CHEMICAL CROSSLINKING EXTENDS AND COMPLEMENTS UV CROSSLINKING IN ANALYSIS OF RNA/DNA NUCLEIC ACID–PROTEIN INTERACTION SITES BY MASS SPECTROMETRY

AUTHORS

Luisa M. Welp[1,2,†], Timo Sachsenberg[3,4,†], Alexander Wulf[1,†], Aleksandar Chernev[1,†], Yehor Horokhovskyi[5,†], Piotr Neumann[6], Martin Pašen[5], Arslan Siraj[3,4], Monika Raabe[1], Sven Johannsson[6], Jana Schmitzova[7], Eugen Netz[3,4], Julianus Pfeuffer[3,4], Yi He[9], Kai Fritzemeier[10], Bernard Delanghe[10], Rosa Viner[9], Seychelle M. Vos[11,12], Patrick Cramer[7], Ralf Ficner[6], Juliane Liepe[5,*], Oliver Kohlbacher[3,4,8,*], Henning Urlaub[1,2,*].

[1] Bioanalytical Mass Spectrometry Group, Max Planck Institute for Multidisciplinary Sciences, 37077 Göttingen, Germany
[2] Bioanalytics Group, Department of Clinical Chemistry, University Medical Center Göttingen, Göttingen, 37075, Germany
3 Institute for Bioinformatics and Medical Informatics, University of Tübingen, Tübingen, 72076, Germany
[4] Applied Bioinformatics, Dept. for Computer Science, University of Tübingen, Tübingen, 72076, Germany
[5] Quantitative and Systems Biology, Max Planck Institute for Multidisciplinary Sciences, Göttingen, 37077, Germany
[6] Department of Molecular Structural Biology, Institute of Microbiology and Genetics, GZMB, Georg-August-University Göttingen, Göttingen, 37077, Germany
[7] Max Planck Institute for Multidisciplinary Sciences, Department of Molecular Biology, Göttingen, 37077, Germany
[8] Institute for Translational Bioinformatics, University Hospital Tübingen, Tübingen, 72076, Germany
[9] Thermo Fisher Scientific, San Jose, CA, 95134, USA
[10] Thermo Fisher Scientific, Bremen, 28199, Germany
[11] Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA
[12] Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA
* To whom correspondence should be addressed. Tel: +49 551 201-1060; Fax: +49 551 201-1197; Email: henning.urlaub@mpinat.mpg.de. Correspondence may also be addressed to Juliane Liepe (Tel: +49 551 201-1471; Email: juliane.liepe@mpinat.mpg.de) or Oliver Kohlbacher (Tel: +49 7071 29 70457; Email: oliver.kohlbacher@uni-tuebingen.de).

† Joint Authors

ABSTRACT

UV (ultra-violet) crosslinking with mass spectrometry (XL-MS) has been established for identifying RNA- and DNA-binding proteins along with their domains and amino acids involved. Here, we explore chemical XL-MS for RNA-protein, DNA-protein, and nucleotide-protein complexes *in vitro* and *in vivo*. We introduce a specialized nucleotide-protein-crosslink search engine, NuXL, for robust and fast identification of such crosslinks at amino acid resolution. Chemical XL-MS complements UV XL-MS by generating different crosslink species, increasing crosslinked protein yields in vivo almost four-fold and thus it expands the structural information accessible via XL-MS. Our workflow facilitates integrative structural modelling of nucleic acid–protein complexes and adds spatial information to the described RNA-binding properties of enzymes, for which crosslinking sites are often observed close to their cofactor-binding domains. *In vivo* UV and chemical XL-MS data from *E. coli* cells analysed by NuXL establish a comprehensive nucleic acid–protein crosslink inventory with crosslink sites at amino acid level for more than 1500 proteins. Our new workflow combined with the dedicated NuXL search engine identified RNA crosslinks that cover most RNA-binding proteins, with DNA and RNA crosslinks detected in transcriptional repressors and activators.

INTRODUCTION

Proteins and RNA (or DNA) form functional complexes termed ribonucleoproteins, RNPs (or deoxyribonucleoproteins, DNPs), which drive and organize key cellular processes. For studying RNA-protein interactions, UV crosslinking combined with mass spectrometry (UV XL-MS) has been established as a standard method, in which the identification of crosslinked RNA-protein heteroconjugates is based either on the inherent reactivity of pyrimidines (mainly uracil) or on UV-reactive derivatives, 4-thiouridine or 6-thioguanine(1-8). In earlier studies, we demonstrated that UV XL-MS can identify the sites of RNA/DNA-protein crosslinking at amino acid resolution and the compositions of the crosslinked (oligo)nucleotide adducts in samples of varying complexity(2,9-12). Other compelling MS-based studies of the RNA-protein interactome in cells have contributed to this research field by establishing dedicated workflows for UV XL-MS. RBDmap and pCLAP protocols have been used to identify RNA-binding proteins and binding domains by identification of peptides neighbouring crosslinked peptides using sequential proteolytic digestion(13,14). Similarly, the CAPRI protocol identifies RNA-crosslinked proteins and peptides neighbouring crosslinked peptides(5). XRNAX and OOPS use UV crosslinking with organic phase separation to identify RNA-crosslinked proteins and peptides by MS(7,8). TRAPP uses silica-based enrichment of UV-crosslinked proteins and for its variant iTRAPP, resolution is increased to the amino acid level(6). Using (p)RBS-ID, Bae *et al.* established chemical digestion of the RNA in UV-crosslinked samples to generate mononucleosides, thus facilitating identification of crosslinked peptides by minimizing the database search space of the MS crosslinking data(3,4). CLIR-MS identifies UV-crosslinked amino acids and the crosslinked nucleotide in the RNA sequence by using stable isotope-labelled RNA *in vitro*(15).

In DNP analysis, UV XL-MS is much less well-established, owing to the reduced UV reactivity of double-stranded DNA compared to RNA(16,17). Alternatively, DNP XL-MS studies use chemicals that react with purines in the DNA(18-23). Chemical crosslinking has also been applied successfully in several biochemical studies of RNA-protein interactions(24), but barely combined with larger XL-MS studies although the potential of alkylating reagents for protein-RNA and DNA crosslinking has been systematically investigated by Scalabrin *et al.* in 2018(25).

To increase coverage of interaction sites we set out to complement UV XL-MS by introducing chemical crosslinking with alkylating agents (mechlorethamine/nitrogen mustard, NM; 1,2:3,4-diepoxybutane, DEB; and formaldehyde, FA), identifying the crosslinking sites, and to compare the results with those from UV XL-MS. We developed NuXL, a specialized, open-source cross-link search tool, with integrated false-discovery rate (FDR) control at the crosslinked spectrum match (CSM) level, and extensive reporting of cross-link results tailored for the analysis of protein-nucleic acid complexes. We demonstrate the feasibility of using XL-MS with NuXL for the integrative modelling of RNPs *in vitro,* based on the combination of UV and chemical crosslink sites with high-resolution structural data from cryo-EM and crystallography. Moreover, using NuXL we could identify more than 1,500 proteins with their chemically or UV-crosslinked amino acids to RNA, DNA, or single nucleotides in *E. coli*. The chemically and UV-induced crosslinking sites identified on proteins confirm that chemical crosslinking complements and extends UV crosslinking at the amino acid and nucleotide levels.

MATERIAL AND METHODS

UV, DEB, NM and FA *E. coli* ribosome crosslinking and sample processing

100 µg of NEB ribosomes purified from *E. coli* (NEB; P0763S) were diluted to 100 µL with 50 mM Hepes/NaOH pH 7.5 prior to crosslinking. For UV XL, samples were pipetted into 96-well polypropylene solid bottom microplates (Greiner Bio-One, 655209) and irradiated for 10 min at a 2 cm distance from four parallel UVC Germicidal T5 lamps (Sylvania; G8W; 0000501; λ=254 nm) as part of an in-house built crosslinking apparatus. DEB crosslinking was performed at a final concentration of 50 mM DEB (Sigma; 202533) for 10 min, 37°C, 300 rpm. NM crosslinking was performed at a final concentration of 1 mM NM (Sigma; 122564) for 10 min, 37°C, 300 rpm. FA crosslinking was performed at 1% [v/v] FA (Thermo Scientific™, 28906) for 10 min, 37°C, 300 rpm. All chemical crosslinking reactions were quenched with 50 mM Tris/HCl pH 7.5 and incubation for 5 min, RT, 300 rpm. After crosslinking and quenching, samples were EtOH-precipitated (75% [v/v] EtOH, 0.3 M NaOAc) and pellets were washed once with 80% [v/v] EtOH. Pellets were dissolved in 45 µL 8 M urea in 50 mM Tris/HCl pH 7.5. Urea concentration was diluted to 1 M using 50 mM Tris/HCl pH 7.5. 1mM $MgCl_2$, 250 U Pierce™ Universal nuclease for Cell Lysis (Thermo Scientific™, 88700), 100 U nuclease P1 (NEB, M0660S) were added, following incubation for 3 h, 37°C, 300 rpm. Trypsin was added at a 1:20 enzyme-to-ribosome

ratio (2.5 µg; Promega, V5111) following incubation ON, 37°C, 300 rpm. Samples were acidified with 2-5 µL 10% [v/v] formic acid to pH 5. 100 U nuclease P1 were added, following incubation for 30 min, 37°C, 300 rpm. 10% [v/v] formic acid and 100% acetonitrile (ACN) were added to reach a final pH of 2 and ACN concentration of 2%. Nucleotide removal was performed using C18 stage tips (Harvard Apparatus; 74-4107) according to manufacturer's instructions. Briefly, columns were equilibrated with each 400 uL of 100% ACN; 50% [v/v] ACN, 0.1% [v/v] formic acid; and three times 0.1% [v/v] formic acid. Samples were loaded twice and washed two times with 0.1% [v/v] formic acid and once with 5% [v/v] ACN, 0.1% [v/v] formic acid. Elution was performed using twice 150 µL of 50% [v/v] ACN, 0.1% [v/v] formic acid and once 100 µL of 80% [v/v] ACN, 0.1% [v/v] formic acid. All centrifugation steps were performed for 2 min at 600*xg*. Peptides were dried in a speed vac concentrator. Enrichment of crosslinked peptide-(oligo)nucleotides was performed as described in (2) using Titansphere TiO_2 5 nm beads (GL Sciences; 5020-75000). Buffer A contained 5% [v/v] glycerol in buffer B; buffer B contained 80% [v/v] ACN, 5% [v/v] trifluoroacetic acid (TFA); buffer B2 contained 60% [v/v] ACN and 0.1% [v/v] TFA; buffer C contained 0.3 M NH_4OH, pH 10.5. Columns were self-packed by adding approximately 3 mg TiO_2 beads in 100 % MeOH to p10 pipette tips equipped with a coffee filter tip. For each replicate, two columns were used. Columns were equilibrated with 60 µL of each 100% MeOH, buffer B and buffer A. Dried samples were dissolved in 60 µL buffer A and loaded twice and washed three times with buffer A, three times with buffer B and once with buffer B2. Elution was performed in three steps of each adding 40 µL buffer C. Elution tubes contained 20 µL of 10% [v/v] FA for pH neutralization. All centrifugation steps were performed for 2-5 min at 900*xg*, to prevent columns from running dry. Eluted crosslinked peptide-(oligo)nucleotides were dried in a speed vac concentrator. Once dry, the samples were dissolved first in 10% [v/v] ACN, 0.1% [v/v] formic acid and diluted to 2% [v/v] ACN, 0.1% [v/v] formic acid and subjected to LC-MS/MS measurements.

Purification of recombinant Hsh49 and Cus1 from *S. cerevisiae*

Recombinant protein complex Cus1-HSH49 was co-expressed from the pETDuet vector in Rosetta II in 2xYT medium ON at 18˚C. Purification was performed as described with some modifications (26). Cells were lysed in 50 mM HEPES-NaOH, pH 7.5, 400 mM NaCl, 500 mM urea, 15% glycerol, 5 mM 2-Mercaptoethanol, 20 mM imidazole supplemented with protease inhibitors EDTA free (Roche) using sonication. 5kU Pierce™ Universal nuclease for Cell Lysis

(Thermo Scientific™, 88700) were added, following incubation for 15 min on ice. Cell debris was removed by centrifugation 17.000 rpm at 4˚C for 1 h (A27 rotor; Thermo Scientific™). The supernatant was loaded on a 5-ml His Trap HP Ni-NTA column (GE Healthcare) equilibrated with lysis buffer A1. Proteins were eluted with a linear imidazole gradient of 20 to 400 mM in buffer B1 (50 mM HEPES-NaOH, pH 7.5, 200 mM NaCl, 400 mM imidazole, 500 mM urea, 10% glycerol, 5 mM 2-Mercaptoethanol) and the main peak fractions containing both proteins was collected and further processed. His-tag was removed by ON TEV-digestion and sample was loaded on the pre-packed His Trap HP Ni-NTA column (5ml, GE Healthcare). The flow through was collected and loaded on a pre-packed Heparin HP column (3x5ml, GE Healthcare). The protein complex was eluted with linear gradient of 0-100% of 20 mM HEPES-NaOH, pH 7.5, 1M NaCl, 10% glycerol, 5 mM 2-Mercaptoethanol. The peak fractions containing purified proteins were pooled and concentrated. Hsh49-Cus1-RNA complexes were reconstituted on ice for 30 min at a protein:RNA molar ratio of 4:1.

Purification of *S. pombe* Dnmt2 and in vitro transcription of tRNA^{Asp}

Saccharomyces pombe Dnmt2 (spDnmt2) was expressed and purified as described (27). Briefly, GST-tagged spDnmt2 was expressed in BL21(DE3) *E. coli* cells, cells were lysed with 50 mM Tris/HCl pH 7.5, 150 mM NaCl, 1 M LiCl, 2 mM DTT following ultracentrifugation at 50.000*xg* for 30 min. Soluble protein fraction was loaded onto Glutathione Sepharose FastFlow (GE Healthcare). GST-spDnmt2 was eluted from the column using 30 mM reduced glutathione. GST-tag was cleaved and removed by heparin-sepharose (GE Healthcare) purification and spDnmt2 was further purified by Superdex S75 size exclusion chromatography (Cytiva™). sptRNA^{Asp} *in vitro* trasncription was performed as described(27). 3.75 nmol Dnmt2 and 3.75 nmol tRNA^{Asp} were incubated in 100 µl 20 mM HEPES, 50 mM NaCl 2 mM DTT, pH 7.5 for 30 min on ice for reconstitution of the complex.

Purification of human NELF complex

The human NELF complex and NELF ΔA, ΔE and ΔAE tentacle mutants was purified as previously described(28). An RNA corresponding to the HIV-1 stem loop was purchased from Integrated DNA technologies with the following sequence 5'-CCAGAUCUGAGCCUGGGAGCUCUCUGG-3'. The RNA was folded by incubating of 50 µM of the RNA in a buffer containing 100 mM NaCl, 20 mM HEPES/NaOH pH 7.4, 3 mM $MgCl_2$,

incubating at 95˚C for 3 min, and transferring to ice for 10 min. NELF (~19 µM final) and RNA (20-30 µM) were mixed, and transferred to a 20 K MWCO Slide-A-Lyzer mini-dialysis cup (Thermo Scientific™) and dialyzed against a buffer containing 100 mM NaCl, 20 mM HEPES/NaOH pH 7.4, 3 mM $MgCl_2$ for 16 h at 4˚C. Fluorescence anisotropy was performed as previously described(28).

UV and chemical crosslinking of Hsh49-Cus1-snRNA, Dnmt2-tRNA^{Asp} and NELF *in vitro* reconstituted complexes and sample processing

Samples containing reconstituted Hsh49-Cus1-snRNA, Dnmt2-tRNA^{Asp} or NELF complex were crosslinked and processed as described above for *E. coli* ribosomes with modifications in crosslinking and digestion conditions. For Hsh49-Cus1 UV crosslinking, samples were irradiated as described above, following EtOH precipitation and washing as described for ribosomes. For DEB and NM crosslinking, 50 mM DEB or 1 mM NM were added, following incubation at 37°C for 30 min and precipitation as described for UV crosslinked samples. Samples were dissolved in 4 M urea, 50 mM Tris/HCl pH 7.5, and diluted to reach 1 M urea, 50 mM Tris/HCl pH 7.5. 2 µg of each Ambion™ RNase A (Invitrogen™, AM2270) and RNase T1 (Invitrogen™, AM2283) were added, following incubation at 37°C for 2 h. Afterwards, 1 mM $MgCl_2$, and 750 U of Benzonase® Nuclease (Millipore, 70664) were added, following incubation at 37°C, for another 2 h. Trypsin digestion, nucleotide removal using C18 stage tips, and crosslinked peptide-(oligo)nucleotide enrichment were performed as described above for *E. coli* ribosomes. For Dnmt2-tRNA^{Asp} UV-crosslinking, samples were irradiated as described above with (UV trypsin replicate 1, UV chymotrypsin replicate 1,2) and without (UV trypsin replicate 2) consecutive acetone-precipitation as described for Hsh49-Cus1 samples. DEB crosslinking was performed using 50 mM DEB and either 4 h incubation on ice (DEB trypsin replicate 1), or for 1 h at 37°C (DEB trypsin replicate 2, chymotrypsin replicate 1,2). NM crosslinking of Dnmt2-tRNA^{Asp} was performed using 10 mM NM and incubation for 2 h at 37°C (NM trypsin replicate 1) or 1 h at 37°C (NM trypsin replicate 2, NM chymotrypsin replicate 1,2). DEB and NM crosslinked Dnmt2-tRNA^{Asp} complexes were acetone precipitated using four volumes of ice-cold acetone and after ON incubation at -20°C washed with 80% ice-cold acetone. Subsequently, samples were digested in 1 M urea, 50 mM Tris/HCl pH 7.5, 1 mM $MgCl_2$, using 1 µg of each Ambion™ RNase A, 1 µg of RNase T1, and 250 U Benzonase® Nuclease HC (Millipore, 71205) for 2 h, at 37°C. Trypsin or chymotrypsin (Promega, V1061)

were added at a 1:20 enzyme-to-protein ratio following incubation ON at 37°C, 300 rpm. The rest of the protocol is identical to what is described for purified *E. coli* ribosomes. The reconstituted NELF-TAR complex was crosslinked and TAR-RNA element was digested as described for Hsh49-Cus1 complex. All crosslinked peptide-(oligo)nucleotide samples were dissolved in 2% [v/v] ACN 0.05% [v/v] TFA for LC-MS/MS measurements.

UV and chemical crosslinking of purified human nucleosomes

1200 µg of purified human nucleosomes (BPS Bioscience) were dialyzed using MF-Millipore® Membrane Filter (0.45 µm pore size; Millipore; HAWP04700) against 10 mL Hepes/NaOH pH 7.5. Volumes were adjusted to 2 mL and 250 µL (150 µg nucleosomes) were used for each replicate. For UV XL, samples were adjusted to 300 µL and split into each three 100 µL aliquots. Crosslinking was performed as described above for *E. coli* ribosomes. DEB, NM, and FA crosslinking were performed using 50 mM DEB, 1 mM NM, or 1% FA, respectively, and incubation for 10 min, 37°C, 300 rpm. For quenching, 50 mM Tris/HCl pH 7.5 was added following incubation for 5 min, RT, 300 rpm. Crosslinked protein-nucleic acid conjugates were EtOH precipitated. DNA was digested in 1 M urea, 50 mM Tris/HCl pH 7.5 with 1mM $MgCl_2$, 500 U Pierce™ Universal nuclease for Cell Lysis (Thermo Scientific™, 88700), 200 U nuclease P1 (NEB, M0660S) and incubation for 3 h, 37°C, 300 rpm. Protein digest, nucleotide removal and TiO_2 based peptide-DNA-(oligo)nucleotide enrichment was performed as described above for *E. coli* ribosomes. Two TiO_2 columns were used per replicate. Peptides were dried and dissolved like *E. coli* ribosomes for LC-MS/MS measurements.

***In vivo* UV and chemical crosslinking of *E. coli* cells**

E. coli BL21 DE3 cells were grown in LB medium to $OD_{600} \approx 0.6$. 15 mL of cell culture (corresponding to roughly 7 x 10^9 cells) were used as starting material per biological replicate for UV, DEB and NM crosslinking. UV, DEB and NM crosslinking were performed in medium. Cells were UV crosslinked for 10 min as described above in glass petri dishes (d = 12 cm, vol. cell suspension ≈ 11 mL, ht. cell suspension ≈ 1 mm) on an ice-cold metal block for two replicates or at RT for one replicate. DEB crosslinking was performed using 50 mM DEB for 10 min, 37°C, 300 rpm. NM crosslinking was performed with 10 mM NM for 10 min, 37°C, 300 rpm. DEB and NM crosslinking reactions were quenched with final 50 mM Tris/HCl pH 7.5 and incubation for 5 min, RT, 300 rpm. FA crosslinking was performed at 1% [v/v] FA for 10 min,

37°C, 300 rpm, followed by quenching using 400 mM Tris/HCl pH 7.5 for 10 min, RT, 300 rpm. Cells were pelleted at 3000*xg* for 10 min and the medium was removed. Cell pellets were either fractionated (see below) or stored at -80 °C. For UV, DEB and NM crosslinked cells, S30/S100 fractionation by ultracentrifugation was performed as follows: Cell pellets were dissolved in 100 µL of 20 mM Tris/HCl pH 7.6, 100 mM NH_4Cl, 10.5 mM $MgCl_2$, 3 mM β-mercaptoethanol, 0.4 mg/mL lysozyme, 10 U/mL RQ1 DNase (Promega, M6101) and cOmplete™ EDTA-free protease inhibitor cocktail (Roche, COEDTAF-RO). Cells were first incubated at 37°C for 30 min and subsequently mechanically disrupted by sonication in a water bath for 15 min, in alternating 30 s 'ON' and 30 s 'OFF' cycles at the highest intensity setting (Bioruptor, Diagenode). Three volumes of B-PER™ bacterial protein extraction reagent (Thermo Scientific™, 78243) were added and samples were briefly mixed by inverting, following incubation at RT for 5 min and another mixing by inverting. After another 5 min incubation at RT, samples were centrifuged at 40,000*xg* for 30 min for S30 fractions, and for another 30 min at 330,000*xg* for S100 fractions. S30 and S100 fractions were EtOH precipitated as described above, pellets were washed and dissolved in 8 M urea, 50 mM Tris/HCl pH 7.5. LysC (Promega, V1671) was added at a 1:100 enzyme-to-protein ratio, following incubation at 25°C, ON, 300 rpm. Urea concentration was diluted to 1 M using 50 mM Tris/HCl pH 7.5. Trypsin was added at a 1:50 enzyme-to-protein ratio, following incubation ON, 30°C, 300 rpm. Samples were EtOH precipitated as described above and dissolved in 4 M urea, 50 mM Tris/HCl pH 7.5. Urea concentration was diluted to 1 M using 50 mM Tris/HCl pH 7.5 and 1mM $MgCl_2$, 500-625 U Pierce™ Universal nuclease, 200-300 U nuclease P1 were added, following incubation for 3 h, 37°C, 300 rpm. Trypsin was added at a 1:50 enzyme-to-protein ratio, following incubation ON, 37°C, 300 rpm. Samples were acidified with formic acid to pH 3-4 and 2% [v/v] ACN was added. Sample clean-up was performed using C18 stage tips as described above. Enrichment of crosslinked peptide-(oligo)nucleotides was performed using 6 to 8 TiO_2 columns as described above. Eluted crosslinked peptide-(oligo)nucleotides were dried in a speed vac and fractionated by high pH reversed phase fractionation (see below).

FA crosslinked *E. coli* cells were resuspended in 300 µL PBS containing 9 mg of lysozyme, following incubation for 5 min, RT. 8 M urea, 100 mM HEPES-NaOH pH 7.9, 20 mM EDTA was added and cells were mechanically lysed by sonication, in alternating 30 s on- and 30 s off-cycles at highest intensity output (Bioruptor, Diagenode). Samples were diluted to 1 M urea and trypsin was added at an estimated 1:20 enzyme-to-protein ratio, following incubation at

37°C, ON, 300 rpm. 5 ml of cell lysate was used for silica enrichment using RNeasy Maxi Kit (Qiagen, 75162) according to manufacturer's instructions. Each centrifugation step was performed at 3.000*xg*. Briefly, 3.8 volumes of RLT Buffer, supplemented with β-mercaptoethanol and 2.8 volumes of EtOH were added to the cell lysate. Samples were loaded and washed once with RW1 Buffer, and twice with RPE Buffer. Peptides were eluted twice with H_2O and 1 M urea, 12.5 mM HEPES-NaOH pH 7.9 were added. Trypsin was added at an estimated 1:20 enzyme-to-protein ratio, following incubation for 2 h at 37°C, 300 rpm, following another round of RNA purification using RNeasy Maxi Kit as described above. Samples were adjusted to 1M urea, 12.5 mM HEPES-NaOH pH 7.9. 1mM $MgCl_2$, 1mM $ZnCl_2$, 1250 U Pierce™ Universal nuclease, 50 U RNase I, 300 U nuclease P1, 2 U RNase T1, and 200 U RNase A were added, following incubation at 37°C, ON, 300 rpm. Sample clean-up was performed using C18 stage tips as described above. Peptides were resuspended in 50 mM NaOAc and 80 U nuclease P1 and 4 U Antarctic phosphatase were added, following incubation at 30°C, 2 h, 300 rpm. Sample clean-up was performed using self-made C18 columns (p200 tips containing three layers of Empore™ C18 SPE Disks (Supelco, 66883-U)). Eluted FA crosslinked peptide-(oligo)nucleotides were dried in a speed vac concentrator, dissolved in 2% [v/v] ACN 0.05% [v/v] TFA, and subjected to LC-MS/MS measurements.

High pH reversed phase sample fractionation

Enriched UV-, DEB-, or NM-crosslinked peptides from *E. coli* were dissolved in 35 µL 10 mM NH_4OH pH 10, 5% [v/v] ACN. Peptides were loaded onto an Xbridge C18 column (Waters, 186003128) using an Agilent 1100 series chromatography system. A flow rate of 60 µl/min was set up using a buffer system consisting of 10 mM NH_4OH pH 10 (buffer A) and 10 mM NH4OH pH10, 80% [v/v] ACN (buffer B). The column was equilibrated with 5% buffer B and developed over 64 min using the following gradient: 5% buffer B (0-7 min), 8-30% buffer B (8-42 min), 30-50% buffer B (43-50 min), 90-95% buffer B (51-56 min), 5% buffer B (57-64 min). The first 6 min were collected as flow-throμgh fraction, followed by 48 x 1 min fractions, which were reduced to 12 fractions by concatenated pooling. Fractions were dried in a speed vac concentrator and dissolved like *E. coli* ribosomes for LC-MS/MS measurements.

LC-MS/MS analyses

All crosslinked peptide-(oligo)nucleotide samples were injected onto a C18 PepMap100-trapping column (0.3 x 5 mm, 5 µm, Thermo Scientific™) connected to an in-house packed C18 analytical column (75 µm x 300 mm; Reprosil-Pur 120C18-AQ, 1.9 µm, Dr Maisch GmbH). Columns were equilibrated using 98% buffer A (0.1% [v/v] formic acid), 2% buffer B (80% [v/v] ACN, 0.08% [v/v] formic acid). Liquid chromatography was performed using an UltiMate-3000 RSLC nanosystem (Thermo Scientific™). Purified, crosslinked *E. coli* ribosome samples and *E. coli* cell UV, DEB and NM XL-MS samples were analysed for 58 min using a linear gradient (10% to 45% buffer B (80% [v/v] ACN, 0.1% [v/v] formic acid in 44 min). Hsh49-Cus1-snRNA and NELF complex samples were analysed for 58 min using a linear gradient from 12% to 50% buffer B in 43 min. Dnmt1-tRNA^{Asp} samples were analysed for 58 min using a linear gradient from 12% to 50% buffer B in 43 min or from 2% to 48% buffer B in 44 min. FA XL-MS *E. coli* samples were analysed for 118 min using a linear gradient from 8% to 44% buffer B in 104 min. HeLa nucleosome samples were analysed for 88 min using a linear gradient (10% to 45% buffer B (80% [v/v] ACN, 0.1% [v/v] formic acid in 74 min). All measurements ended with an approximately 5 min high-percentage buffer B washing step. Eluting peptides were analysed on an Orbitrap Exploris 480 (Thermo Scientific™; *E. coli* ribosomes, UV, DEB and NM XL-MS *E. coli* cells, HeLa nucleosomes), Q-Exactive HF (Thermo Scientific™; Dnmt2-tRNA^{Asp}, NELF complex), Q-Exactive HF-X (Thermo Scientific™; Hsh49-Cus1-snRNA, Dnmt2-tRNA^{Asp}, NELF complex, UV XL-MS *E. coli* cells), Orbitrap Fusion (Thermo Scientific™; Dnmt2-tRNA^{Asp}, FA XL-MS *E. coli* cells), or Orbitrap Fusion Lumos (Thermo Scientific™; Dnmt2-tRNA^{Asp}) instrument. The following MS settings were used for *E. coli* ribosome, UV, DEB and NM crosslinked *E. coli* cell and HeLa nucleosome samples: MS1 scan range, 350–1600 *m/z*; MS1 resolution, 120,000 FWHM; AGC target MS1, 1E6; maximum injection time MS1, 60 ms; intensity threshold, 1E4; isolation window, 1.6 Th; normalized collision energy, 30% (DEB, NM, FA) or 28% (UV); charge states, 2+ to 6+; dynamic exclusion, 9 s; Top 20 most abundant precursors were selected for fragmentation; MS2 resolution, 30,000, AGC target MS2, 1E5; maximum injection time MS2, 120 ms. For Hsh49-Cus1-snRNA and NELF complex samples, the same settings as above were used, except for: MS1 scan range, 380–1580 *m/z*; maximum injection time MS1, 104 ms; isolation window, 1.4 Th; normalized collision energy, 30%; charge states, 2+ to 7+; AGC target MS2, 5E5; maximum injection time MS2, 104 ms. For Dnmt2-tRNA^{Asp} samples, the same

settings as for *E. coli* ribosomes were used, except for: MS1 scan range, 380-1580/350-1600 *m/z*; maximum injection time MS1, 50 ms; intensity threshold, 1E4/5E4; isolation window, 1.4/1.6 Th; normalized collision energy, 30%; dynamic exclusion, 3/7 s; Cycle time (top speed method), 3 s or Top20 method; maximum injection time MS2, 120/128 ms. For FA XL-MS *E. coli* samples, the same settings as for *E. coli* ribosomes were used, except for: MS1 scan range, 350–1580 *m/z*; MS1 resolution, maximum injection time MS1, 50 ms; intensity threshold, 2E5; isolation window, 1.4 Th; normalized collision energy, 36%; dynamic exclusion, 20 s; Cycle time (top speed method), 3 s; AGC target MS2, 1.1E5; maximum injection time MS2, 80 ms.

XL-MS data analysis using NuXL

E. coli ribosome, Hsh49-Cus1, Dnmt2 and NELF *in vitro* complexes, HeLa nucleosome*, E. coli in vivo* MS raw data were analysed with the NuXL node in the OpenMS open-source C++ library (version OpenMS-3.0.0-pre-HEAD-2022-04-27-Win64) or in Proteome Discoverer using the following databases: i) the reference proteome canonical sequence database for *E. coli* strain downloaded from the UniProt repository on 10.03.2022 (4400 protein sequences) including MaxQuant contaminants (version 2.0.3.0.) for *E. coli* ribosome and *E. coli in vivo* MS data, ii) Hsh49-Cus1 protein sequences for Hsh49-Cus1 data, iii) the reference proteome canonical sequence database for *E. coli* strain downloaded from the UniProt repository on 10.03.2022 (4400 protein sequences) plus MaxQuant contaminants (version 2.0.3.0.) including *S. pombe* Dnmt2 protein sequence for Dnmt2 data, iv) NELF subunit protein sequences for NELF data, and v) the reference proteome canonical sequence database for *H. sapiens* downloaded from the UniProt repository on 02.03.2021 (20381 protein sequences) plus MaxQuant contaminants (version 2.0.3.0.) for HeLa nucleosome data. The following presets were selected: *E. coli* ribosome, RNA-UV (UCGA), RNA-DEB, RNA-NM, RNA-FA; Hsh49-Cus1, RNA-UV (UCGA), RNA-DEB, RNA-NM; Dnmt2, RNA-UV (UCGA), RNA-DEB, RNA-NM; NELF, RNA-UV (UCGA), RNA-DEB, RNA-NM; HeLa nucleosome, DNA-UV, DNA-DEB, DNA-NM, DNA-FA; *E. coli in vivo*, RNA-UV (UCGA), RNA-DEB, RNA-NM, RNA-FA, DNA-UV, DNA-DEB, DNA-NM, DNA-FA. Other parameters were set to default or as follows: enzyme, Trypsin/P or Chymotrypsin/P (only for Dnmt2-tRNA^{Asp} datasets derived from chymotryptic cleavage); max. missed cleavages, 2; min. peptide length, 5; max. peptide length, 30; peptideFDR, 0.01; nucleotide adduct length, 2 (for XL-MS datasets from *E. coli* cells) or 3 (for all other datasets); dynamic modifications, Oxidation (Met).

Label-free quantitative MS data analysis using Maxquant

Proteomics data from *E. coli* cells were analysed with MaxQuant (version 2.0.3.0) using the reference proteome canonical sequence database for *E. coli* strain downloaded from the UniProt repository on 10.03.2022 (4400 protein sequences) including MaxQuant contaminants (version 2.0.3.0.). Search parameters were set to default or as follows: enzyme, Trypsin/P; max. missed cleavages, 2; dynamic modifications, Oxidation (Met), Acetlyation (protein N-term); static modification, Carbamidomethylation (Cys).

Downstream data analysis

Barplots were generated using OriginPro 2019 (version 9.6.0.172), violin plots were generated using Python (version 3.10.3). Protein 3D structures were plotted in PyMOL Molecular Graphics System (version 2.5.4) using Python (version 3.10.3) scripts for crosslink mapping and distance calculations. Figure 1 was created with BioRender.com.

spDnmt2-tRNA^{Asp} docking model generation

The model for spDnmt2-tRNA^{Asp} was created as previously described(29). The specific procedure is detailed in the following sections.

tRNA^{Asp} modeling

Atomic models of *S. pombe* tRNA^{Asp} have been predicted *de novo* using RNA modelling protocol as implemented in ROSETTA 2022.11 suite (released on March the 13th, 2022, https://www.rosettacommons.org/software)(30). This protocol utilizes a parameter-optimized version of Fragment Assembly of RNA with Full Atom Refinement (FARFAR) known as FARFAR2(30). Predicted tRNA secondary structure in dot-parens notation, that is necessary for the modelling process, has been obtained from the RNApdbee server (http://rnapdbee.cs.put.poznan.pl/)(31). This server requires a three-dimensional (3D) atomic model as input. Hence, *S. pombe* tRNA^{Asp} has been manually generated in Coot based on the *E. coli* tRNA^{Asp} crystal structure (PDB: 6ugg)(32). Subsequent regularization of the initially modeled *S. pombe* tRNA^{Asp} was performed in ERRASER(33).

Over 1100 de novo modeled S. pombe tRNA^{Asp} decoys were generated using Rosetta's FARFAR2 protocol, utilizing fragments derived from the regularized model of *S. pombe* tRNA^{Asp}. These decoys were subsequently clustered and sorted according to Rosetta's energy score.

Clustering revealed significant structural variations, particularly in the anticodon arm containing the modified cytosine at position 38, as well as less pronounced differences in the TΨC-arm, acceptor stem, and discriminator base. Two tRNA models (decoys 3 and 38, when sorted based on Rosetta's score), representing cluster centroids of two clusters with the most deviated anticodon arm conformations compared to the ERRASER-regularized *S. pombe* $tRNA^{Asp}$ initial model, were used for the majority of docking experiments. This approach was intended to mimic the presumed and yet unknown extent of anticodon arm deformation required for productive binding to Dnmt2 methyltransferase (with decoy 3 showing moderate and decoy 38 the strongest structural alterations). Interestingly, the model with the best Rosetta energy score (decoy number 1) exhibited a tRNA conformation identical to the *E. coli* $tRNA^{Asp}$ crystal structure and thus to the initially modeled *S. pombe* $tRNA^{Asp}$. This $tRNA^{Asp}$ decoy was used during initial tests to set up the calculation pipeline, involving the computation of several thousand docking models.

spDnmt2 modelling

The atomic model of *S. pombe* Dnmt2 (PDB: 6fdf, chain A) was used as one of three protein templates for docking calculations. The other two protein models, which differ in the conformation of a flexible loop (residues 79-96) near the active site, were generated using Rosetta loop modelling(34). This approach was chosen to minimize the risk of failing to obtain a productive tRNA-protein docking solution due to potential clashes between $tRNA^{Asp}$ and the loop region, which could be influenced by crystal packing. Notably, residues 79-96, comprising the flexible loop, were modeled only in one (chain A) of the four Dnmt2 monomers in the crystallographic asymmetric unit, highlighting their high flexibility, which is likely necessary for tRNA binding.

$tRNA^{Asp}$ and spDnmt2 docking experiments

Molecular in silico docking experiments yielding 3D models of *S. pombe* $tRNA^{Asp}$-Dnmt2 complexes have been performed using Rosetta fold-and-dock method optimized for predicting RNA-protein (RNP) complexes(30,35). This protocol combines RNA folding with RNA-protein docking protocol, which is both time and computationally intensive. Thus, two $tRNA^{Asp}$ models (cluster centroids, decoys 3 and 38), differing in conformation of the anticodon arm region, were docked to three Dnmt2 models varying in the conformation of a flexible loop

(residues 79:96). Specifically, $tRNA^{Asp}$ decoy number 38 (tRNA-38) was docked to Dnmt2 model 1 (Dnmt2-1, PDB: 6fdf, chain A) and model 2 (remodeled loop number 2), while $tRNA^{Asp}$ decoy number 3 (tRNA-3) was docked to Dnmt2 model 3 (Dnmt2-3, remodeled loop number 3) and Dnmt2 model 1. Both blind docking (docking a ligand to the entire surface of a protein without prior knowledge of the target pocket) and site-constrained docking experiments were conducted. The latter, which accounted for about 75% of all decoys, utilized knowledge of the Dnmt2 binding site defined by the SAH (S-adenosyl-L-homocysteine) ligand present in the crystal structure, as well as biochemical evidence identifying cytosine 38 as the modification site on tRNAAsp. This prior knowledge was encoded in the docking experiments as Euclidean distances between cytosine 38 and protein residues 75 (18 Å), 304 (21 Å), and 73 (20 Å), which are located near the methyl-donating SAH moiety. These spatial constraints were incorporated using a flat harmonic function, penalizing models only if the Euclidean distance constraint exceeded the specified threshold. Additional parameters included a flat bottom width of 2 Å and a standard deviation of 1 Å. In total, over 94,700 decoys were generated.

Verification of spDnmt2-$tRNA^{Asp}$ docking results based on XL-MS data

Almost 45.000 modelled *S. pombe* $tRNA^{Asp}$-Dnmt2 complexes, selected based on Rosetta's energy score, have been individually analysed in terms of their compatibility with crosslinking data using a protocol adopted from (cite https://doi.org/10.1093/nar/gkad1171). This was achieved using a set of Bash and AWK shell scripts, optimized for execution in a multi-core environment with GNU parallel (https://www.gnu.org/software/parallel/). The same procedure was applied to each set of crosslinking data (with minimum localization scores of ≥1 and ≥2) as well as to the subsets of UV crosslinks and chemical crosslinks.

The crosslinking dataset was filtered for a CSM count of ≥3 per crosslink site. Each crosslink was then assessed for compatibility with every $tRNA^{Asp}$-Dnmt2 model by calculating Euclidean distances between the protein residue (known position) and the linked nucleotide (unknown position). This process identified single (one protein residue-nucleotide pair) and multiple (several pairs) crosslink occurrences, which were counted and recorded. The Euclidean distance thresholds were based on the crosslinker type: UV (10 Å), DEB, and NM (15 Å). Distances were measured between atom N3 (nucleotides U or C) or N7 (other nucleotides) of tRNA and the Cα-position of the crosslinked protein residue. The number of crosslinker occurrences within these thresholds was termed "occurrences_d0" (with 0 indicating no slack, i.e., ideal distances).

To account for protein and tRNA flexibility that was not modeled during docking, additional "protein residue-nucleotide" pairs were checked by increasing the threshold distances by slacks of 2, 3, and 5 Å. These assessments were termed d2, d3, and d5, respectively. For each of the 44,694 $tRNA^{Asp}$-Dnmt2 complexes, sets of single and multiple crosslinker occurrences (d0, d2, d3, d5) were obtained and subjected to further analysis.

To identify a set of $tRNA^{Asp}$-Dnmt2 models supporting the highest number of crosslinks, we analyzed a table comprising individual decoys with both ideal and distance-slacked single and multiple crosslink occurrences. Initially, the decoys were filtered based on Rosetta's energy score, retaining 21,911 models (the top 80% best scoring decoys). These were further filtered using a spatial distance constraint of 23 Å between cytosine 38 and the Cα-positions of protein residues 73, 75, and 304 (spatial constraints from docking experiments). This ensured that only biologically relevant decoys, with the anticodon arm region near the active site, were considered.

The decoys were then sorted in ascending order based on single and multiple crosslink occurrences (d0, d2, d3), resulting in six individually sorted lists of decoys. These lists were truncated to retain a maximum 100 top-scoring models. In the final step, the six truncated lists were analyzed to identify common decoys across all lists. These decoys were selected as the best set of $tRNA^{Asp}$-Dnmt2 models compatible with the highest number of crosslinks. This approach aimed to avoid selecting decoys based on a single distance threshold, which could lead to incorrect assignments. For example, a crosslinked protein residue located near several nucleotides of the same type within a given threshold might artificially increase the score. Thus, the best decoys should consistently appear among the top solutions across different slacks (d0, d2, d3) and crosslink occurrences - single (one best protein-tRNA pair) and multiple.

The analysis revealed that both crosslinking datasets (minimum localization score ≥1 and ≥2) and their subsets comprising only chemical crosslinks produced two sets of decoys representing a similar binding mode of $tRNA^{Asp}$ to Dnmt2. Notably, the dataset with a minimum localization score ≥2 yielded a set of $tRNA^{Asp}$-Dnmt2 models that exclusively shared the same binding mode. In contrast, the dataset with a minimum localization score ≥1 also produced alternative binding modes, albeit less populated. In conclusion, analyzing the two crosslinking datasets filtered with varying stringency for localization scores serves as a cross-validation approach.

Protein annotation sources

The protein sequences were retrieved from UniProt on 2022-03-10(36) and UniProt identifiers were used to transfer the information from additional sources. Domain annotation was downloaded from InterPro(37). In addition to that, single-residue annotation of conserved and functionally important residues was obtained using the InterProScan software version 5.56-89.0 on the above mentioned FASTA protein sequences, with residue annotation mode enabled and minimal feature size set to 1(38). Manual review identified 202 relevant InterPro signatures that were manually grouped into 20 categories, corresponding to the molecule class that the protein feature is interacting with (i.e. ATP, FAD/NAD/NADP). Secondary structure was predicted from protein sequences using the S4PRED(39) using the default settings. Residue-level features were transferred from the DESCRIBEPROT(40) and the continuous scores were binarized into feature and non-feature using the default database thresholds.

Protein abundance in ppm for E.coli K12 strain and integrated abundance across strains were downloaded from the PaxDb(41). In addition to the pre-computed reference abundance in parts per million (ppm), the MS1 intensities derived from label-free quantification of the non-crosslinked control *E. coli* dataset were converted into ppm using the PaxDb web-interface. The annotation of RNA-related proteins was derived based on RBP2GO database of putative RNA-binding proteins(42).

Sampling and statistical testing of feature enrichment

Statistical testing was performed to quantify the enrichment of amino-acid (AA) composition, secondary structure and other residue features. The foreground was defined as a set of crosslinked peptide residues, either exactly at the crosslink site or belonging to the whole peptide. On the site level, only those crosslink-spectrum matches (CSMs) with the localization threshold >= 1 were kept. The number of peptides in the foreground set corresponds to the number of the CSMs per identified sequence. The background was derived by an *in silico* tryptic cleavage of the proteome. When comparing the crosslinked residue distributions, one random residue per background peptide was chosen.

Since not all the possible tryptic peptides are observable in the MS experiment and 85% of identified experimental sequences were previously reported in PeptideAtlas(43) and PaxDb(41), a mass-spectrometry-informed background was defined as a union of all the experimental sequences and all the tryptic *E. coli* peptides from PeptideAtlas(43) and PaxDb(41), and was used to sample the background distribution, omitting the non-detectable peptides. A global background consisting of all the MS-detectable peptides was used for the 3D distance enrichment analysis, whilst for feature count enrichment analysis, samples were drawn with replacement from the background distribution to match the experimental distribution of peptide lengths and missed cleavages. This propensity score-based covariate matching was done via matchRanges(44) using the "nearest" matching option.

Quantification and correction of the protein abundance bias

In order to assess whether the crosslink detection is favouring more abundant proteins we correlated the available protein abundance metrics (below) to the RBP2GO score and the number of observed CSMs in the crosslinking data. In most cases we observed a moderate 0.42-0.53 Pearson correlation values between the number of CSMs and protein abundance and 0.48-0.54 when comparing the number of CSMs to the RBP2GO score. When comparing the abundance distributions of the proteins annotated in RBP2GO, the RNA-related proteins appeared to be on average more abundant that other proteins. From this we concluded that the measured protein abundance could be confounded by the protein abundance.

We have modified the sampling scheme to compensate for the abundance bias, by associating every peptide with 4 metrics of the source protein abundance: PaxDB integrated (ppm) and K12 reference abundance (ppm), log10-transformed label-free MS1 intensity of our non-crosslinked *E. coli* proteome control and quantile-normalised ppm values derived from the label-free MS1 intensity of our non-crosslinked *E. coli* proteome control. A combination of multiple features was chosen in order to obtain a balanced abundance estimate informed by both in-house and public data and cover the most proteins. The abundance-corrected background resulted in comparable amino acid enrichment values, but the frequency of putative RNA-binding residues was reduced, supporting the hypothesis of RNA-related proteins being naturally high-abundant in *E. coli*.

Statistical testing for residue feature enrichment

The difference of the distribution mean feature counts per residue between the foreground and background was assessed via a t-test (45) as implemented in R language. In addition, since most of the protein features are distributed not normally, non-parametric goodness-of-fit tests(46) (Anderson-Darling and Cramer-von Mises) are used to validate the statistical significance. Furthermore, since many features are discrete, the discrete adaptations of nonparametric tests are used (47,48). The significance threshold for the feature enrichment was set to the adjusted p-value of 0.01 (BH method(49)) to be surpassed by both the t-test and non-parametric goodness-of-fit test. Sampling was repeated 10 times and the resulting statistics were aggregated by mean. The log2FoldChanges between the foreground and background distribution mean feature counts per residue were formatted into a matrix and plotted via the "pheatmap" R library(50).

3D distance analysis

The protein structures were downloaded from the AlphaFold Protein Structure Database (51,52) and 3D Euclidean distances were retrieved between the pairs of Cα atoms using the Xwalk(53) version 0.6 with maximal distance 100 Å. All the amino acids with min. localisation score ≥ 1 were checked for the presence of annotated protein features in the 10 Å radius. Since some of the features of interest occur infrequently, the background was not sampled from the proteome, but a full set of all annotated residues was used as a background distribution. A combination of a t-test and nonparametric goodness-of-fit-test was applied to test for enrichment, same as for the feature counts enrichment.

Over-representation analysis

The STRING database was used to define the gene sets for *E. coli*. All the identified proteins with at least 1 unique peptide were grouped into sets corresponding to different crosslinking methods each(54). The hypergeometric test for term over-representation was done via the "fgsea" R package(55). A union set of every protein identified in any of RNA-XL datasets or in the non-crosslinked control dataset was used as an enrichment background. The minimal gene set size was set to 4 and maximal to 1000. After testing and p.value adjustment with the Benjamini-Hochberg method, a set of non-redundant terms was derived(49). All the terms that

pass the significance threshold of 0.05 p.adj BH were tested to not enrich on the background of each other, using the "fgsea" package implementation.

Gene set over-representation heatmaps

The over-represented STRING terms were selected for heatmap display as long as they passed the p.adj BH significance threshold of 0.01 except the Domains set, where it was set to 0.05. In addition, a query and background set overlap size of at least 5 proteins was required. Upon the filtering, the manually selected top representative terms were visualised using the "pheatmap" R library(50).

Protein feature and coverage plots

We have compiled a residue-level annotation of E. coli proteome and visualised the co-localization of protein features and RNA-XL coverage. Tryptic cleavage sites correspond to Lys and Arg residues in the proteins. Protein annotation was converted into per-residue GenomicRanges format using the protein coordinates and used for further plotting(56).

All the CSMs with the localization threshold >= 1 and their source peptides were plotted for all RNA-XL datasets except FA, where no localization information is available. The positions of RNA-XL sites and peptide coverage were plotted along the protein sequence together with the domain and residue annotation(56-61).

RESULTS

Assessment of MS2 spectra and crosslinked species from UV and chemical XL-MS

Our first goal was to identify DEB-, NM- and FA-induced crosslink sites at the amino-acid and nucleotide level in MS analysis. To this end, purified *E. coli* ribosomes were subjected to *in vitro* UV, DEB, NM, and FA crosslinking, nuclease and trypsin digestion, followed by C18 clean-up for removal of non-crosslinked (oligo)nucleotides and TiO_2 bead-based peptide-RNA (oligo)nucleotide enrichment followed LC-MS analysis (Figure 1A). We manually annotated MS1 and MS2 data, applying known mass-adduct settings for UV and calculating the mass of hypothetical crosslinker-RNA (oligo)nucleotide adducts of various compositions, based on the detected mass adducts and on previous studies suggesting possible reaction mechanisms

(Figure 1A, Supplementary Figure S1, Supplementary Table S1)(2,62-65). Individual mass adducts on the precursor and fragment ions and on several marker ions (crosslinked immonium, nucleobase, and nucleotide ions) were assembled and guided the development of a fast, reliable FDR-controlled computational analysis tool for identifying crosslinking sites, which we named NuXL.

Using NuXL, we determined the occurrence of predominant crosslink types by comparison of CSM counts (Figure 1B). For UV, the majority of peptide-RNA (oligo)nucleotides showed crosslinks of Tyr to uridines, in agreement with earlier studies(66). Lys was frequently crosslinked to either uracil with neutral loss of H_2O, or cytidine with loss of NH_3 (these products are isobaric; Figure 1C; Supplementary Table S2). Cytosine and uracil crosslinks are distinguishable when no neutral loss of NH_3 or H_2O is identified. Other crosslinked amino acids identified in UV-crosslinked ribosomes are Met, Cys, Ile, Thr, and Leu. While most UV crosslinks involved uracil or cytosine, fewer CSMs matched with peptides crosslinked to adenosine or guanosine (Figure 1B; Supplementary Table S2). Unlike UV XL-MS, CSMs from DEB and NM XL-MS ribosome data identified mainly peptides crosslinked to purines. 75% (DEB) and 67% (NM) of unique crosslink site-nucleotide pairs (unique crosslinks) involved guanine, and 20% (DEB) and 21% (NM) adenine. Only few crosslinks were observed to pyrimidines. For both DEB and NM XL-MS, more than 75% of unique crosslinks identified were to Cys, Met or His. In addition, crosslinks to Lys, Arg or Glu (with DEB) and Lys, Ser, Thr, Ala, Leu, Glu, Phe, Pro, Tyr, or Asp (with NM) were identified (Figure 1B; Supplementary Table S2). MS2 spectra of His crosslinks show prominent immonium ions of His with mass adducts of the crosslinker remnant and/or crosslinker remnant plus the crosslinked nucleobase/nucleotide of adenine and guanine (Figure 1C). For peptides with Met crosslinked by DEB/NM, we frequently observed specific loss of C_2H_2S from Met on precursors and fragment ions (Figure 1C). As a result, the MS2 spectra contained high-intensity marker ions comprising DEB/NM, the crosslinked nucleobase/nucleotide plus C_2H_2S. Furthermore, marker ions of only DEB/NM plus C_2H_2S were observed.

In MS2, spectra of FA crosslinks differed fundamentally from those of DEB and NM. FA MS2 spectra revealed specific and highly abundant marker ions (FA plus nucleobase/nucleotide) and comprehensive peptide sequence coverage by unmodified fragment ions (Figure 1C). The absence of crosslinked peptide fragment ions, presumably due to the unstable peptide-FA bond, simplified the MS2 spectra and improved database search efficiency, but prevents

assignment of the crosslinked amino acid(67). Therefore, FA presets contained only a small number of fragment adducts (Supplementary Table S1). 70% of unique FA crosslinks involved adenine, followed by 15% guanine, and 15% cytosine (Figure 1B; Supplementary Table S2).
In conclusion, chemical XL-MS using DEB, NM, or FA, adds complementary information by crosslinking amino acids mainly to purines, while UV crosslinks amino acids (mainly) to pyrimidines. The incorporation of assembled crosslink mass adducts within dedicated presets in NuXL facilitates rapid, FDR-controlled assignment of crosslinking sites.

NuXL – a nucleic acid–protein XL-MS search engine for automated and sensitive crosslink site identification

Our novel software tool NuXL identifies crosslinked peptide-RNA/DNA (oligo)nucleotides and crosslinked amino acids rapidly and sensitively. NuXL is an open-source software and has been developed using the OpenMS software framework(68). NuXL is available as standalone version as part of the upcoming OpenMS version and also available as a free node for the Proteome Discoverer (PD) software platform (for online tutorials, see https://openms.de/applications/nuxl/)(69). NuXL generates extensive reports at the CSM, peptide, and protein levels, including output for visualization of fragment annotations in the separate viewer application TOPPView for the standalone version of NuXL and in PD(70). NuXL allows the user to choose between crosslinker-specific presets (conventional UV, DEB, NM, FA, 4SU, and 6SG; Figure 2A and Supplementary Table S1) that describe the relation between fragment adducts and cross-linked precursor adducts. In the standalone version of NuXL, these presets can be extended by the user to support novel protocols. The NuXL algorithm processes XL-MS data by MS2 spectra pre-processing, theoretical spectrum generation and spectrum-matching, crosslinked amino acid localization, scoring, and subsequent protein inference (Figure 2B; see Supplementary Information for extended description). For spectrum matching and scoring, we designed subscores that consider XL-MS-specific spectrum features: precursor ions with crosslinker-specific (oligo)nucleotide mass adducts and their associated neutral losses; precursor adduct and cross-linked nucleotide-dependent MS2 fragment ions with (oligo)nucleotide mass adducts (including losses); and singly charged marker ions (i.e., unmodified or chemically crosslinked nucleobase, nucleoside, nucleotide and immonium ions; Figure 2C; Supplementary Table S3). To increase CSM identification rates, subscores are combined and rescored using the Percolator algorithm(71). Lastly, the CSM-FDR is calculated;

protein groups are inferred from all peptides. NuXL localizes crosslink sites from the pattern of unmodified and crosslinked fragment ions and from the presence of crosslinked immonium ions. Spectra with peak annotations can be visualized in TOPPView and in PD. Comprehensive tutorials are provided online for both the standalone version and the PD version of NuXL (https://openms.de/applications/nuxl/).

We evaluated NuXL performance using datasets generated for this study including *in vitro* and *in vivo* UV, DEB, NM and FA XL-MS data (Supplementary Tables S2, S4-6). Compared with our previous approach in RNPxl, the XL-MS data analysis time with NuXL is decreased by several orders of magnitude; this is achieved by implementing NuXL as a specialized search engine with full control over the internal search process, fragment-ion generation, scoring, optimized spectrum comparison, parallelization and caching of MS2 spectra generated *in silico*(2,72). Score combination of XL-specific subscores and rescoring using Percolator is fundamental to NuXL's sensitivity and improved crosslink identification rates at the CSM level for all protocols (1.2- to 8.1-fold higher identification rates). We compared NuXL to the established MSFragger custom offset workflow (XRNAX MassOffset(8) plus Percolator rescoring, mass shifts from NuXL presets) using our *E. coli* crosslink data (see below). Our findings revealed a high degree of overlap in identified crosslinks (mean 81% shared CSMs; 88% shared proteins, across all protocols). NuXL showed a higher yield in in CSMs (104% CSMs uniquely identified by NuXL vs. 19% CSMs uniquely identified by MSFragger; 74% distinct proteins unique to NuXL vs. 12% unique to MSFragger using the XRNAX MassOffset protocol)(8). When we compared NuXL with MSFragger labile search(73), NuXL achieved comparable or higher gains (28% CSMs unique to NuXL vs. 23% CSMs unique to MSFragger (labile); 36% distinct proteins unique to NuXL vs. 15% unique to MSFragger labile protocol). Percentages are relative to the total number of CSMs or proteins reported by the MSFragger protocols.

UV combined with chemical crosslinking provides spatial constraints in structural modelling of protein-RNA complexes

UV XL-MS resolved at the amino-acid level provides high-resolution spatial information on RNA interaction interfaces of proteins, which proved to be valuable to understand the three-dimensional arrangement of biomolecular complexes(2,9,11,16,66),(27). We applied UV, DEB and NM XL-MS for the analysis of different protein-RNA complexes reconstituted *in vitro*, highlighting XL MS's role in structural analysis, particularly for RNPs in which the 3D protein

structure but not the 3D RNA structure was resolved. FA XL-MS was excluded due to its lack of crosslink site information.

We analysed yeast spliceosomal protein Hsh49 in complex with Cus1 and U2 47Nt-snRNA(26). Hsh49 contains two RNA recognition motifs (RRM). Hsh49 is thus a good example of a protein with a well-studied RNA-binding domain(26,74). For all crosslinking conditions, we identified crosslink sites located in both RRM domains as well as in regions not resolved including crosslinks in Cus1 protein (Figure 3A, Supplementary Table S4).

We also applied UV, DEB and NM XL-MS to obtain spatial restraints for the support of *in silico* docking models of *S. pombe* methyltransferase Dnmt2 in complex with its cognate $tRNA^{Asp}$ (27). Atomic models of *S. pombe* $tRNA^{Asp}$ and Dnmt2 were utilized in docking calculations. Filtering 45,000 docking models by crosslink compatibility, we identified distinct crosslink sites for each method: two unique sites (all uracil) by UV crosslinking; nine sites (all purines) by DEB XL-MS, and 15 sites (to cytosine, adenine, and guanine) by NM XL-MS (localisation score ≥2, Supplementary Table S4). Both sets of crosslinks, 'UV plus chemical' and 'exclusively chemical' crosslinks, singled out two sets of decoys sharing the same three models representing a similar binding mode of $tRNA^{Asp}$ to Dnmt2, which differs from a previously published model of $tRNA^{Asp}$-Dnmt2 complex using UV crosslink information only (Figure 3B)(27). Hence, chemical crosslinking provides spatial information complementary to that from UV crosslinking, and can be used in crosslink-based modelling of RNP complexes(29).

Finally, we analysed the negative elongation factor (NELF) complex with its four subunits NELF-A/B/D/E, in complex with HIV TAR RNA(28). XL-MS with UV, DEB and NM identified 27, 50 and 38 unique crosslinks, respectively (Figure 3C, Supplementary Table S4). We identified numerous crosslink sites in the tentacle regions of NELF-A/E that so far have not been visualized. Fluorescence anisotropy experiments with mutants of the NELF complex showed significantly reduced TAR RNA binding upon deleting intrinsically disordered regions of either NELF-A or NELF-E, and an even greater loss in binding affinity when both regions were deleted (Figure 3D). Overall, our data demonstrate that UV and chemical crosslinking cover different sites within RNPs reconstituted *in vitro*, and hence are valuable for mapping extended RNA-binding interfaces in proteins and for providing suitable restraints for in silico docking of RNA to proteins.

NuXL facilitates identification of UV and chemical *crosslinking sites* in *E. coli*

To assess chemical XL-MS and NuXL *in vivo*, we applied our method to *E. coli* cells. After crosslinking with UV, DEB, NM, or FA, cells were lysed, enriched for crosslinked peptide-(oligo)nucleotides, and analysed with LC-MS. NuXL identified 4,924 unique crosslinked peptides across 1,502 proteins, with varying numbers of peptides and proteins identified by each XL-MS method (1,007, 2,715, 2,805 and 670 crosslinked peptides, and 393, 952, 1,186 and 302 crosslinked proteins, were identified by UV, DEB, NM and FA XL-MS, respectively) (Figure 4A, Supplementary Table S5). 162 proteins were found crosslinked under all four crosslinking conditions. UV crosslinked mostly pyrimidines (51% cytosine/uracil, 39% uracil, 2% cytosine) and the chemical crosslinkers most frequently gave purine crosslinks (63%, 70%, 51% guanine, and 32%, 26%, 38% adenine for DEB, NM and FA, respectively; Figure 4B). The most frequently identified crosslinked amino acids in unique crosslinks in UV XL-MS *E. coli* data were Lys (41%), Tyr (14%) and Cys (7%). For DEB XL-MS, we identified Met (42%), His (14%) and Lys (11%); for NM XL-MS, Met (64%), His (9%) and Cys (5%) (Figure 4B).

We mapped *in vivo* UV, DEB and NM crosslink sites and FA crosslinked peptides of ribosomal proteins onto the 70S *E. coli* ribosome structure (Figure 4C)(75). For this, we filtered crosslink sites for localisation score ≥1, as judging from manual spectrum validation, this threshold secures reliable amino acid localisation. We also calculated distances between Cα atoms of crosslinked amino acids and the closest atom of the closest nucleotide for UV, DEB and NM. The median distances found for UV, DEB, and NM were 6.8, 8.9, and 8 Å, respectively, which agrees well with calculated spacer lengths and consideration of structural flexibility (we accept maximum distances of 10, 15 and 15 Å for UV, DEB and NM, respectively) and underscores the specificity of chemical crosslinking for *in vivo* crosslink analysis (Figure 4C).

We also located the *in vivo* XL-MS sites of the polynucleotide phosphorylase (PNPase) on available 3D structures of the protein trimer in complex with six copies of HfQ and RNA (Figure 4D)(76,77). PNPase UV-crosslinking sites are positioned in the cavity of the PNPase tetramer core, and their locations align with the amino acids responsible for the RNA degradation activity of the PNPase(77). Chemical crosslinking sites are located mainly in, or very close to the S1/KH domain of the PNPase that binds to (small) RNAs and guides them towards degradation. PNPase forms the *E. coli* degradasome with proteins RNaseE, RhlB, and Enolase and we also identified crosslinking sites in these proteins (Supplementary Table S5,

Supplementary Data S1)(78). Enolase also plays an essential part in the glycolysis pathway; the crosslink sites are scattered across the enzyme, with no visible tendency to cluster, for example, at positively charged surface patches. Several crosslinks are located in the cleft between two enolase monomers in close vicinity to the active site of the enzyme (Figure 4E).

Properties of crosslinked *E. coli* proteins

Because of the large number of detected crosslinking sites, we aimed to assess whether crosslinking is associated with molecular features of the crosslinked proteins. For this, we integrated secondary structure, residue properties and 3D structure annotations. We also investigated the abundance of proteins detected across XL-MS methods, revealing that crosslinked proteins are higher abundant proteins in our *E. coli* samples as determined by label-free MS (Supplementary Figure S2, Supplementary Table S6). We statistically tested the enrichment of amino acid residue features of crosslinked amino acids and peptides against an expected background distribution of *E. coli* peptides, considering peptide length and tryptic missed cleavages and protein abundance (see Methods for details). The DESCRIBEPROT database protein annotations were transferred for all identified crosslinked peptides (for FA XL-MS) and sites (for UV, DEB and NM XL-MS) at the residue level. We found that secondary structure elements were only mildly associated with crosslinking events (Figure 5A, Supplementary Table S7). The analysis of residue properties revealed the under-representation of putative disordered residues at the crosslink sites, but an overall enrichment of disordered RNA and DNA binders on the crosslinked peptides (Figure 5A). The strongest enrichment was observed for specific amino acids at crosslinked sites (Supplementary Figure S3), confirming the amino acid preferences displayed in Figure 1B and Figure 4B.

We analysed crosslinked protein sets under different MS-XL conditions for gene ontology (GO) term enrichment using STRING database annotations (Figure 5B, Supplementary Table S8). The most enriched GO terms are related to the cytosol, nucleic acid binding, and translation and those for biosynthetic processes. Regarding the latter, we found almost all proteins involved in glycolysis and the tricarboxylic acid cycle (TCA) and their crosslinked amino acids linked to mono- and dinucleotides. Other studies of the RNA interactome in *E. coli* and eukaryotes also identified metabolic enzymes crosslinked to RNA, but barely with their crosslinked amino acid (Supplementary Figure S4A)(6,79),(80,81). In glycolysis, 27% of enzymes were crosslinked across all XL-MS conditions, and 80% under at least one (Supplementary Figure S5A). Over half

of these enzymes were linked to dinucleotides or cytosine/uracil, indicating RNA-binding properties. Similar trends were observed in the TCA pathway (Supplementary Figure S5B). Our analysis also revealed an overrepresentation of coding and non-coding RNA-related and metabolic terms and domains, including RNA-binding domains, as well as S1 and OB fold domains, and cofactor-binding domains such as the AAA domain (Figure 5B).

To globally determine if (RNA) crosslinking in cofactor-binding metabolic proteins occurs at consistent or varying sites, we mapped UV, DEB, and NM crosslink sites onto the sequences and structures of proteins (Supplementary Data S1, Supplementary Figure S6). We assessed co-localization of crosslinked residues and functionally important amino acids from DESCRIBEPROT and InterPro databases. We compared the 3D distances between C_α of the crosslinked amino acids and the C_α of residues with annotated features against a background distribution (see Methods for details) and found that RNA-binding residues, disordered regions, conserved residues and molecular recognition features (i.e., protein interactions) were enriched within a 4 Å radius of crosslink sites (Supplementary Figure S6A-G, Supplementary Table S7, S9). Crosslinked sites were significantly closer to RNA-binding regions and further away from DNA-binding areas. Additionally, crosslink sites were enriched adjacent to Coenzyme A and FAD/NAD/NADP. Crosslink sites across all methods were close to ATP-binding residues (Supplementary Figure S6H-L).

To further provide structural insights, we show protein structures with their bound mononucleotides/co-enzyme A in Figure 5C&D. For inosine 5'-monophosphate dehydrogenase (IMDH; AF-P0ADG7-F1), an enzyme involved in guanosine-nucleotide metabolism, amino acids crosslinked chemically to adenine and guanine are in close proximity to the Bateman domain of the predicted 3D structure (Figure 5C). The Bateman domain is known to be involved in RNA and ssDNA binding(82,83). In addition, several crosslinks close to conserved residues in the active site, which binds IMP, were identified. The crosslink data support the suggestion of regulatory roles for RNA associated with ATP and GTP binding in the Bateman domain(84). We further mapped crosslinks to the structure of succinyl-CoA synthetase subunits alpha and beta (Figure 5D). In SUCD, crosslinks were identified close to bound coenzyme A, including a dinucleotide crosslink of unambiguous RNA origin. In SUCC,

crosslinks are rather scattered across the whole structure, without any apparent tendency to cluster at the ADP-binding site.

Chemical *XL*-MS identification of protein-DNA crosslinks – DNA-specific NuXL settings

We adapted our chemical NuXL presets for the analysis of protein-DNA crosslink sites (Figure 6, Supplementary Table S10) by comparing chemical and UV XL-MS with dsDNA *in vitro* using human nucleosomes. Sample-processing, including crosslink enrichment, was performed as described above for the analysis of *E. coli* ribosomes. In addition to peptides crosslinked to the nucleobase moiety of DNA, we observed precursor-ion mass adducts of 178.00311 Da ($C_5H_7O_5P$) or 196.01368 Da ($C_5H_9O_6P$), matching deoxyribose-phosphate adducts with and without loss of H_2O, respectively. These are derived from abasic sites in the DNA generated by the crosslinker reaction and lead to generation of a reactive aldehyde that crosslinks to the proteins (Supplement figure S7A-C)(85,86). We did not observe a similar species in protein-RNA crosslinking.

UV, DEB, NM and FA XL-MS sites were mapped to structures of the human nucleosome core particle (Figure 6A). For UV, DEB and NM, crosslinked amino acids were found to cluster in close proximity to the nucleosomal DNA. We measured distances of Cα atoms of crosslinked amino acid residues that were resolved in the structures to the closest atom of the closest nucleotide. The median of amino-acid-dNTP distances was 9.4, 9.9 and 9.8 Å for UV, DEB and NM, respectively, which agrees well with UV crosslinking and calculated spacer lengths.

UV XL-MS mostly identified unique crosslinks to deoxyribose (73%), thymine (15%) and cytosine (9%) with cysteine as the most crosslinked amino acid (31%; Figure 6B; Supplementary Table S10). DEB XL-MS identifies mostly crosslinks to deoxyribose (84% of unique crosslinks), followed by guanine (10%) involving mainly Cys (39%; Figure 6B, Supplementary Table S6). For NM XL-MS, we identified mostly deoxyribose crosslinks (92%). 46% of the unique crosslink sites involved Cys (Supplementary Table S10, Figure 6B). As in RNA-protein XL-MS experiments, in DNA-protein FA XL-MS the crosslinked amino acid could not be identified in most cases. FA XL-MS identified deoxyribose (38% of crosslinks), and purine crosslinks (adenine, 40%; guanine, 12%; Figure 6B, Supplementary Table S10). We observed that the yield of chemical crosslinking exceeded that of UV crosslinking; particularly NM showed a high yield of identified crosslink sites.

Crosslinking of DNA-binding proteins in *E. coli*

We searched our *in vivo* crosslink *E. coli* dataset with NuXL and DNA presets (Supplementary Table S11). We identified the cold shock domain containing Csp protein family members crosslinked to RNA and DNA according to chaperone properties in binding single-stranded RNA and DNA(87),(88). The crosslinking sites to RNA and DNA are at the same amino acids in these proteins. A similar pattern is present in the ssDNA binding protein SSB, in which the same residues crosslink to DNA and RNA. We also identified RNA and DNA crosslinks in the nucleoid double-stranded (ds)DNA-binding proteins H-NS, DBHA/DBHB, FIS, and IHF (Figure 7A). We mapped IHFA and IHFB crosslinks to the EM 3D structure of the Cas1-Cas2-IHF-DNA holo-complex (Figure7A). The UV and chemical RNA crosslinking sites are located in close proximity to the dsDNA in the structure. Potential RNA-binding was previously suggested in acid-resistance studies(89). All other nucleoid proteins show RNA and DNA crosslinking at the same positions in the 3D structures, which indicates RNA-binding properties of nucleoid proteins in general (Figure 7A).

UV XL-MS generally unambiguously distinguishes between RNA and DNA in crosslinks, because thymidine and deoxyribose crosslinks generate DNA-specific mass adducts. In contrast, many chemical crosslinks, involving purines, are ambiguous as deoxyguanosine and adenosine are isobaric (Supplementary Table S12). The detection of adducts of two or more nucleotides harbouring cytidine, uridine or thymidine reduces such ambiguity. Also, deoxyribose crosslinks originated unambiguously from DNA (see above). However, their presence, in particular in a cellular context, must be still considered with caution. It is not entirely clear whether crosslinking of these reactive deoxyribose derivatives might be (entirely) reversible and could lead to crosslinking to proteins not necessarily involved in direct DNA-binding. In protein TOP1, which releases negative supercoiling in DNA, we identified unambiguous UV crosslinks to deoxydinucleotides, within a peptide encompassing amino acid positions 774-788 (Figure 7B)(90). The peptide contains Phe780 and Phe786, which are very close to the ssDNA in the 3D structure. Ambiguous crosslinks derived from chemical crosslinking (Met376, Met438, Met522 and Met669), were all located near conserved residues involved in DNA binding(90). Deoxyribose crosslinks were found to Cys401 and Cys763 in the DNA- and ssDNA-binding region, respectively.

Overall, we identified 59 crosslinked DNA-binding transcriptional regulators (according to the Regulon database (91)) in *E. coli*, exclusively after chemical crosslinking. Crosslinking sites were found within or close to described DNA-binding domains of these regulators, such as the helix-turn-helix (H-T-H) motif (FIS, ARGP, NANR, SLYA, YBAQ, ROB, ASNC, LRP, DEOR, LLDR, CYTR, OXYR, MPRA, ZNTR, YJDC, EVGA, NAGC, UHPA, GATR, HYFR, BLUR, YCJW, MQSA), OmpR/PhoB-type domain (ARCA, OMPR, CPXR, PHOB, PHOP, BASR, YHAJ, BAER), or the LytTr DNA-binding domain (BTSR). In approximately half of the transcription regulators, the crosslinking site is not located in the H-T-H motif (ULAR, GLPR, EXUR, YIHL, PURR, RBSR, TDCA, YEIE, ENVR, UIDR, ICLR, YAFC, YIEP, NARP, CSQR, DCTR, YIAJ, DSDC, YGBI, COMR, YCIT, LSRR, KDGR) but in N- or C-terminal regions of the protein. These potential functional domains have not yet been defined, however, our data suggests a nucleic acid-binding role of these sequences. Recently, the RNA-binding properties of H-T-H transcription regulators have been confirmed in *S. aureus*(92). Owing to the ambiguity in DNA/RNA chemical crosslinking, we currently cannot finally conclude, whether the crosslinks in the H-T-H domain or the other, as yet undefined, domains originated from RNA or DNA.

DISCUSSION

In a protein-centric analysis, we regard the detection/identification of crosslinked amino acids by XL-MS as the most direct approach to localizing the interactions of proteins with RNA and/or DNA. Several excellent proteome-wide crosslinking studies applied UV XL-MS and obtained remarkable inventories of proteins bound to (m)RNAs(5-7,13,14).

Compared to previous work, our study makes three major contributions: (1) establishing chemical XL-MS as a method to increase the number of cross-links, (2) introduce a dedicated search engine for crosslink identification and localization, and (3) reporting the to date largest data set of crosslinking sites for *E. coli* proteins.

UV crosslinking is the most frequently applied method in studying RNPs by MS, while chemical XL-MS is favored for application to DNPs(18,93). UV XL-MS requires, for example, 6-thio-G or purine derivatives incorporated into cellular RNAs in order to capture RNP interaction with bases other than uracil(3,94). We show that the reaction with alkylating chemicals offers an alternative, as guanine and adenosine are linked to proteins. We anticipate a broad application

of chemical XL-MS in global interaction mapping and also in modelling of DNA/RNA to its binding proteins in RNPs.

The first search engine for nucleotide-protein crosslinks, RNPxl, was restricted to UV XL-MS and required additional manual evaluation/annotation of XL-MS data(2,72). Such shortcomings have been overcome in our tool NuXL. While general-purpose peptide search engines such as MS Fragger(95) have also been applied in several UV-XL MS studies, NuXL has been developed specifically for protein-nucleotide crosslink assignment.

NuXL also allows the incorporation of other protein-nucleotide crosslinkers, for example, SDA, whose NHS ester group reacts with lysine residues in RNPs and upon photoactivation forms a carbene nucleotide in close vicinity to the derivatized lysine residue (Supplementary Figure S8A-C for reaction mechanism, adduct sum formulae, and MS2).

We compared our results with results from three other UV XL-MS studies (TRAPP(6), OOPS(7) and RIC(96)) of protein-RNA interactions in *E. coli* (Supplementary Figure S4). Our results overlap substantially with these studies (Supplementary Figure S4B), but our dataset additionally reveals the sites of crosslinking for all identified proteins. In addition, we detected 548 crosslinked proteins and their crosslinking sites that were not identified in previous studies. 74% of these proteins were detected under only one of the four crosslinking conditions (12%, 20%, 39% and 3% for UV, DEB, NM and FA XL-MS, respectively; Supplementary Figure S4C), emphasizing the advantage of the combination of UV and chemical MS-XL. For example, crosslinking sites were identified in 9 of 10 hitherto poorly characterized RBPs described in the RIC(96) study and in 12 of 17 putative RBPs investigated in a recent OOPS study(97); in nine of these exclusively upon chemical crosslinking.

Unlike in the above-mentioned studies, we did not apply a dedicated purification of e.g. RNA after crosslinking. Instead, we chose a comprehensive approach aimed to enrich any (deoxy)ribonucleotide crosslink species to show the feasibility of chemical crosslinking and NuXL in RBP and DBP analysis. Based on this, particularly chemical crosslinks show ambiguities, i.e. crosslinks that could be derived from RNA, DNA, mononucleotides or enzymatic co-factors. Nonetheless, our data largely overlaps with RNA-crosslinked proteins in other *E. coli* studies and we show that crosslinks of RNA origin coincide with nucleotide/co-factor-binding sites. Our workflow is advantageous, for comparing crosslink sites originating from RNA or DNA with assigned nucleotide/co-factor-binding sites in a single experiment. The identification of

nearly all proteins of glycolysis and the TCA cycle and their crosslinking sites (with the key enzymes crosslinked by UV, DEB, NM, and FA) strengthen the view that metabolic enzymes are RNA-binding proteins, as previously reported in pro- and eukaryotes(80,81).

FIGURES

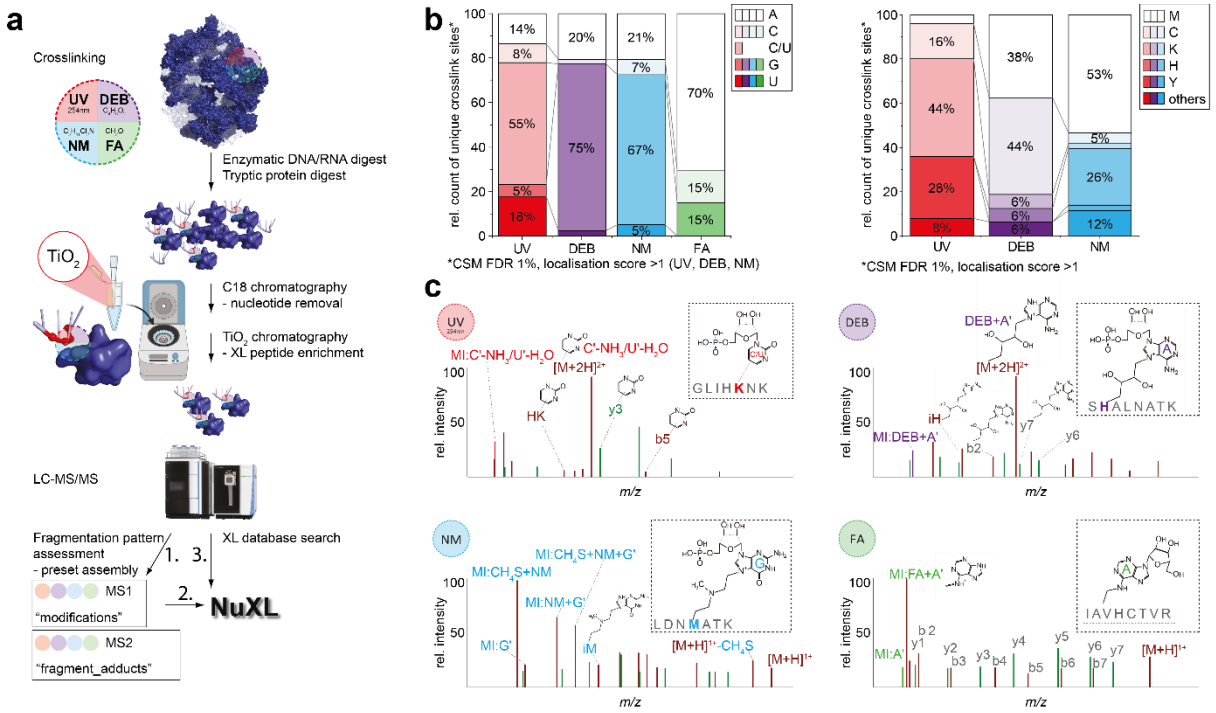

Figure 1. UV and chemical XL-MS workflow for establishing MS. **(a)** XL-MS processing and analysis workflow. Purified *E. coli* ribosomes were crosslinked with UV, DEB, NM, or FA. DNA, RNA and proteins were digested enzymatically and free nucleotides were removed by C18 chromatography. Crosslinked peptides were enriched by TiO_2 chromatography and analysed by LC-MS/MS. Spectra were evaluated manually and crosslinker-specific adducts were assembled into presets. MS data was then analyzed using NuXL. **(b)**, left panel, stacked bar plot (normalised to 100%) of relative counts of observed crosslinked nucleobases in unique crosslinked peptide-(oligo)nucleotide pairs resulting from UV, DEB, NM and FA XL-MS of *E. coli* ribosomes. CSMs were filtered for 1% FDR and for UV, DEB and NM a minimum localisation score ≥1. Ambiguous, isobaric C–NH_3/U–H_2O crosslinks were counted as C/U. Labels are displayed only for values of 5% and above. Right panel, 100% stacked bar plot of relative counts of observed crosslinked amino acids in unique crosslinked peptide-(oligo)nucleotide pairs resulting from UV, DEB, NM and FA XL-MS of *E. coli* ribosomes. CSMs were filtered for 1% FDR and a localisation score ≥1. Labels are displayed only for values of 5% and above. **(c)**, Schematic crosslink MS2 spectra for UV-, DEB-, NM- and FA-crosslinked peptide–RNA(oligo)nucleotides. Peptide fragment b- and y-type ions are shown in red and green, respectively. UV crosslink MS2 spectrum of GLIHKNK-CMP-NH_3/UMP-H_2O. Marker ion (C'–NH_3 or U'–H_2O), fragment ions and internal fragment (HK, His-Lys) modified with C'–NH_3 or U'–H_2O indicate a crosslink located at Lys-5. DEB crosslink MS2 spectrum of peptide–

RNA(oligo)nucleotide SHALNATK-DEB-AMP. Marker ion (DEB plus adenine, DEB+A'), fragment ions b2 and y7 modified with DEB+A', and unmodified y6 indicate crosslink localisation at His-2. NM crosslink MS2 spectrum of crosslinked peptide-RNA(oligo)nucleotide LDNMATK-NM-GMP. Methionine-specific losses of CH_4S on singly charged precursor ($[M+H]^{1+}$) ion and marker ions (CH_4S plus NM plus guanine, CH_4S+NM+G'; CH_4S+NM; NM+G'; G') unambiguously identify a Met-4 crosslink to G'. FA crosslink MS2 spectrum of crosslinked peptide-RNA(oligo)nucleotide IAVHCTVR-FA-AMP-HPO_3. Marker ions (FA plus adenine, FA+A'; A') identify A' crosslink. Unmodified fragment ions (b2–b7 in red, y1–y7 in green) cover the peptide sequence.

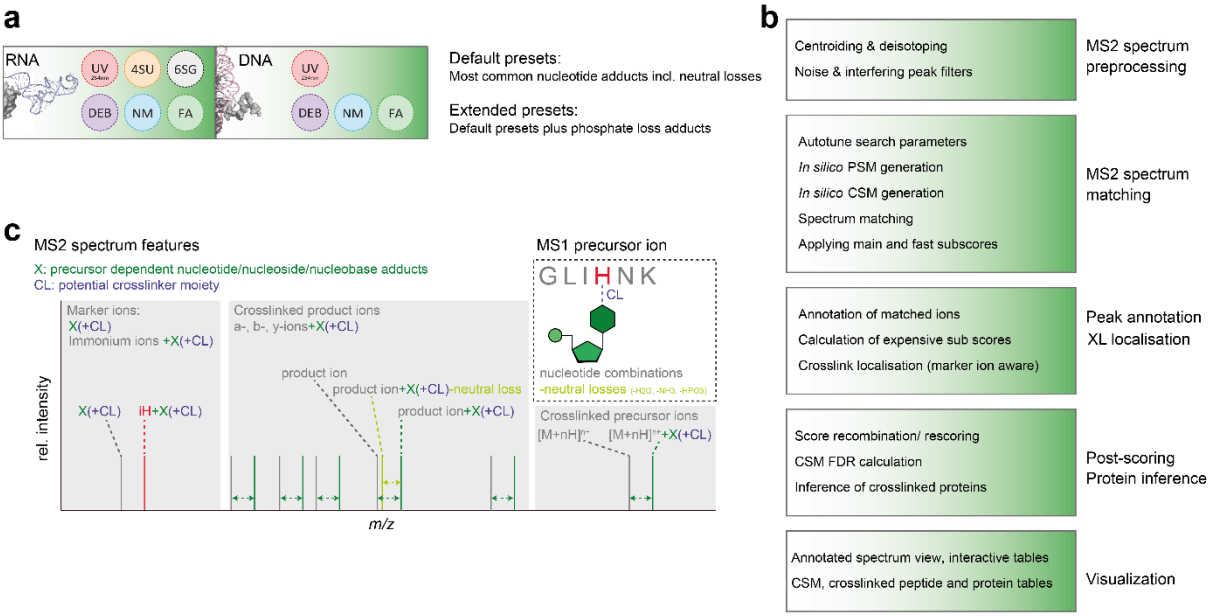

Figure 2. NuXL software tool for the analysis of XL-MS data. **(a)** Crosslinking reagents supported by NuXL database search tool. Available RNA-, DNA- and crosslinker-specific default and extended NuXL presets. **(b)** NuXL pipeline. XL-MS data-processing includes MS2 spectra pre-processing and spectrum-matching; peak annotation and localisation of crosslinked amino acid; post-scoring using the Percolator algorithm and protein inference; visualization of the annotated spectra and generation of tables of results. **(c)** MS2 spectrum features considered in (sub)score calculation. Main score and subscores used by the Percolator algorithm consider the presence and intensities of MS1 and MS2 precursor ions with crosslinker-specific (oligo)nucleotide mass adducts (X(+CL)) including various oligonucleotide combinations and related neutral losses, precursor adduct-dependent MS2 fragment ions with unmodified or chemically crosslinked (oligo)nucleotide mass adducts (a/b/y-ions+X(+CL)), and singly charged marker ions (nucleobase/nucleoside/nucleotide, X(+CL); immonium ions, iX+X(+CL)).

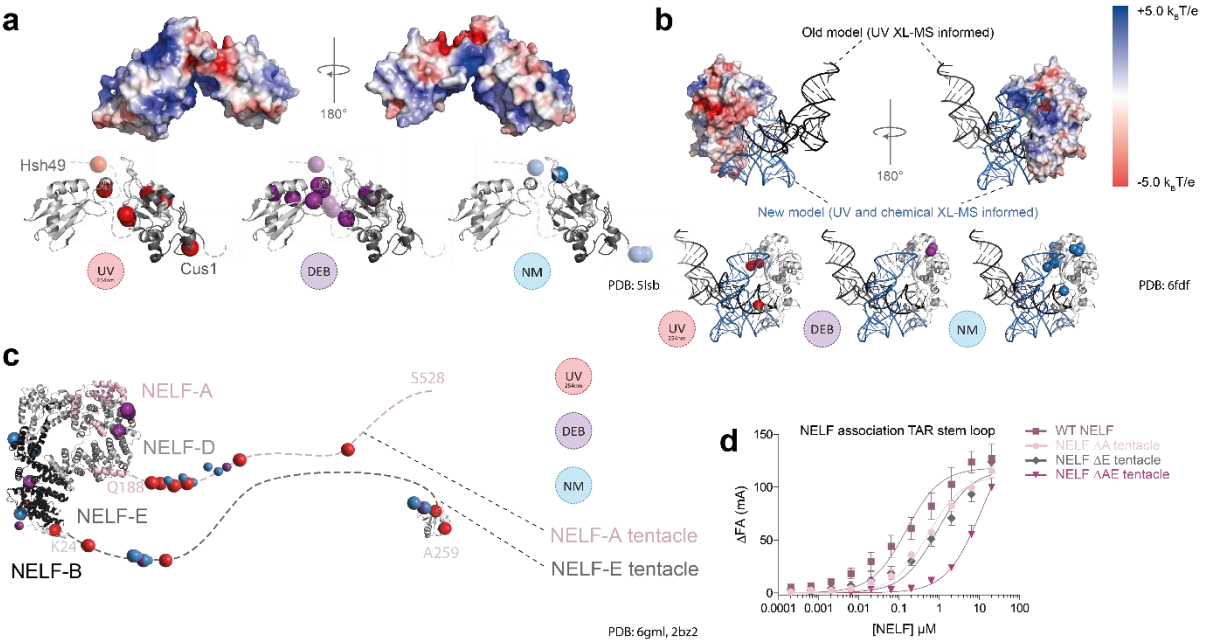

Figure 3. UV, DEB, NM XL-MS in integrated structural modelling. **(a)** 3D structure of yeast spliceosomal protein Hsh49 in complex with Cus1 (PDB: 5lsb(26)). Surface representation of electrostatic potentials depicted as a colour gradient of ±5.0 k_BT/e. Hsh49-Cus1 RNA-(oligo)nucleotide-crosslinked amino acids (Cα) highlighted as red (UV), purple (DEB) or blue (NM) spheres based on CSMs filtered for 1% FDR, 60% peptide sequence coverage of ions of ≥5% relative intensity, and localisation score ≥1. **(b)** 3D docking model of *S. pombe* methyltransferase Dnmt2 (PDB: 6fdf(27)) in complex with $tRNA^{Asp}$. Black cartoon tRNA representation: published model based on UV XL-MS data. Blue cartoon representation: New model based on UV and DEB/NM XL-MS data based on CSMs filtered for 1% FDR, 60% peptide sequence coverage of ions of ≥5% relative intensity, and localisation score ≥2. RNA-(oligo)nucleotide-crosslinked amino acids are highlighted as in **a**. **(c)** 3D structure of human negative elongation factor (NELF) complex (PDB: 6gml(98), 2bz2(99)) including RNA-(oligo)nucleotide-crosslinked amino acids highlighted and filtered as in **a**. Localisation score ≥0.5. Intrinsically disordered regions (tentacles) of NELF-A and NELF-E are indicated by grey and pink lines including relative RNA-(oligo)nucleotide-crosslinked amino acid positions. **(d)** Binding of WT NELF, NELF ΔA tentacle, NELF ΔE tentacle and ΔAE tentacle to TAR stem loop measured by fluorescence anisotropy.

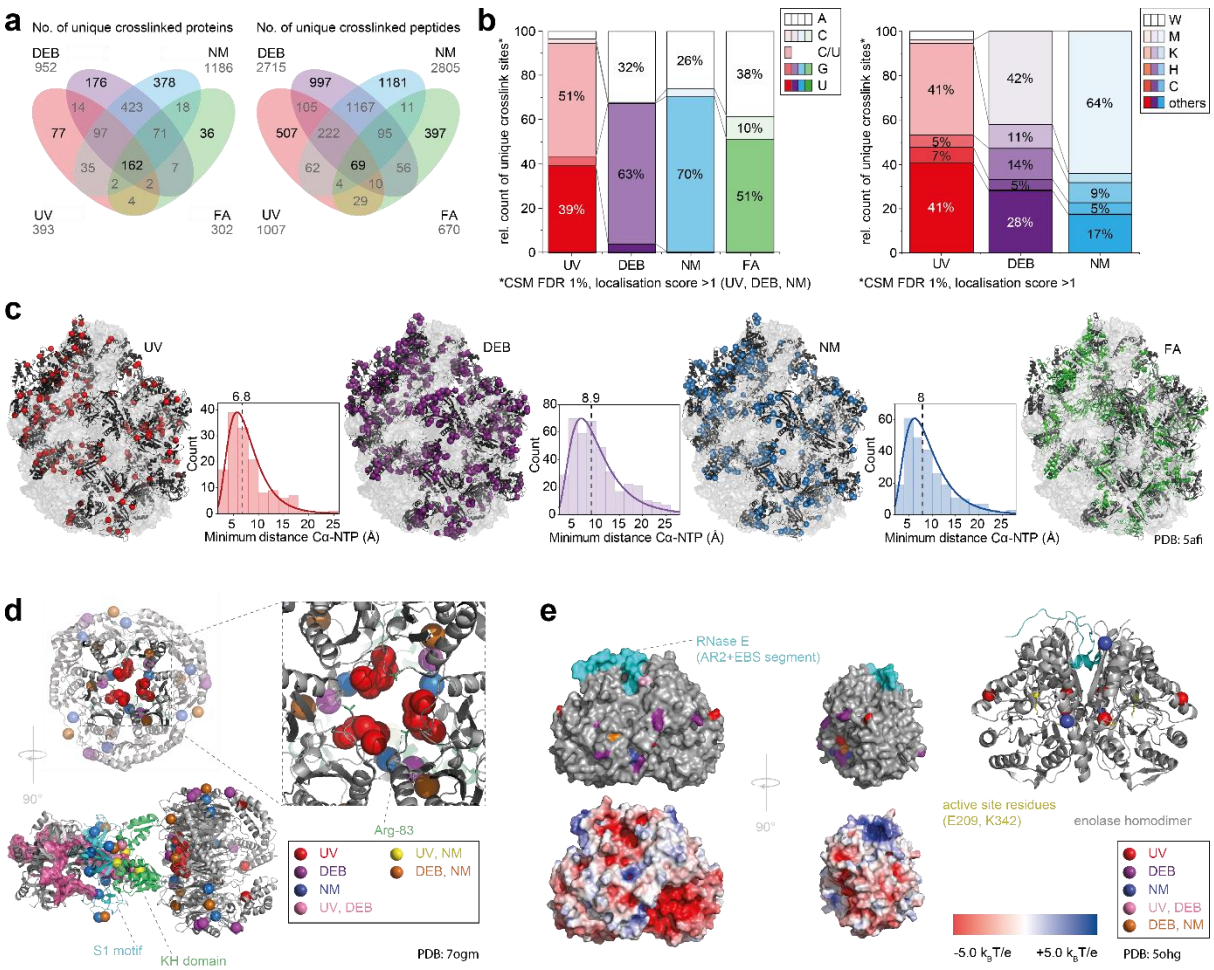

Figure 4. In vivo UV and chemical XL-MS of *Escherichia coli* cells. **(a)** Venn diagrams of unique crosslinked proteins (left) and peptides (right) for UV, DEB, NM and FA XL-MS of *E. coli* cells. Identifications are filtered for 1% FDR at CSM level. **(b)** 100% stacked bar plots (normalised to 100%) of relative counts of observed crosslinked DNA nucleotides (left panel) and amino acid residues (right panel) in unique crosslinked peptide-(oligo)nucleotide pairs resulting from UV, DEB, NM and FA XL-MS of *E. coli* cells. CSMs were filtered for 1% FDR and for UV, DEB and NM a minimum localisation score of 1. Labels are displayed only for values of 5% and above. **(c)** Cartoon representation of cryo-EM 3D structure of *E. coli* 70S ribosome-EF-TU complex (PDB: 5afi(75)) and histograms of crosslink distances. In the structures, RNA-(oligo)nucleotide-crosslinked amino acids are highlighted as spheres (Cα atoms) in red (UV), purple (DEB), or blue (NM); or as green cartoon (FA). Crosslink sites are based on CSMs filtered for 1% FDR for all conditions, and localisation score ≥1 for UV, DEB and NM. Histograms display distances (in Å) between UV (left), DEB (centre left) and NM (centre right) crosslinked amino acids (Cα) and closest atom of the closest nucleotide in the *E. coli* 70S ribosome-EF-TU complex structure.

(d) Cartoon representation of the cryo-electron microscopy structure of a trimer of *E. coli* polynucleotide phosphorylase (PNPase; grey cartoon) in complex with six copies of *E. coli* RNA-binding protein Hfq (grey cartoon) and the 3'external transcribed spacer RNA of leuZ (pink cartoon/surface; PDB: 7ogm(77)). PNPase KH domain and S1 motif are highlighted in limegreen and cyan, respectively. The Arg-83 residue involved in RNA binding is highlighted by green sticks(76). Alpha-carbon atoms (Cα) of RNA-(oligo)nucleotide crosslinked amino acids are highlighted as red (UV), purple (DEB), blue (NM), pink (UV and DEB), yellow (UV and NM), or orange (DEB and NM) spheres. Crosslinks are filtered for a localisation score ≥1. **(e)** Surface/cartoon representation and electrostatic surface of the crystal structure of enolase homodimer (grey surface/cartoon) in complex with RNaseE fragment (cyan surface/cartoon; PDB: 5ogh(100)). RNA-(oligo)nucleotide-crosslinked amino acids (Cα) are highlighted as purple (DEB), blue (NM), orange (DEB and NM) or light pink (UV, DEB and NM) surface or spheres. Mapped crosslink positions are based on CSMs filtered for 1% FDR, and localisation score ≥1.

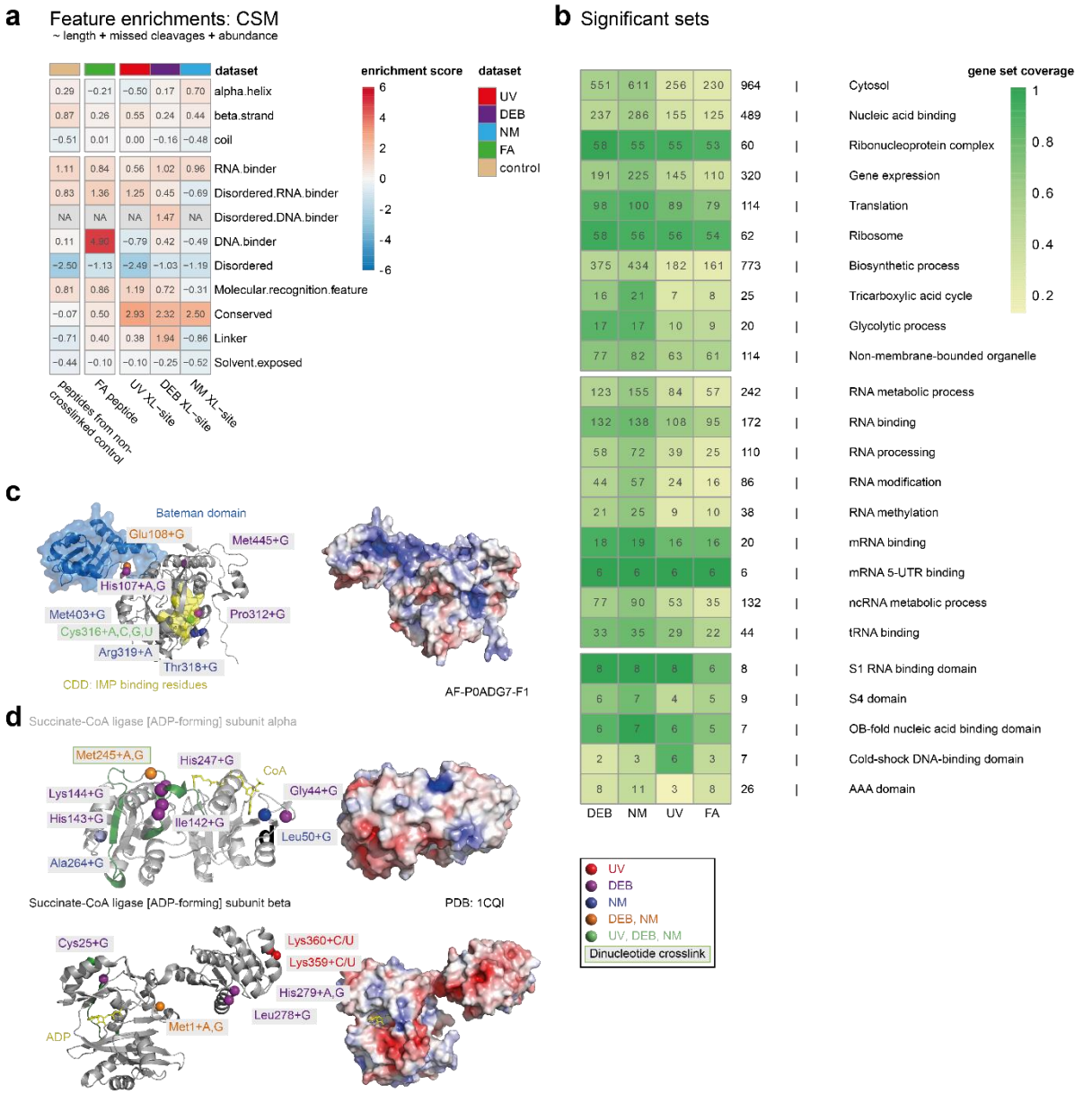

Figure 5. Crosslink-assisted functional and structural protein characterization. **(a)** A heatmap of log2FoldChanges between the experimental and an *in-silico* background distribution mean feature counts per residue. The background sampling was performed with regards to the experimentally observed peptide length and missed cleavage distribution and source protein quantity to compensate for the abundance bias. The non-crosslinked control and FA datasets were analysed on the peptide level, whilst the UV, NM and DEB enrichment was analysed on crosslinked residue level. The NA values correspond to the cases where the sampling or statistical testing could not be performed due to the feature distribution violating the method assumptions. **(b)** A gene set over-representation by the crosslinked protein sets. Reported are the terms significant in at least one crosslinking method (p.adjusted BH <= 0.01 for all sets

and 0.05 for protein domain sets). The gene set coverage is colour coded, with the number of identified proteins noted in the heatmap cells and the full gene set size next to the corresponding rows **(c)** Cartoon/surface representation and electrostatic surface of AlphaFold-predicted 3D structure of *E. coli* inosine-5'-monophosphate dehydrogenase (AF-P0ADG7-F1) with RNA-(oligo)nucleotide-crosslinked amino acids (Cα) highlighted as purple (DEB), blue (NM), orange (DEB and NM) or lime-green (UV, DEB and NM) spheres. Mapped crosslink positions are based on CSMs filtered for 1% FDR, and localisation score ≥1. The Bateman domain is highlighted as a transparent blue surface. Conserved residues (derived from conserved domain database CDD) involved in IMP binding are highlighted by a yellow transparent surface. **(d)** Cartoon representation and electrostatic surface of the crystal structure of succinate-CoA ligase subunit alpha (upper panel) with bound coenzyme A (yellow sticks) and subunit beta (lower panel) with bound ADP (yellow sticks) (PDB: 1cqi(101)). Crosslinks are mapped as in **c** with FA-crosslinked peptide displayed as green cartoon and UV-crosslinked amino acids (Cα) highlighted as red spheres. M-245 crosslink found to a dinucleotide is marked by a green-outlined box.

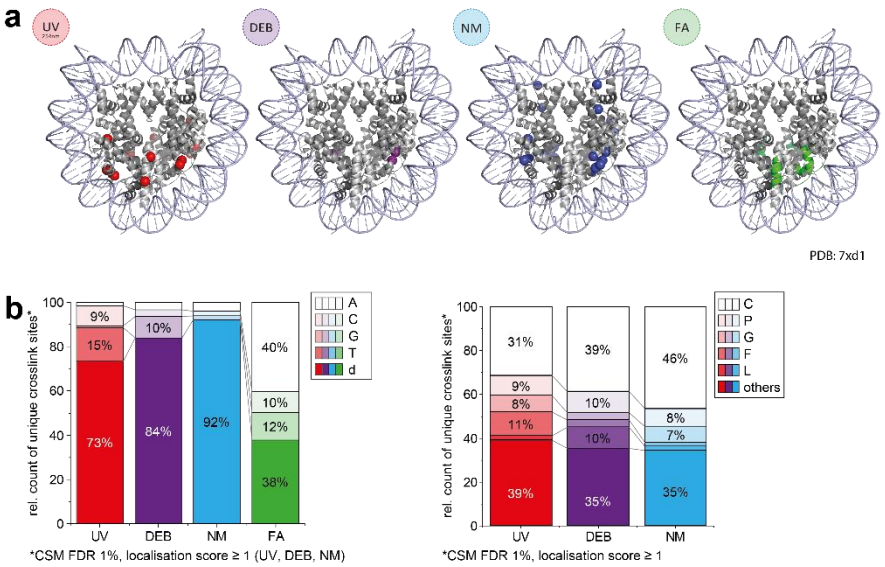

Figure 6. Protein–DNA UV and chemical XL-MS of human nucleosomes. (**a)** Human nucleosome core-particle structure (PDB: 7xd1(102)) with DNA-(oligo)nucleotide-crosslinked amino acids (Cα for UV,DEB and NM) or peptides (for FA) highlighted as red (UV), purple (DEB), blue (NM) spheres or (for FA) with a green cartoon based on CSMs filtered for 1% FDR, and localisation score ≥1 (for UV, DEB and NM). **(b)** Left panel: Stacked bar plot (normalised to 100%) of relative counts of observed crosslinked DNA nucleotides/deoxyribose (d) in unique crosslinked peptide-(oligo)nucleotide pairs resulting from UV, DEB, NM and FA XL-MS of purified human nucleosomes. CSMs were filtered for 1% FDR and for UV, DEB and NM with a localisation score ≥1. Labels are displayed only for values ≥5%. Right panel: Similar bar plot of relative counts of observed crosslinked amino acids in unique crosslinked peptide-(oligo)nucleotide pairs resulting from UV, DEB, NM and FA XL-MS of purified human nucleosomes.

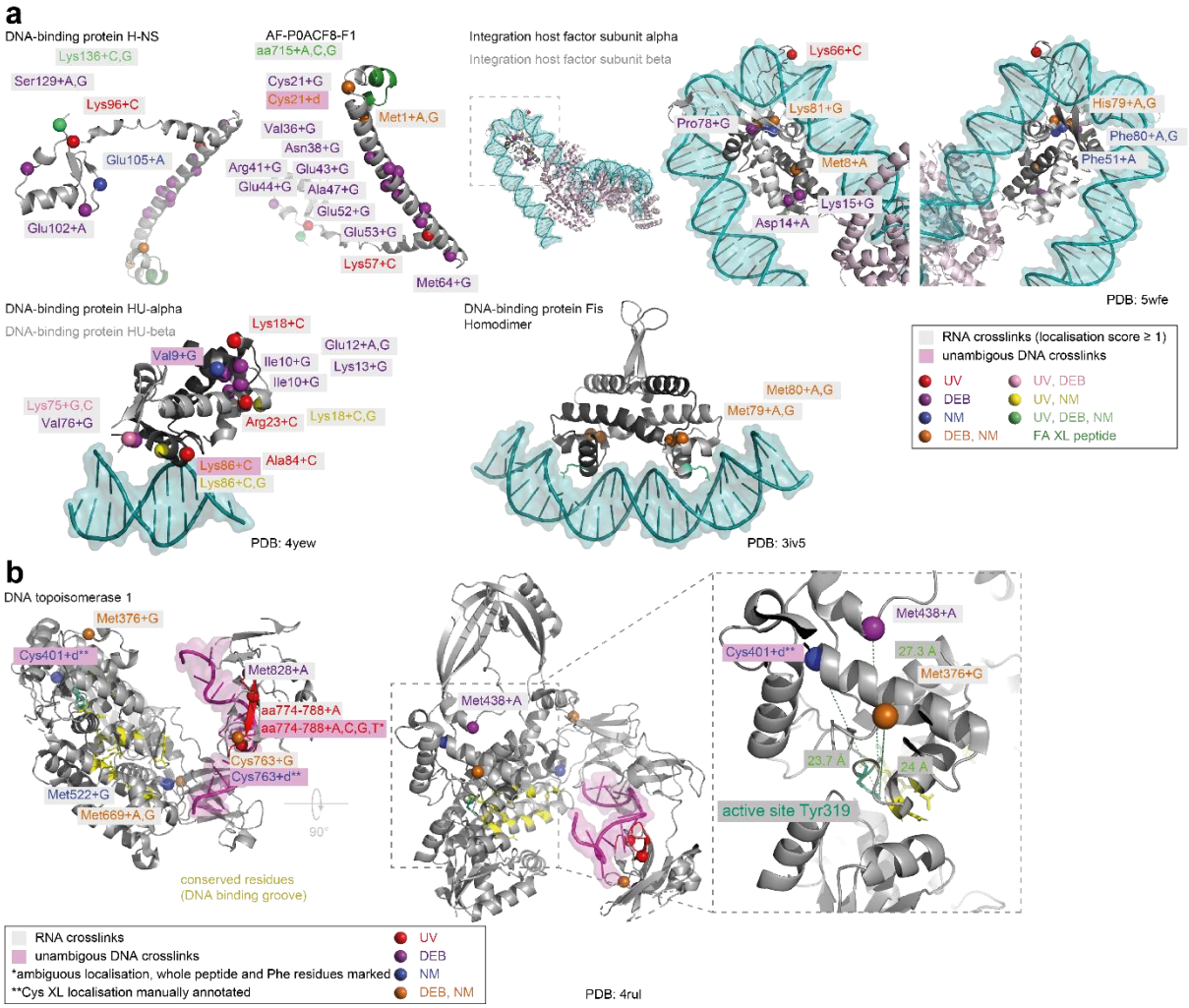

Figure 7. Examples of crosslinked DNA-binding proteins in E. coli. ***(a)*** Cartoon representations of nucleoid proteins HN-S (AF-P0ACF8-F1), integration host factor (subunit alpha and beta; *PDB: 5wfe(103)*), HU (alpha and beta; PDB: 4yew(104)), and FIS (as homodimer; PDB: 3iv5(105)) crosslinked to RNA (and DNA) in *E. coli*. DNA is displayed as dark-cyan cartoon/surface. RNA- and unambiguous DNA-(oligo)nucleotide-crosslinked amino acids (Cα) are highlighted as red (UV), purple (DEB), blue (NM), light-pink (UV and DEB), yellow (UV and NM), orange (DEB and NM) or lime-green (UV, DEB and NM) spheres. FA-crosslinked peptides are displayed as green cartoon. RNA- and unambiguous DNA-crosslink labels are coloured grey and purple, respectively. Labels indicate crosslinked amino acids and directly crosslinked nucleic acid moiety. Crosslink sites are based on CSMs filtered for 1% FDR for all conditions, and localisation score ≥ 1 for UV, DEB and NM peptide-RNA(oligonucleotide)-crosslinks. **(b)** Cartoon representation of the crystal structure of topoisomerase I (4rul(106)) with crosslinks and labels as in a. In the case of UV-crosslinked peptide spanning amino acids 774 to 788, cartoon is coloured in red and Cα of putatively crosslinked phenylalanines Phe-780 and Phe-

786 are highlighted by spheres. For Cys-401, and Cys-763, the crosslink sites were manually annotated based on the reactivity of cysteine residues. Crosslinks are not filtered based on localisation score. Conserved residues of the DNA binding groove (CDD1: cd00186) and the active site Tyr-319 are highlighted by yellow and cyan sticks, respectively. The distances between Cα of crosslinked residues Cys-401 (23.7 Å) and Met-376 (24 Å) and the hydroxyl group of the active site Tyr-319 are indicated. A, adenine; G, guanine; C, cytosine; T, thyrosine; d, deoxyribose-phosphate.

DATA AVAILABILITY

The datasets (.raw, .mzML, .idxML, and .txt files) generated and analysed for this study are deposited in the MassIVE repository and submitted to ProteomeXchange Consortium and will be available using the following identifiers MSV000094879 (MassIVE), PXD052614 (ProteomeXchange) upon peer-reviewed publication. Dnmt2-tRNA^{Asp} models will be available via figshare.

SUPPLEMENTARY DATA

Supplementary Data are available at BioRxiv online.

ACKNOWLEDGEMENTS

We thank Monika Raabe, Ralf Pflanz, Tamara Dehne and Sabine König for excellent technical assistance and support in LC-MS/MS analyses. We are grateful for Fanni Bazsó's excellent contribution in FA XL-MS *E. coli* experiments and corresponding method development.

FUNDING

H.U. and R.F. were supported by grants of the Deutsche Forschungsgemeinschaft (H.U.: SFB860, FOR1193, SFB1565, R.F.: SFB860 (Z01), and Germany's Excellence Strategy (EXC 2067/1-390729940)). T.S. and O.K. acknowledge funding by Federal Ministry of Education and Research

in the frame of de.NBI/ELIXIR-DE (W-de.NBI-022). T.S. is supported by the Ministry of Science, Research and Arts Baden-Württemberg. J.L. received financial support from the Max Planck Society and an ERC Starting Grant (ERC-StG 945528 IMAP). A.S. is part of the MSCA-ITN-2020 PROTrEIN project, which received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No: 956148. S.M.V is a Freeman Hrabowski Scholar of the Howard Hughes Medical Institute. Research in the Vos lab is funded by the Smith Family Foundation, NIH DP2GM146254, and Alex's Lemonade Foundation Crazy 8 Initiative. Funding for open access charge: Open Access Publication Fund of the Max Planck Society.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

1. Urdaneta, E.C. and Beckmann, B.M. (2020) Fast and unbiased purification of RNA-protein complexes after UV cross-linking. *Methods*, **178**, 72-82.
2. Kramer, K., Sachsenberg, T., Beckmann, B.M., Qamar, S., Boon, K.L., Hentze, M.W., Kohlbacher, O. and Urlaub, H. (2014) Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins. *Nat Methods*, **11**, 1064-1070.
3. Bae, J.W., Kim, S., Kim, V.N. and Kim, J.S. (2021) Photoactivatable ribonucleosides mark base-specific RNA-binding sites. *Nat Commun*, **12**, 6026.
4. Bae, J.W., Kwon, S.C., Na, Y., Kim, V.N. and Kim, J.S. (2020) Chemical RNA digestion enables robust RNA-binding site mapping at single amino acid resolution. *Nat Struct Mol Biol*, **27**, 678-682.
5. Panhale, A., Richter, F.M., Ramirez, F., Shvedunova, M., Manke, T., Mittler, G. and Akhtar, A. (2019) CAPRI enables comparison of evolutionarily conserved RNA interacting regions. *Nat Commun*, **10**, 2682.
6. Shchepachev, V., Bresson, S., Spanos, C., Petfalski, E., Fischer, L., Rappsilber, J. and Tollervey, D. (2019) Defining the RNA interactome by total RNA-associated protein purification. *Mol Syst Biol*, **15**, e8689.
7. Queiroz, R.M.L., Smith, T., Villanueva, E., Marti-Solano, M., Monti, M., Pizzinga, M., Mirea, D.M., Ramakrishna, M., Harvey, R.F., Dezi, V. *et al.* (2019) Comprehensive identification of RNA-protein interactions in any organism using orthogonal organic phase separation (OOPS). *Nat Biotechnol*, **37**, 169-178.
8. Trendel, J., Schwarzl, T., Horos, R., Prakash, A., Bateman, A., Hentze, M.W. and Krijgsveld, J. (2019) The Human RNA-Binding Proteome and Its Dynamics during Translational Arrest. *Cell*, **176**, 391-403 e319.
9. Cretu, C., Schmitzova, J., Ponce-Salvatierra, A., Dybkov, O., De Laurentiis, E.I., Sharma, K., Will, C.L., Urlaub, H., Luhrmann, R. and Pena, V. (2016) Molecular Architecture of SF3b and Structural Consequences of Its Cancer-Related Mutations. *Mol Cell*, **64**, 307-319.

10. Stützer, A., Welp, L.M., Raabe, M., Sachsenberg, T., Kappert, C., Wulf, A., Lau, A.M., David, S.S., Chernev, A., Kramer, K. *et al.* (2020) Analysis of protein-DNA interactions in chromatin by UV induced cross-linking and mass spectrometry. *Nat Commun*, **11**, 5250.
11. Keil, P., Wulf, A., Kachariya, N., Reuscher, S., Hühn, K., Silbern, I., Altmüller, J., Keller, M., Stehle, R., Zarnack, K. *et al.* (2023) Npl3 functions in mRNP assembly by recruitment of mRNP components to the transcription site and their transfer onto the mRNA. *Nucleic Acids Res*, **51**, 831-851.
12. Jia, J., Hilal, T., Bohnsack, K.E., Chernev, A., Tsao, N., Bethmann, J., Arumugam, A., Parmely, L., Holton, N., Loll, B. *et al.* (2023) Extended DNA threading through a dual-engine motor module of the activating signal co-integrator 1 complex. *Nature Communications*, **14**, 1886.
13. Castello, A., Fischer, B., Frese, C.K., Horos, R., Alleaume, A.M., Foehr, S., Curk, T., Krijgsveld, J. and Hentze, M.W. (2016) Comprehensive Identification of RNA-Binding Domains in Human Cells. *Mol Cell*, **63**, 696-710.
14. Mullari, M., Lyon, D., Jensen, L.J. and Nielsen, M.L. (2017) Specifying RNA-Binding Regions in Proteins by Peptide Cross-Linking and Affinity Purification. *J Proteome Res*, **16**, 2762-2772.
15. Dorn, G., Leitner, A., Boudet, J., Campagne, S., von Schroetter, C., Moursy, A., Aebersold, R. and Allain, F.H. (2017) Structural modeling of protein-RNA complexes using crosslinking of segmentally isotope-labeled RNA and MS/MS. *Nat Methods*, **14**, 487-490.
16. Stutzer, A., Welp, L.M., Raabe, M., Sachsenberg, T., Kappert, C., Wulf, A., Lau, A.M., David, S.S., Chernev, A., Kramer, K. *et al.* (2020) Analysis of protein-DNA interactions in chromatin by UV induced cross-linking and mass spectrometry. *Nat Commun*, **11**, 5250.
17. Reim, A., Ackermann, R., Font-Mateu, J., Kammel, R., Beato, M., Nolte, S., Mann, M., Russmann, C. and Wierer, M. (2020) Atomic-resolution mapping of transcription factor-DNA interactions by femtosecond laser crosslinking and mass spectrometry. *Nat Commun*, **11**, 3019.
18. Tretyakova, N.Y., Groehler, A.t. and Ji, S. (2015) DNA-Protein Cross-Links: Formation, Structural Identities, and Biological Outcomes. *Acc Chem Res*, **48**, 1631-1644.
19. Michaelson-Richie, E.D., Ming, X., Codreanu, S.G., Loeber, R.L., Liebler, D.C., Campbell, C. and Tretyakova, N.Y. (2011) Mechlorethamine-induced DNA-protein cross-linking in human fibrosarcoma (HT1080) cells. *J Proteome Res*, **10**, 2785-2796.
20. Michaelson-Richie, E.D., Loeber, R.L., Codreanu, S.G., Ming, X., Liebler, D.C., Campbell, C. and Tretyakova, N.Y. (2010) DNA-protein cross-linking by 1,2,3,4-diepoxybutane. *J Proteome Res*, **9**, 4356-4367.
21. Loeber, R.L., Michaelson-Richie, E.D., Codreanu, S.G., Liebler, D.C., Campbell, C.R. and Tretyakova, N.Y. (2009) Proteomic analysis of DNA-protein cross-linking by antitumor nitrogen mustards. *Chem Res Toxicol*, **22**, 1151-1162.
22. Groehler, A.t., Degner, A. and Tretyakova, N.Y. (2017) Mass Spectrometry-Based Tools to Characterize DNA-Protein Cross-Linking by Bis-Electrophiles. *Basic Clin Pharmacol Toxicol*, **121 Suppl 3**, 63-77.
23. Gherezghiher, T.B., Ming, X., Villalta, P.W., Campbell, C. and Tretyakova, N.Y. (2013) 1,2,3,4-Diepoxybutane-induced DNA-protein cross-linking in human fibrosarcoma (HT1080) cells. *J Proteome Res*, **12**, 2151-2164.
24. Urlaub, H., Thiede, B., Müller, E.C., Brimacombe, R. and Wittmann-Liebold, B. (1997) Identification and sequence analysis of contact sites between ribosomal proteins and rRNA in Escherichia coli 30 S subunits by a new approach using matrix-assisted laser desorption/ionization-mass spectrometry combined with N-terminal microsequencing. *J Biol Chem*, **272**, 14547-14555.
25. Scalabrin, M., Dixit, S.M., Makshood, M.M., Krzemien, C.E. and Fabris, D. (2018) Bifunctional cross-linking approaches for mass spectrometry-based investigation of nucleic acids and protein-nucleic acid assemblies. *Methods*, **144**, 64-78.

26. van Roon, A.M., Oubridge, C., Obayashi, E., Sposito, B., Newman, A.J., Seraphin, B. and Nagai, K. (2017) Crystal structure of U2 snRNP SF3b components: Hsh49p in complex with Cus1p-binding domain. *RNA*, **23**, 968-981.
27. Johannsson, S., Neumann, P., Wulf, A., Welp, L.M., Gerber, H.D., 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. *Sci Rep*, **8**, 8880.
28. Vos, S.M., Pollmann, D., Caizzi, L., Hofmann, K.B., Rombaut, P., Zimniak, T., Herzog, F. and Cramer, P. (2016) Architecture and RNA binding of the human negative elongation factor. *Elife*, **5**.
29. Wicke, D., Neumann, P., Gößringer, M., Chernev, A., Davydov, S., Poehlein, A., Daniel, R., Urlaub, H., Hartmann, R.K., Ficner, R. *et al.* (2024) The previously uncharacterized RnpM (YlxR) protein modulates the activity of ribonuclease P in Bacillus subtilis in vitro. *Nucleic Acids Res*, **52**, 1404-1419.
30. Watkins, A.M., Rangan, R. and Das, R. (2020) FARFAR2: Improved De Novo Rosetta Prediction of Complex Global RNA Folds. *Structure*, **28**, 963-976 e966.
31. Zok, T., Antczak, M., Zurkowski, M., Popenda, M., Blazewicz, J., Adamiak, R.W. and Szachniuk, M. (2018) RNApdbee 2.0: multifunctional tool for RNA structure annotation. *Nucleic Acids Res*, **46**, W30-W35.
32. Emsley, P., Lohkamp, B., Scott, W.G. and Cowtan, K. (2010) Features and development of Coot. *Acta Crystallogr D Biol Crystallogr*, **66**, 486-501.
33. Chou, F.C., Sripakdeevong, P., Dibrov, S.M., Hermann, T. and Das, R. (2013) Correcting pervasive errors in RNA crystallography through enumerative structure prediction. *Nat Methods*, **10**, 74-76.
34. Mandell, D.J., Coutsias, E.A. and Kortemme, T. (2009) Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nat Methods*, **6**, 551-552.
35. Das, R. and Baker, D. (2007) Automated de novo prediction of native-like RNA tertiary structures. *Proc Natl Acad Sci U S A*, **104**, 14664-14669.
36. UniProt, C. (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*, **47**, D506-D515.
37. Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res*, **37**, D211-215.
38. Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236-1240.
39. Moffat, L. and Jones, D.T. (2021) Increasing the accuracy of single sequence prediction methods using a deep semi-supervised learning framework. *Bioinformatics*, **37**, 3744-3751.
40. Zhao, B., Katuwawala, A., Oldfield, C.J., Dunker, A.K., Faraggi, E., Gsponer, J., Kloczkowski, A., Malhis, N., Mirdita, M. and Obradovic, Z. (2021) DescribePROT: database of amino acid-level protein structure and function predictions. *Nucleic Acids Research*, **49**, D298-D308.
41. Huang, Q., Szklarczyk, D., Wang, M., Simonovic, M. and von Mering, C. (2023) PaxDb 5.0: curated protein quantification data suggests adaptive proteome changes in yeasts. *Molecular & Cellular Proteomics*, **22**.
42. Caudron-Herger, M., Jansen, R.E., Wassmer, E. and Diederichs, S. (2021) RBP2GO: a comprehensive pan-species database on RNA-binding proteins, their interactions and functions. *Nucleic acids research*, **49**, D425-D436.
43. Desiere, F., Deutsch, E.W., King, N.L., Nesvizhskii, A.I., Mallick, P., Eng, J., Chen, S., Eddes, J., Loevenich, S.N. and Aebersold, R. (2006) The peptideatlas project. *Nucleic acids research*, **34**, D655-D658.

44. Davis, E.S., Mu, W., Lee, S., Dozmorov, M.G., Love, M.I. and Phanstiel, D.H. (2023) matchRanges: generating null hypothesis genomic ranges via covariate-matched sampling. *Bioinformatics*, **39**, btad197.
45. R Core Team, R. (2013) R: A language and environment for statistical computing.
46. Faraway, J., Marsaglia, G., Marsaglia, J. and Baddeley, A. (2019) goftest: Classical goodness-of-fit tests for univariate distributions. *R package version*, 1.2-2.
47. Arnold, T.B. and Emerson, J.W. (2011) Nonparametric goodness-of-fit tests for discrete null distributions. *R Journal*, **3**.
48. Dowd, C. (2018) twosamples: Fast permutation based two sample tests. *R package version 1.0. 0*.
49. Benjamini, Y., Hochberg, Y. . (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. . *Journal of the Royal statistical society: series B (Methodological)*, **57(1)**.
50. Kolde, R., Kolde, M. R. . (2015) Package 'pheatmap'. *R package*, **1(7)**.
51. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G. and Laydon, A. (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, **50**, D439-D444.
52. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A. and Potapenko, A. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583-589.
53. Kahraman, A., Malmström, L. and Aebersold, R. (2011) Xwalk: computing and visualizing distances in cross-linking experiments. *Bioinformatics*, **27**, 2163-2164.
54. Szklarczyk, D., Gable, A.L., Nastou, K.C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N.T., Legeay, M., Fang, T., Bork, P. *et al.* (2021) The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res*, **49**, D605-D612.
55. Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M. N., Sergushichev, A. (2016) Fast gene set enrichment analysis. *BioRxiv*, **060012**.
56. Lee, S., Cook, D. and Lawrence, M. (2019) plyranges: a grammar of genomic data transformation. *Genome Biol*, **20**, 4.
57. Ou, J. and Zhu, L.J. (2019) trackViewer: a Bioconductor package for interactive and integrative visualization of multi-omics data. *Nat Methods*, **16**, 453-454.
58. Wickham, H. (2011) ggplot2. *Wiley interdisciplinary reviews: computational statistics*, 180-185.
59. Arnold, J.B., Arnold, M. J. B. . (2015) Package 'ggthemes'.
60. Slowikowski, K. (2024) ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'. .
61. Auguie, B., Antonov, A., Auguie, M. B. . (2017) Package 'gridExtra'. Miscellaneous Functions for "Grid" Graphics.
62. Feldman, M.Y. (1973) Reactions of nucleic acids and nucleoproteins with formaldehyde. *Prog Nucleic Acid Res Mol Biol*, **13**, 1-49.
63. Lu, K., Ye, W., Zhou, L., Collins, L.B., Chen, X., Gold, A., Ball, L.M. and Swenberg, J.A. (2010) Structural Characterization of Formaldehyde-Induced Cross-Links Between Amino Acids and Deoxynucleosides and Their Oligomers. *Journal of the American Chemical Society*, **132**, 3388-3399.
64. Loeber, R., Rajesh, M., Fang, Q., Pegg, A.E. and Tretyakova, N. (2006) Cross-linking of the human DNA repair protein O6-alkylguanine DNA alkyltransferase to DNA in the presence of 1,2,3,4-diepoxybutane. *Chem Res Toxicol*, **19**, 645-654.
65. Rajski, S.R. and Williams, R.M. (1998) DNA Cross-Linking Agents as Antitumor Drugs. *Chemical Reviews*, **98**, 2723-2796.

66. Knörlein, A., Sarnowski, C.P., de Vries, T., Stoltz, M., Götze, M., Aebersold, R., Allain, F.H.T., Leitner, A. and Hall, J. (2022) Nucleotide-amino acid π-stacking interactions initiate photo cross-linking in RNA-protein complexes. *Nature Communications*, **13**, 2719.
67. Hoffman, E.A., Frey, B.L., Smith, L.M. and Auble, D.T. (2015) Formaldehyde Crosslinking: A Tool for the Study of Chromatin Complexes*. *Journal of Biological Chemistry*, **290**, 26404-26411.
68. Pfeuffer, J., Bielow, C., Wein, S., Jeong, K., Netz, E., Walter, A., Alka, O., Nilse, L., Colaianni, P.D., McCloskey, D. *et al.* (2024) OpenMS 3 enables reproducible analysis of large-scale mass spectrometry data. *Nat Methods*, **21**, 365-367.
69. Rost, H.L., Sachsenberg, T., Aiche, S., Bielow, C., Weisser, H., Aicheler, F., Andreotti, S., Ehrlich, H.C., Gutenbrunner, P., Kenar, E. *et al.* (2016) OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat Methods*, **13**, 741-748.
70. Sturm, M. and Kohlbacher, O. (2009) TOPPView: an open-source viewer for mass spectrometry data. *J Proteome Res*, **8**, 3760-3763.
71. The, M., MacCoss, M.J., Noble, W.S. and Kall, L. (2016) Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. *J Am Soc Mass Spectrom*, **27**, 1719-1727.
72. Veit, J., Sachsenberg, T., Chernev, A., Aicheler, F., Urlaub, H. and Kohlbacher, O. (2016) LFQProfiler and RNP(xl): Open-Source Tools for Label-Free Quantification and Protein-RNA Cross-Linking Integrated into Proteome Discoverer. *J Proteome Res*, **15**, 3441-3448.
73. Polasky, D.A., Geiszler, D.J., Yu, F., Li, K., Teo, G.C. and Nesvizhskii, A.I. (2023) MSFragger-Labile: A Flexible Method to Improve Labile PTM Analysis in Proteomics. *Molecular & Cellular Proteomics*, **22**.
74. Spadaccini, R., Reidt, U., Dybkov, O., Will, C., Frank, R., Stier, G., Corsini, L., Wahl, M.C., Luhrmann, R. and Sattler, M. (2006) Biochemical and NMR analyses of an SF3b155-p14-U2AF-RNA interaction network involved in branch point definition during pre-mRNA splicing. *RNA*, **12**, 410-425.
75. Fischer, N., Neumann, P., Konevega, A.L., Bock, L.V., Ficner, R., Rodnina, M.V. and Stark, H. (2015) Structure of the E. coli ribosome–EF-Tu complex at <3 Å resolution by Cs-corrected cryo-EM. *Nature*, **520**, 567-570.
76. Shi, Z., Yang, W.Z., Lin-Chao, S., Chak, K.F. and Yuan, H.S. (2008) Crystal structure of Escherichia coli PNPase: central channel residues are involved in processive RNA degradation. *RNA*, **14**, 2361-2371.
77. Dendooven, T., Sinha, D., Roeselová, A., Cameron, T.A., De Lay, N.R., Luisi, B.F. and Bandyra, K.J. (2021) A cooperative PNPase-Hfq-RNA carrier complex facilitates bacterial riboregulation. *Molecular Cell*, **81**, 2901-2913.e2905.
78. Bruce, H.A., Du, D., Matak-Vinkovic, D., Bandyra, K.J., Broadhurst, R.W., Martin, E., Sobott, F., Shkumatov, A.V. and Luisi, B.F. (2018) Analysis of the natively unstructured RNA/protein-recognition core in the Escherichia coli RNA degradosome and its interactions with regulatory RNA/Hfq complexes. *Nucleic Acids Res*, **46**, 387-402.
79. Huppertz, I., Perez-Perri, J.I., Mantas, P., Sekaran, T., Schwarzl, T., Russo, F., Ferring-Appel, D., Koskova, Z., Dimitrova-Paternoga, L., Kafkia, E. *et al.* (2022) Riboregulation of Enolase 1 activity controls glycolysis and embryonic stem cell differentiation. *Mol Cell*, **82**, 2666-2680 e2611.
80. Castello, A., Hentze, M.W. and Preiss, T. (2015) Metabolic Enzymes Enjoying New Partnerships as RNA-Binding Proteins. *Trends Endocrinol Metab*, **26**, 746-757.
81. Wegener, M. and Dietz, K.J. (2022) The mutual interaction of glycolytic enzymes and RNA in post-transcriptional regulation. *Rna*, **28**, 1446-1468.
82. McLean, J.E., Hamaguchi, N., Belenky, P., Mortimer, S.E., Stanton, M. and Hedstrom, L. (2004) Inosine 5'-monophosphate dehydrogenase binds nucleic acids in vitro and in vivo. *Biochem J*, **379**, 243-251.

83. Cornuel, J.F., Moraillon, A. and Gueron, M. (2002) Participation of yeast inosine 5'-monophosphate dehydrogenase in an in vitro complex with a fragment of the C-rich telomeric strand. *Biochimie*, **84**, 279-289.
84. Gedeon, A., Ayoub, N., Brule, S., Raynal, B., Karimova, G., Gelin, M., Mechaly, A., Haouz, A., Labesse, G. and Munier-Lehmann, H. (2023) Insight into the role of the Bateman domain at the molecular and physiological levels through engineered IMP dehydrogenases. *Protein Sci*, **32**, e4703.
85. Nejad, M.I., Johnson, K.M., Price, N.E. and Gates, K.S. (2016) A New Cross-Link for an Old Cross-Linking Drug: The Nitrogen Mustard Anticancer Agent Mechlorethamine Generates Cross-Links Derived from Abasic Sites in Addition to the Expected Drug-Bridged Cross-Links. *Biochemistry*, **55**, 7033-7041.
86. Thompson, P.S. and Cortez, D. (2020) New insights into abasic site repair and tolerance. *DNA Repair (Amst)*, **90**, 102866.
87. Jiang, W., Hou, Y. and Inouye, M. (1997) CspA, the major cold-shock protein of Escherichia coli, is an RNA chaperone. *J Biol Chem*, **272**, 196-202.
88. Phadtare, S. and Inouye, M. (1999) Sequence-selective interactions with RNA by CspB, CspC and CspE, members of the CspA family of Escherichia coli. *Mol Microbiol*, **33**, 1004-1014.
89. Bi, H. and Zhang, C. (2014) Integration host factor is required for the induction of acid resistance in Escherichia coli. *Curr Microbiol*, **69**, 218-224.
90. Chen, S.-J. and Wang, J.C. (1998) Identification of Active Site Residues in Escherichia coli DNA Topoisomerase I*. *Journal of Biological Chemistry*, **273**, 6050-6056.
91. Salgado, H., Gama-Castro, S., Lara, P., Mejia-Almonte, C., Alarcón-Carranza, G., López-Almazo, A.G., Betancourt-Figueroa, F., Peña-Loredo, P., Alquicira-Hernández, S., Ledezma-Tejeida, D. *et al.* (2024) RegulonDB v12.0: a comprehensive resource of transcriptional regulation in E. coli K-12. *Nucleic Acids Res*, **52**, D255-d264.
92. Chu, L.-C., Arede, P., Li, W., Urdaneta, E.C., Ivanova, I., McKellar, S.W., Wills, J.C., Fröhlich, T., von Kriegsheim, A., Beckmann, B.M. *et al.* (2022) The RNA-bound proteome of MRSA reveals post-transcriptional roles for helix-turn-helix DNA-binding and Rossmann-fold proteins. *Nature Communications*, **13**, 2883.
93. Esteban-Serna, S., McCaughan, H. and Granneman, S. (2023) Advantages and limitations of UV cross-linking analysis of protein–RNA interactomes in microbes. *Molecular Microbiology*, **120**, 477-489.
94. Heindel, A.J., Brulet, J.W., Wang, X., Founds, M.W., Libby, A.H., Bai, D.L., Lemke, M.C., Leace, D.M., Harris, T.E., Hafner, M. *et al.* (2023) Chemoproteomic capture of RNA binding activity in living cells. *Nature Communications*, **14**, 6282.
95. Kong, A.T., Leprevost, F.V., Avtonomov, D.M., Mellacheruvu, D. and Nesvizhskii, A.I. (2017) MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat Methods*, **14**, 513-520.
96. Stenum, T.S., Kumar, A.D., Sandbaumhüter, F.A., Kjellin, J., Jerlström-Hultqvist, J., Andrén, P.E., Koskiniemi, S., Jansson, E.T. and Holmqvist, E. (2023) RNA interactome capture in Escherichia coli globally identifies RNA-binding proteins. *Nucleic Acids Res*, **51**, 4572-4587.
97. Monti, M., Herman, R., Mancini, L., Capitanchik, C., Davey, K., Dawson, C.S., Ule, J., Thomas, G.H., Willis, A.E., Lilley, K.S. *et al.* (2024) Interrogation of RNA-protein interaction dynamics in bacterial growth. *Mol Syst Biol*, **20**, 573-589.
98. Vos, S.M., Farnung, L., Urlaub, H. and Cramer, P. (2018) Structure of paused transcription complex Pol II–DSIF–NELF. *Nature*, **560**, 601-606.
99. Rao, Jampani N., Neumann, L., Wenzel, S., Schweimer, K., Rösch, P. and Wöhrl, Birgitta M. (2006) Structural studies on the RNA-recognition motif of NELF E, a cellular negative transcription elongation factor involved in the regulation of HIV transcription. *Biochemical Journal*, **400**, 449-456.

100. Prats-Ejarque, G., Blanco, J.A., Salazar, V.A., Nogués, V.M., Moussaoui, M. and Boix, E. (2019) Characterization of an RNase with two catalytic centers. Human RNase6 catalytic and phosphate-binding site arrangement favors the endonuclease cleavage of polymeric substrates. *Biochimica et Biophysica Acta (BBA) - General Subjects*, **1863**, 105-117.
101. Joyce, M.A., Fraser, M.E., James, M.N.G., Bridger, W.A. and Wolodko, W.T. (2000) ADP-Binding Site of Escherichia coli Succinyl-CoA Synthetase Revealed by X-ray Crystallography. *Biochemistry*, **39**, 17-25.
102. Ai, H., Sun, M., Liu, A., Sun, Z., Liu, T., Cao, L., Liang, L., Qu, Q., Li, Z., Deng, Z. *et al.* (2022) H2B Lys34 Ubiquitination Induces Nucleosome Distortion to Stimulate Dot1L Activity. *Nature Chemical Biology*, **18**, 972-980.
103. Wright, A.V., Liu, J.J., Knott, G.J., Doxzen, K.W., Nogales, E. and Doudna, J.A. (2017) Structures of the CRISPR genome integration complex. *Science*, **357**, 1113-1118.
104. Hammel, M., Amlanjyoti, D., Reyes, F.E., Chen, J.H., Parpana, R., Tang, H.Y., Larabell, C.A., Tainer, J.A. and Adhya, S. (2016) HU multimerization shift controls nucleoid compaction. *Sci Adv*, **2**, e1600650.
105. Stella, S., Cascio, D. and Johnson, R.C. (2010) The shape of the DNA minor groove directs binding by the DNA-bending protein Fis. *Genes Dev*, **24**, 814-826.
106. Tan, K., Zhou, Q., Cheng, B., Zhang, Z., Joachimiak, A. and Tse-Dinh, Y.-C. (2015) Structural basis for suppression of hypernegative DNA supercoiling by E. coli topoisomerase I. *Nucleic Acids Research*, **43**, 11031-11046.