2008. doi:10.1074/jbc.R800034200. https://pubmed.ncbi.nlm.nih.gov/18544537/.

210. Tretyakova, N. Y., Groehler, A. & Ji, S. DNA-Protein Cross-Links: Formation, Structural Identities, and Biological Outcomes. *Accounts of Chemical Research* **48,** 1631–1644. ISSN: 15204898 (June 2015).

211. Kreuzer, K. N. DNA damage responses in prokaryotes: Regulating gene expression, modulating growth patterns, and manipulating replication forks. *Cold Spring Harbor Perspectives in Biology* **5.** ISSN: 19430264. doi:`10.1101/cshperspect.a012674`. `/pmc/articles/PMC3809575/%20/pmc/articles/PMC3809575/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3809575/` (Nov. 2013).

212. Phadtare, S. & Inouye, M. Sequence-selective interactions with RNA by CspB, CspC and CspE, members of the CspA family of Escherichia coli. *Molecular Microbiology* **33,** 1004–1014. ISSN: 0950382X (1999).

213. Luo, X. *et al.* Structural and Functional Analysis of the E. coli NusB-S10 Transcription Antitermination Complex. *Molecular Cell* **32,** 791–802. ISSN: 10972765 (Dec. 2008).

214. Liljas, A. & Sanyal, S. *The enigmatic ribosomal stalk* Jan. 2018. doi:`10.1017/S0033583518000100`. `https://pubmed.ncbi.nlm.nih.gov/30912488/`.

215. Valle, M. *et al.* Locking and unlocking of ribosomal motions. *Cell* **114,** 123–134. ISSN: 00928674 (July 2003).

216. Hebert, A. S. *et al.* Comprehensive Single-Shot Proteomics with FAIMS on a Hybrid Orbitrap Mass Spectrometer. *Analytical Chemistry* **90,** 9529–9537. ISSN: 15206882 (Aug. 2018).

217. Schnirch, L. *et al.* Expanding the Depth and Sensitivity of Cross-Link Identification by Differential Ion Mobility Using High-Field Asymmetric Waveform Ion Mobility Spectrometry. *Analytical Chemistry* **92,** 10495–10503. ISSN: 15206882 (Aug. 2020).

218. Schönbrunn, E. *et al.* The fungal product terreic acid is a covalent inhibitor of the bacterial cell wall biosynthetic enzyme UDP- N -acetylglucosamine 1- carboxyvinyltransferase (MurA). *Biochemistry* **49,** 4276–4282. ISSN: 00062960 (May 2010).

219. Golla, V. K. *et al.* Fosfomycin Permeation through the Outer Membrane Porin OmpF. *Biophysical Journal* **116,** 258–269. ISSN: 15420086 (Jan. 2019).

220. Xia, B., Etchegaray, J. P. & Inouye, M. Nonsense Mutations in cspA Cause Ribosome Trapping Leading to Complete Growth Inhibition and Cell Death at Low Temperature in Escherichia coli. *Journal of Biological Chemistry* **276,** 35581–35588. ISSN: 00219258 (Sept. 2001).

221. Garcin, E. D. *GAPDH as a model non-canonical AU-rich RNA binding protein* Feb. 2019. doi:`10.1016/j.semcdb.2018.03.013`.

222. Joyce, M. A. *et al.* Probing the nucleotide-binding site of Escherichia coli succinyl-CoA synthetase. *Biochemistry* **38,** 7273–7283. ISSN: 00062960 (June 1999).

223. Gaudet, P., Livstone, M. S., Lewis, S. E. & Thomas, P. D. Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Briefings in Bioinformatics* **12,** 449–462. ISSN: 14675463 (Sept. 2011).

224. Scott, J. M., Ju, J., Mitchell, T. & Haldenwang, W. G. The Bacillus subtilis GTP binding protein Obg and regulators of the σ(B) stress response transcription factor cofractionate with ribosomes. *Journal of Bacteriology* **182,** 2771–2777. ISSN: 00219193 (May 2000).

225. Luo, D., Leautey, J., Grunberg-Manago, M. & Putzer, H. Structure and regulation of expression of the Bacillus subtilis valyl- tRNA synthetase gene. *Journal of Bacteriology* **179,** 2472–2478. ISSN: 00219193 (1997).

226. Putzer, H., Brakhage, A. A. & Grunberg-Manago, M. Independent genes for two threonyl-tRNA synthetases in Bacillus subtilis. *Journal of Bacteriology* **172,** 4593–4602. ISSN: 00219193 (1990).

227. Xue, H., Xue, Y., Doublié, S. & Carter Jr., C. W. Chemical modifications of Bacillus subtilis tryptophanyl-tRNA synthetase. *Biochemistry and Cell Biology* **75,** 709–715. ISSN: 0829-8211 (Dec. 1997).

228. Xin, Y., Li, W. & First, E. A. The 'KMSKS' motif in tyrosyl-tRNA synthetase participates in the initial binding of tRNA(Tyr). *Biochemistry* **39,** 340–347. ISSN: 00062960 (Jan. 2000).

229. Saxild, H. H. & Nygaard, P. Gene-enzyme relationships of the purine biosynthetic pathway in Bacillus subtilis. *MGG Molecular & General Genetics* **211,** 160–167. ISSN: 00268925 (Jan. 1988).

230. Trapero, A. *et al.* Covalent inactivation of Mycobacterium thermoresistibile inosine-5'-monophosphate dehydrogenase (IMPDH). *Bioorganic & Medicinal Chemistry Letters* **30,** 126792. ISSN: 0960894X (Jan. 2020).

231. Hedstrom, L. IMP dehydrogenase: Structure, mechanism, and inhibition. *Chemical Reviews* **109,** 2903–2928. ISSN: 00092665 (July 2009).

232. Hazan, N. P. *et al.* Article Nucleosome Core Particle Disassembly and Assembly Kinetics Studied Using Single-Molecule Fluorescence. doi:10.1016/j.bpj.2015.07.004. http://dx.doi.org/10.1016/j.bpj.2015.07.004 (2015).

233. Baltz, A. G. *et al.* The mRNA-Bound Proteome and Its Global Occupancy Profile on Protein-Coding Transcripts. *Molecular Cell* **46,** 674–690. ISSN: 10972765 (June 2012).

234. Wilson, D. N. & Cate, J. H. The structure and function of the eukaryotic ribosome. *Cold Spring Harbor Perspectives in Biology* **4,** 5. ISSN: 19430264 (May 2012).

235. Takagi, M., Absalon, M. J., McLure, K. G. & Kastan, M. B. Regulation of p53 translation and induction after DNA damage by ribosomal protein L26 and nucleolin. *Cell* **123,** 49–63. ISSN: 00928674 (Oct. 2005).

236. Park, S. & Tretyakova, N. Structural Characterization of the Major DNA-DNA Cross-Link of 1,2,3,4-Diepoxybutane. *Chemical Research in Toxicology* **17,** 129–136. ISSN: 0893-228X (Feb. 2004).

13 Appendix

13.1 Appendix A

13.1.1 Emitted energy from UV-Lamps

In general,

$$E_{\text{total}} = Lumosity * distance * time \tag{2}$$

$$\text{fraction of light exposed} = \frac{\text{(cross-sectional area of sample)}}{\text{(imaginary shell around lamp at distance d of sample)}} \tag{3}$$

$$\text{fraction of light exposed} = \frac{\pi r^2}{4\pi d^2} \tag{4}$$

$$\text{fraction of light exposed} = \frac{1}{4}(\frac{r}{d})^2 \tag{5}$$

And:

$$E_{\text{total}} = L_{\text{lamp}} * \frac{1}{4}(\frac{r}{d})^2 t \tag{6}$$

Where L is the lumosity of the lamp, r is the cross-sectional radius of the sample, and d is the distance to the lamp. Now,

$$\frac{E_{\text{total}}}{A} = L_{\text{lamp}} \frac{t}{4\pi}(\frac{1}{d})^2 \tag{7}$$

Where A is the area. This way, the final units will be Jcm$_{\text{-2}}$. The effective lumosity is given by:

$$L_{\text{lamp}} = \text{ordinal lumosity} * \text{efficiency factor} \tag{8}$$

Where the ordinal lumosity = 8W and the efficiency factor = 0.85 according to the manufacturer. So, it all combines into:

$$\frac{E_{\text{total}}}{A} = L_{\text{lamp}} \frac{t}{4\pi}(\frac{1}{d})^2 * 0.85 \tag{9}$$

Equation 9 is used to calculate the amount of energy per cross-area emitted from the UV-lamps built into the custom-made UV-apparatus.

Different setups were trialed in this work, but ultimately setup number III was set as standard procedure:

I Duration of UV radiation: t1 = 2 min = 120 s , d = 2 cm

II Duration of UV radiation: t1 = 2 min = 120 s , d = 3 cm

III Duration of UV radiation: t1 = 10 min = 600 s , d = 2 cm

IV Duration of UV radiation: t1 = 10 min = 600 s , d = 3 cm

Exemplary calculation: Duration of UV radiation: t1 = 2 min = 120 s, d = 2 cm.

$$\frac{E_{\text{total}}}{A} = L_{\text{lamp}}\frac{t}{4\pi}(\frac{1}{d})^2 * 0.85 \tag{10}$$

$$\frac{E_{\text{total}}}{A} = 8W\frac{120s}{4\pi}(\frac{1}{d})^2 * 0.85 \tag{11}$$

$$\frac{E_{\text{total}}}{A} = 16.2\frac{J}{cm_{\text{-2}}} \tag{12}$$

Computing energies for the four different setups yields:

I 16.2 Jcm^{-2}

II 7.2 Jcm^{-2}

III 81.2 Jcm^{-2}

IV 36.1 Jcm^{-2}

13.1.2 Table of nucleotide specific mass shifts used in RNP$_{\mathrm{xl}}$ searches

Table 16: Nucleotide specific adducts from UV crosslinking. UV-induced adducts of different nucleotides correspond to a specific mass shift that are used by RNP$^{\mathrm{xl}}$ to identify nucleotide cross-linked peptides. Chemical forumulae of adduct compositions are used by RNP$^{\mathrm{xl}}$ to annotate the spectra which can be inspected in TOPPVIEW.

nucleotide	adduct formula	adduct composition	mass shift m/z	nucleic acid	crosslinker
U	$C_9H_{12}N_2O_6$	U-HPO_3	244.0695	RNA	UV
U	$C_4H_4N_2O_2$	U'	112.0273	RNA	UV
U	$C_4H_2N_2O$	U'-H_2O	94.01671	RNA	UV
U	C_3O	C_3O	51.99491	RNA	UV
U	$C_9H_{13}N_2O_9P$	U	324.0359	RNA	UV
U	$C_9H_{11}N_2O_8P$	U-H_2O	306.0253	RNA	UV
U	$C_9H_{10}N_2O_5$	U-H_3PO_4	226.0590	RNA	UV
G	$C_{10}H_{13}N_5O_5$	G-HPO_3	283.0917	RNA	UV
G	$C_5H_5N_5O$	G'	151.0494	RNA	UV
G	$C_5H_3N_5$	G'-H_2O	133.0388	RNA	UV
G	$C_5H_2N_2O$	G'-NH_3	106.0167	RNA	UV
G	$C_{10}H_{14}N_5O_8P$	G	363.0580	RNA	UV
G	$C_{10}H_{11}N_4O_8P$	G-NH_3	346.0315	RNA	UV
G	$C_{10}H_{12}N_5O_7P$	G-H_2O	345.0474	RNA	UV
G	$C_{10}H_{11}N_5O_4$	G-H_3PO_4	265.0811	RNA	UV
C	$C_9H_{13}N_3O_5$	C-HPO_3	243.0855	RNA	UV
C	$C_4H_5N_3O$	C'	111.0433	RNA	UV
C	$C_4H_2N_2O$	C'-NH_3	94.0167	RNA	UV
C	$C_4H_3N_3$	C'-H_2O	93.0327	RNA	UV
C	$C_9H_{14}N_3O_8P$	C	323.0519	RNA	UV
C	$C_9H_{11}N_2O_8P$	C-NH_3	306.0253	RNA	UV
C	$C_9H_{12}N_3O_7P$	C-H_2O	305.0413	RNA	UV
C	$C_9H_{11}N_3O_4$	C-H_3PO_4	225.0750	RNA	UV
A	$C_{10}H_{13}N_5O_4$	A-HPO_3	267.0968	RNA	UV
A	$C_5H_5N_5$	A'	135.0545	RNA	UV
A	$C_{10}H_{14}N_5O_7P$	A	347.0631	RNA	UV
A	$C_{10}H_{11}N_5O_3$	A-H_3PO_4	249.2265	RNA	UV
A	$C_{10}H_{12}N_5O_6P$	A-H_2O	329.0525	RNA	UV
A	$C_5H_2N_4$	A'-NH_3	118.0279	RNA	UV
A	$C_{10}H_{11}N_4O_7P$	A-NH_3	330.0365	RNA	UV

Table 17: Nucleotide specific adducts from 2-iminothiolane crosslinking. 2IT mediated crosslinking yields adducts of different nucleotides attached to the peptide by 2IT. Corresponding specific mass shifts are used by RNP^{xl} to identify nucleotide cross-linked peptides and the chemical forumulae of adduct compositions are used by RNP^{xl} to annotate the spectra. Annotated spectra can be inspected in TOPPVIEW. 2IT: 2-iminothiolane

nucleotide	formula of adduct mass	composition of added mass	mass shift m/z	crosslinker
None	C_4H_7NS	2IT	101.1701	2IT
Adenine	$C_{14}H_{18}N_6O_3$	2IT+A-H_3PO_4	350.3961	2IT
Adenine	$C_9H_{12}N_6$	2IT+A’	236.2968	2IT
Adenine	$C_{14}H_{21}N_6O_7PS$	2IT+A	448.3913	2IT
Adenine	$C_{14}H_{19}N_6O_6PS$	2IT+A-H_2O	430.3760	2IT
Adenine	$C_{14}H_{17}N_6O_5PS$	2IT+A-2H_2O	412.3607	2IT
Adenine	$C_{14}H_{17}N_6O_5S$	2IT+A-HPO_3	368.1267	2IT
None	C_4H_7NS	2IT	101.1701	2IT
Cytosine	$C_{13}H_{18}N_4O_4S$	2IT+C- H_3PO_4	326.3714	2IT
Cytosine	$C_8H_{12}N_4OS$	2IT+C’	212.0732	2IT
Cytosine	$C_8H_{10}N_4S$	2IT+C’- H_2O	194.0626	2IT
Cytosine	$C_{13}H_{21}N_4O_8PS$	2IT+C	424.0818	2IT
Cytosine	$C_{13}H_{19}N_4O_7PS$	2IT+C- H_2O	406.0712	2IT
Cytosine	$C_{13}H_{17}N_4O_6PS$	2IT+C-2 H_2O	388.0606	2IT
Cytosine	$C_{13}H_{18}N_4O_5S$	2IT+C- HPO_3	342.0998	2IT
None	C_4H_7NS	2IT	101.1701	2IT
Guanosine	$C_{14}H_{18}N_6O_4S$	2IT+G- H_3PO_4	366.1110	2IT
Guanosine	$C_9H_{12}N_6OS$	2IT+G’	252.0793	2IT
Guanosine	$C_9H_{10}N_6S$	2IT+G’H_2O	234.0688	2IT
Guanosine	$C_{14}H_{21}N_6O_8PS$	2IT+G	464.0879	2IT
Guanosine	$C_{14}H_{19}N_6O_7PS$	2IT+G-H_2O	446.0774	2IT
Guanosine	$C_{14}H_{17}N_6O_6PS$	2IT+G-2H_2O	428.0668	2IT
Guanosine	$C_{14}H_{20}N_6O_5S$	2IT+G-HPO_3	384.1216	2IT
None	C_4H_7NS	2IT	101.1701	2IT
Uracil	$C_{13}H_{18}N_4O_4S$	2IT+U-H_3PO_4	327.0889	2IT
Uracil	$C_8H_{11}N_3O_2S$	2IT+U’	213.0572	2IT
Uracil	$C_8H_9N_3OS$	2IT+U’-H_2O	195.0466	2IT
Uracil	$C_8H_7N_3S$	2IT+U’-2H_2O	177.0361	2IT
Uracil	C_7H_7NOS	2IT+C_3O	153.0248	2IT
Uracil	$C_{13}H_{20}N_3O_9PS$	2IT+U	425.0658	2IT
Uracil	$C_{13}H_{18}N_3O_8PS$	2IT+U-H_2O	407.0552	2IT
Uracil	$C_{13}H_{16}N_3O_7PS$	2IT+U-2H_2O	389.0447	2IT
Uracil	$C_{13}H_{19}N_3O_6S$	2IT+U-HPO_3	345.0995	2IT

Table 18: Nucleotide specific adducts from DEB crosslinking. DEB mediated crosslinking yields adducts of different nucleotides linked to the peptide by DEB. Corresponding specific mass shifts are used by RNP^{xl} to identify nucleotide cross-linked peptides and the chemical forumulae of adduct compositions are used by RNP^{xl} to annotate the spectra. Annotated spectra can be inspected in TOPPVIEW. DEB: 1,2,3,4-diepoxybutane

nucleotide	formula of adduct mass	composition of added mass	mass shift m/z	crosslinker
None	$C_4H_6O_2$	DEB	86.03678	DEB
None	C_4H_4	DEB-H_2O	68.07396	DEB
Adenine	$C_{14}H_{19}N_5O_5$	DEB+A-H_3PO_4	337.1386	DEB
Adenine	$C_9H_{11}N_5O_2$	DEB+A'	221.0913	DEB
Adenine	$C_9H_9N_5O$	DEB+A'-H_2O	203.0807	DEB
Adenine	$C_9H_7N_5O$	DEB+A'-2xH_2O	185.0701	DEB
Adenine	$C_{14}H_{22}N_5O_9P$	DEB+A	435.1155	DEB
Adenine	$C_{14}H_{20}N_5O_8P$	DEB+A-H_2O	417.1049	DEB
Adenine	$C_{14}H_{18}N_5O_7P$	DEB+A-2H_2O	399.0944	DEB
Adenine	$C_{14}H_{21}N_5O_6$	DEB+A-HPO_3	355.1492	DEB
None	$C_4H_6O_2$	DEB	86.03678	DEB
None	C_4H_4	DEB-H_2O	68.07396	DEB
Cytosine	$C_{13}H_{17}N_3O_6$	DEB+C-H_3PO_4	311.1117	DEB
Cytosine	$C_8H_{11}N_3O_3$	DEB+C'	197.0800	DEB
Cytosine	$C_8H_9N_3O_2$	DEB+C'-H_2O	179.0695	DEB
Cytosine	$C_8H_7N_3O$	DEB+C'-2xH_2O	161.0589	DEB
Cytosine	$C_{13}H_{20}N_3O_{10}P$	DEB+C	409.0886	DEB
Cytosine	$C_{13}H_{18}N_3O_9P$	DEB+C-H_2O	391.0781	DEB
Cytosine	$C_{13}H_{16}N_3O_8P$	DEB+C-2xH_2O	373.06756	DEB
Cytosine	$C_{13}H_{19}N_3O_7$	DEB+C-HPO_3	329.1223	DEB
None	$C_4H_6O_2$	DEB	86.03678	DEB
None	C_4H_4	DEB-H_2O	68.07396	DEB
Guanosine	$C_{14}H_{17}N_5O_6$	DEB+G-H_3PO_4	351.1179	DEB
Guanosine	$C_9H_{11}N_5O_3$	DEB+G'	237.0862	DEB
Guanosine	$C_9H_9N_5O_2$	DEB+G'-H_2O	219.0756	DEB
Guanosine	$C_9H_7N_5O$	DEB+G'-2xH_2O	201.0651	DEB
Guanosine	$C_{14}H_{20}NO_{10}P$	DEB+G	449.0948	DEB
Guanosine	$C_{14}H_{18}N_5O_9P$	DEB+G-H_2O	431.0842	DEB
Guanosine	$C_{14}H_{16}N_5O_8P$	DEB+G-2xH_2O	413.0736	DEB
Guanosine	$C_{14}H_{19}N_5O_7$	DEB+G-HPO_3	369.1284	DEB
None	$C_4H_6O_2$	DEB	86.03678	DEB
None	C_4H_4	DEB-H_2O	68.07396	DEB
Uracil	$C_{13}H_{16}N_2O_7$	DEB+U-H_3PO_4	312.0958	DEB
Uracil	$C_8H_{10}N_2O_4$	DEB+U'	198.0641	DEB
Uracil	$C_8H_8N_2O_3$	DEB+U'-H_2O	180.0535	DEB
Uracil	$C_8H_6N_2O_2$	DEB+U'-2xH_2O	162.0429	DEB
Uracil	$C_7H_6O_3$	DEB+C_3O	138.0317	DEB
Uracil	$C_{13}H_{19}N_2O_{11}P$	DEB+U	410.0726	DEB
Uracil	$C_{13}H_{17}N_2O_{11}P$	DEB+U-H_2O	408.0570	DEB
Uracil	$C_{13}H_{15}N_2O_{10}P$	DEB+U-2xH_2O	390.0464	DEB
Uracil	$C_{13}H_{18}N_2O_8$	DEB+U-HPO_3	330.1063	DEB

Table 19: Nucleotide specific adducts from NM crosslinking. NM mediated crosslinking yields adducts of different nucleotides linked to the peptide by nitrogen mustard. Corresponding specific mass shifts are used by RNP^{xl} to identify nucleotide cross-linked peptides and the chemical forumulae of adduct compositions are used by RNP^{xl} to annotate the spectra. Annotated spectra can be inspected in TOPPVIEW. NM: nitrogen mustard

Parent nucleotide	formula of adduct mass	composition of added mass	mass shift m/z	crosslinker
None	C_5H_9N	NM	83.07350	NM
Adenine	$C_{15}H_{20}N_6O_3$	NM+A-H_3PO_4	332.1597	NM
Adenine	$C_{10}H_{14}N_6$	NM+A’	218.1280	NM
Adenine	$C_{15}H_{23}N_6O_7P$	NM+A	430.1366	NM
Adenine	$C_{15}H_{21}N_6O_6P$	NM+A-H_2O	412.1260	NM
Adenine	$C_{15}H_{19}N_6O_5P$	NM+A-2xH_2O	394.1155	NM
Adenine	$C_{15}H_{22}N_6O_4$	NM+A-HPO_3	350.1703	NM
None	C_5H_9N	NM	83.07350	NM
Cytosine	$C_{14}H_{20}N_4O_4$	NM+C-H_3PO_4	308.1485	NM
Cytosine	$C_9H_{14}N_4O$	NM+C’	194.1168	NM
Cytosine	$C_9H_{12}N_4$	NM+C’-H_2O	176.1062	NM
Cytosine	$C_8H_7N_3O$	NM+C’-2xH_2O	161.0589	NM
Cytosine	$C_{14}H_{23}N_4O_8P$	NM+C	406.3282	NM
Cytosine	$C_{14}H_{21}N_4O_7P$	NM+C-H_2O	388.1148	NM
Cytosine	$C_{14}H_{19}N_4O_6P$	NM+C-2xH_2O	370.1042	NM
Cytosine	$C_{14}H_{22}N_4O_5$	NM+C-HPO_3	326.1590	NM
None	C_5H_9N	NM	83.07350	NM
Guanosine	$C_{15}H_{20}N_6O_5$	NM+G-H_3PO_4	348.1546	NM
Guanosine	$C_{10}H_{14}N_6O$	NM+G’	234.1229	NM
Guanosine	$C_{10}H_{12}N_6$	NM+G’-H_2O	216.1123	NM
Guanosine	$C_{15}H_{23}N_6O_8P$	NM+G	446.1315	NM
Guanosine	$C_{15}H_{21}N_6O_7P$	NM+G-H_2O	428.1209	NM
Guanosine	$C_{15}H_{19}N_6O_6P$	NM+G-2xH_2O	410.1104	NM
Guanosine	$C_{15}H_{22}N_6O_5$	NM+G-HPO_3	366.1652	NM
None	C_5H_9N	NM	83.07350	NM
Uracil	$C_{14}H_{19}N_3O_5$	NM+U-H_3PO_4	312.1559	NM
Uracil	$C_9H_{13}N_3O_2$	NM+U’	195.1008	NM
Uracil	$C_9H_{11}N_3O$	NM+U’-H_2O	177.0902	NM
Uracil	$C_9H_9N_3$	NM+U’-2xH_2O	159.0796	NM
Uracil	C_8H_9NO	NM+C_3O	135.0684	NM
Uracil	$C_{14}H_{22}N_3O_9P$	NM+U	438.0831	NM
Uracil	$C_{14}H_{20}N_3O_8P$	NM+U-H_2O	389.0988	NM
Uracil	$C_{14}H18N_3O_7P$	NM+U-2xH_2O	371.0882	NM
Uracil	$C_{14}H_{21}N_3O_6$	NM+U-HPO_3	327.1430	NM

13.1.3 Possible crosslinker adducts from rearrangment reaction.

The following figures 60 and 61 illustrate potential crosslinker adducts, from 1,2,3,4-diepoxybutane (DEB) and nitrogen mustard (NM), respectively. Adducts are steming from single reactions between the crosslinkers and a nucleophilic amino acid of a peptide. Whilst RNP^{xl} identifies true crosslinking events, mono adduct reactions leading to a single modification on the peptide also occurs frequently. Since both DEB and NM are highly reactive towards nucleophiles, the aqueous solution provides ample water molecules that quench the crosslinker by hydrolysis reactions. Open searches on MS1 precursor level with a tolerance window of 500 Da revealed that peptides were modified by other massess than true crosslinking events. Apart from neutral losses on the peptide, masses corresponding to the hypothesized adducts were identified. Required energies for dehydroxylation reactions involving DEB may be provided by high temperatures during the ionization process. Alternatively, high in-source temperatures also facilitate otherwise endergonic reactions by overcoming endothermic thresholds.

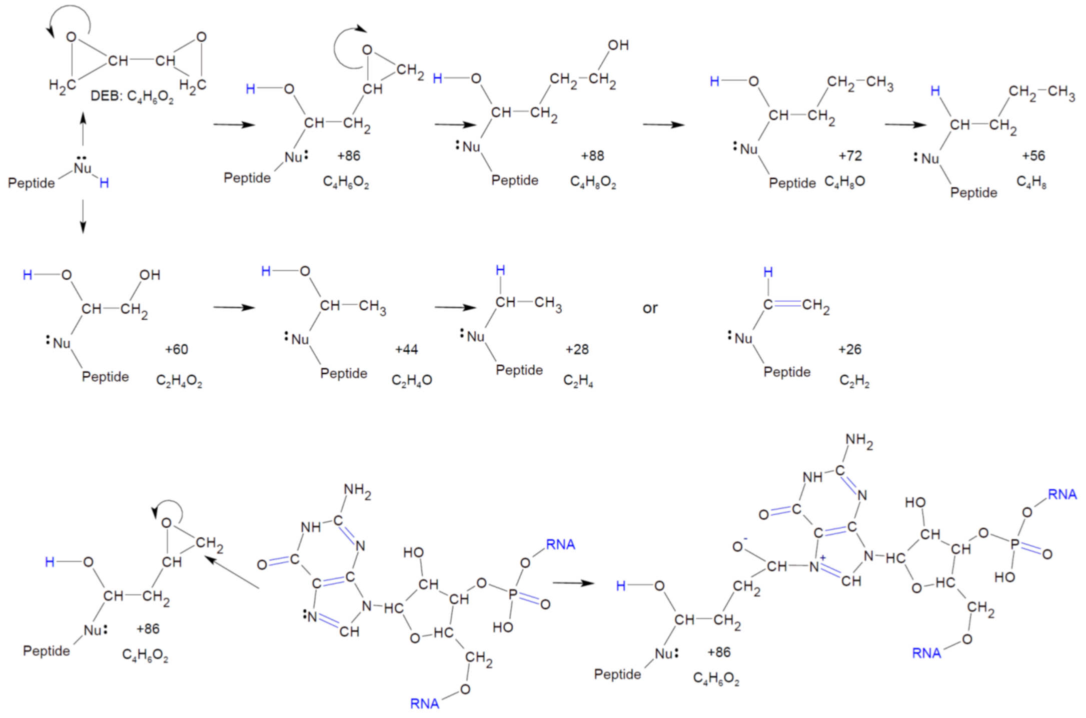


Figure 60: Chemical structures of potential DEB adducts. DEB can react with nucleophilic amino acids to generate mono cross-linked adducts akin to dead-end crosslinking events seen in protein-protein crosslinking. Due to its reactivity and the aqueous solvent used, proposed structures gravitate towards hydrolysis products. Required energies for some reactions may have been contributed by high temperatures during the ionization process in ESI. Nominal mass next to the structures denote the *m/z* shift that were identified in an open search on MS1 precrusor level. A true crosslinking event between peptide and nucleic acid is also shown at the bottom.

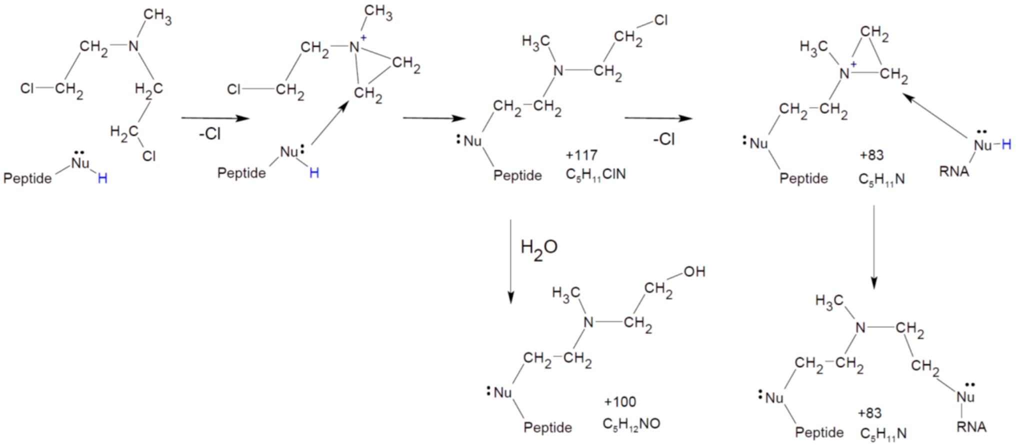


Figure 61: Chemical structures of potential NM adducts. Analogous to DEB, NM can react with nucleophilic amino acids to generate mono cross-linked adducts. These dead-end crosslinking events, usually seen in protein-protein crosslinking, also occur in protein-nucleic acid crosslinking. NM hydrolysis products were primiarily investigated, but high temperatures during the ionization process in ESI may foster the formation of additional adducts not depicted here. Nominal mass next to the structures denote the *m/z* shift that were identified in an open search on MS1 precrusor level. In addition to the dead-end reactions, a true crosslinking event between peptide and nucleotide is depicted at the bottom right.

13.2 Appendix B

13.2.1 Table of numeric features from RNP$_{xl}$ data

Table 20: List of numeric features used correlational analyses and in supervised machine learning approaches. Features 1-31 were part of the RNP^{xl} output file, but features 32-39 were computes using the appropriate R script. Note that all 39 features were numeric to allow building a k-NN classifyer. All features except peptide length, number of charged/polar/apolar amino acids, pI of the peptide, aliphatic index, hydrophobicity index, and molecular weight of the peptide were computed from the peptide sequence post RNP$_{xl}$ output.

Numeric features used in machine learning	
features 1-31 from RNP^{xl} output	features 32-39 from R-based computations
1 RT (retention time)	21 RNP$_{xl}$ peptide length Morpheus
2 Precursor	22 RNP$_{xl}$ peptide length error metric
3 Score (Percolator)	23 RNP$_{xl}$ peptide length immonium ion MIC
4 charge	24 RNP$_{xl}$ peptide length modification score
5 precursor error (ppm)	25 RNP$_{xl}$ peptide length precursor MIC
6 precursor intensity	26 RNP$_{xl}$ precursor score
7 OMS precursor mz error(ppm)	27 RNP$_{xl}$ adjusted score
8 RNP$_{xl}$ Da difference	28 RNP$_{xl}$ total MIC
9 RNP$_{xl}$ MIC (matched ion current)	29 RNP$_{xl}$ total loss score
10 RNP$_{xl}$ Morpheus score	30 RNP$_{xl}$ original score
11 RNP$_{xl}$ RNA MASS z$_0$ score	31 precursor intensity log$_{10}$
12 RNP$_{xl}$ best localization score	32 Peptide length
13 RNP$_{xl}$ error score	33 number of charged amino acids in peptide
14 RNP$_{xl}$ immonium score	34 number of polar amino acids in peptide
15 RNP$_{xl}$ marker ions score	35 number of apolar amino acids in peptide
16 RNP$_{xl}$ mass error peptide	36 pI of the peptide
17 RNP$_{xl}$ modification score	37 aliphatic index
18 RNP$_{xl}$ partial loss score	38 hydrophobicity index
19 RNP$_{xl}$ peptide mass z$_0$	39 molecular weight of the peptide
20 RNP$_{xl}$ peptide length MIC	

13.3 Appendix C

13.3.1 R-scripts and their pdf reports

The following list enumerates the R-scripts (in .Rmd format) and their corresponding report (in .pdf format) that were used in supervised machine learning approaches and can be found on the accompanying disc. The nomenclature is as follows: The session number is used to match .Rmd script with .pdf report. UV vs. DEB denotes the crosslinker used. RNA specifies which type of nucleic acid was classified. FAIMS/no_FAIMS is used to distinguish between FAIMS and non-FAIMS runs. S30/S100 denotes the sample type of *E. coli* cell extracts used as input sample.

1. Session4_UV_RNA_FAIMS_S30
2. Session5_UV_RNA_FAIMS_S100
3. Session6_UV_RNA_no_FAIMS_S30
4. Session7_UV_RNA_no_FAIMS_S100
5. Session6B_DEB_RNA_no_FAIMS_S30
6. Session7B_DEB_RNA_no_FAIMS_S100
7. Session9_DEB_RNA_FAIMS_S30
8. Session8_DEB_RNA_FAIMS_S100

13.3.2 Tables of CSMs for all pertinent experiments

CSM data was combined into different tables according to their respective experiments. The following tables are available on the accompanying CD:

1. X_positions_Hsh49_pNTs (section 6.2.1)
2. Hsh49_dinucleotide_preference (section 6.2.2)
3. X_positions_Hsh49_Cus1_U2s (section 6.2.3)
4. Dnmt2_Xlinks_results (section 6.3)
5. NELF_NM_TAR_results (section 6.5)
6. NELF_DEB_TAR_results (section 6.5)
7. HCD_HFX_Ecoli_all_NuXL (section 6.6)
8. HCD_targeted_Ecoli_all_NuXL (section 6.12)
9. FAIMS_Ecoli_all_NuXL (section 6.13)
10. Bacillus_subtilis_DEB_XL_NuXL (section 8)
11. HeLa_DEB_XL_NuXL (section 9)

14 Acknowledgements

A lot of people have continuously supported me over the years, helping me to grow professionally, as well as personally on this journey to a PhD in mass spectrometry. I thank all my friends, family, and colleagues wholeheartedly.

First of all, I would like to thank Prof. Dr. Henning Urlaub who gave me the opportunity work on such a fascinating project. Having been trained in proteomics and some PTM analysis, he convinced me to dip my toes into the water of protein-RNA crosslinking. I was bewildered to enter a world, where manual annotation of spectra was absolutely necessary and soon found it to be a really addicting thrill (once I detected my first crosslinking amino acids in a spectrum). I am very thankful for having worked with a lot of spectral data which has made me a better mass spectrometrist.

I thank Prof. Dr. Markus Bohnsack and Prof. Dr. Patrick Cramer for being part of my thesis advisory committee. The discussions were always very enlightening and steered the project into the right directions.

I want to extend my gratitude to Dr.Rosa Viner from Thermo Fisher Scientific who invited me to work with her at Thermo's mass spec facility in San Jose to record data using the novel FAIMS device when it was just released.
I would like to also thank Prof. Dr. Ralf Ficner for providing me the opportunity to have such as fruitful collaboration and model protein-RNA complex to study. Similarly, I am deeply grateful to Dr. Seychelle Vos who did not only provided me with an ideal protein-RNA complex to refine my method, but also long discussion about the findings and its implications.

I am so lucky and thankful for the GGNB program "Biomolecules: Structure - Function - Dynamics" that provided an excellent framework for my PhD with interesting courses to take, as well as an enjoyable atmosphere at GGNB retreats and get-togethers outside the lab. The non-scientific courses and the people I have met have helped me grow immensely as a person.

I am incredibly lucky to have had such excellent mentors such as Dr. Kundan Sharma who taught me the basics of protein-RNA crosslinking in record time before he was unfortunately leaving the group, and Dr. Alexandra Stützer who supported me after Kundan had left. Our scientific conversations helped immensely in directing the project, as well as your guidance in learning how to purify proteins quickly and efficiently. Thank you so much for teaching me the 2% of illustrator that I now know to generate nicer figures.

And of course, a special "thank you" to Luisa M. Welp, the best Master student one could ask for who grew into a much valued colleague whom I can always count on. Working with you has been pure bliss and I have learned a lot from you - more than the occasional trick in Illustrator. Thank you so much!

I am very grateful to the scientific input from Dr. Aleksandar Chernev in discussing this project and proofreading this thesis. Additionally, I am grateful for having worked with Dr. Kuan-Ting Pan with whom I embarked on an exciting journey to Thermo Fischer Scientific in San Jose. Our mission to learn the newest technological developments in mass spectrometry felt so much more manageable with him around being able to seamlessly explain everything. Our discussions late at night have expanded my appreciation of mass spectrometry even more.

I would also thank Dr. Andreia Cepeda for proofreading my thesis and helping me wherever she could. The same is true for Dr. Momchil Ninov. I am also very grateful to have worked with such wonderful people in the lab: Monika Raabe, Dr. Ralf Pflanz, Dr. Sabine König, and Uwe Pleßmann have not only provided a nice atmosphere, but also showed me more than one little trick in the lab that always fixed the problem. I would not dare to forget to thank the rest of the Bioanalytical Mass Spectrometry group for Having provided an excellent atmosphere both scientifically and interpersonally: Iwan, Fanni, Sofia, Ivan, Pan, Juliane, Pan, Christof, Yalong, Christian, Sunit, and Olex.

Additionally, I would like to thank the other members of the "5-Freunde": Marcel, Malte, Jan and Sebastian who have reminded me that there is a life outside of the lab and who have supported me for over twenty years by being the best friends one could ask for. Similarly, I would like to thank my friends Jesse, Sam, and Ada who live across the ocean but never let that prevent them from supporting me however they could.

Lastly, I would like to thank my family for always having been there all the way through my time as a PhD student.

Publications pertinent to this thesis

Johannsson, S., Neumann, P., **Wulf, A.**, Welp, L., Gerber, H., Krull, M., Diederichsen, U., Urlaub, H. and Ficner, R., 2018. Structural insights into the stimulation of S. pombe Dnmt2 catalytic efficiency by the tRNA nucleoside queuosine. Scientific Reports, **8**.

Sonnleitner, E., **Wulf, A.**, Campagne, S., Pei, X., Wolfinger, M., Forlani, G., Prindl, K., Abdou, L., Resch, A., Allain, F., Luisi, B., Urlaub, H. and Bläsi, U., 2017. Interplay between the catabolite repression control protein Crc, Hfq and RNA in Hfq-dependent translational regulation in Pseudomonas aeruginosa. Nucleic Acids Research, **43**, pp.1470-1485.

Stützer, A., Welp, L., Raabe, M., Sachsenberg, T., Kappert, C., **Wulf, A.**, Lau, A., David, S., Chernev, A., Kramer, K., Politis, A., Kohlbacher, O., Fischle, W. and Urlaub, H., 2020. Analysis of protein-DNA interactions in chromatin by UV induced crosslinking and mass spectrometry. Nature Communications, **11**.

Özcan, A., Pausch, P., Linden, A., **Wulf, A.**, Schühle, K., Heider, J., Urlaub, H., Heimerl, T., Bange, G. and Randau, L., 2018. Type IV CRISPR RNA processing and effector complex formation in Aromatoleum aromaticum. Nature Microbiology, **4**, pp.89-96.