

## SUPPLEMENTARY INFORMATION

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

#### AUTHORS

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

<sup>1</sup> Bioanalytical Mass Spectrometry Group, Max Planck Institute for Multidisciplinary Sciences, 37077 Göttingen, Germany

<sup>2</sup> Bioanalytics Group, Department of Clinical Chemistry, University Medical Center Göttingen, Göttingen, 37075, Germany

<sup>3</sup> Institute for Bioinformatics and Medical Informatics, University of Tübingen, Tübingen, 72076, Germany

<sup>4</sup> Applied Bioinformatics, Dept. for Computer Science, University of Tübingen, Tübingen, 72076, Germany

<sup>5</sup> Quantitative and Systems Biology, Max Planck Institute for Multidisciplinary Sciences, Göttingen, 37077, Germany

<sup>6</sup> Department of Molecular Structural Biology, Institute of Microbiology and Genetics, GZMB, Georg-August-University Göttingen, Göttingen, 37077, Germany

<sup>7</sup> Max Planck Institute for Multidisciplinary Sciences, Department of Molecular Biology, Göttingen, 37077, Germany

<sup>8</sup> Institute for Translational Bioinformatics, University Hospital Tübingen, Tübingen, 72076, Germany

<sup>9</sup> Thermo Fisher Scientific, San Jose, CA, 95134, USA

<sup>10</sup> Thermo Fisher Scientific, Bremen, 28199, Germany

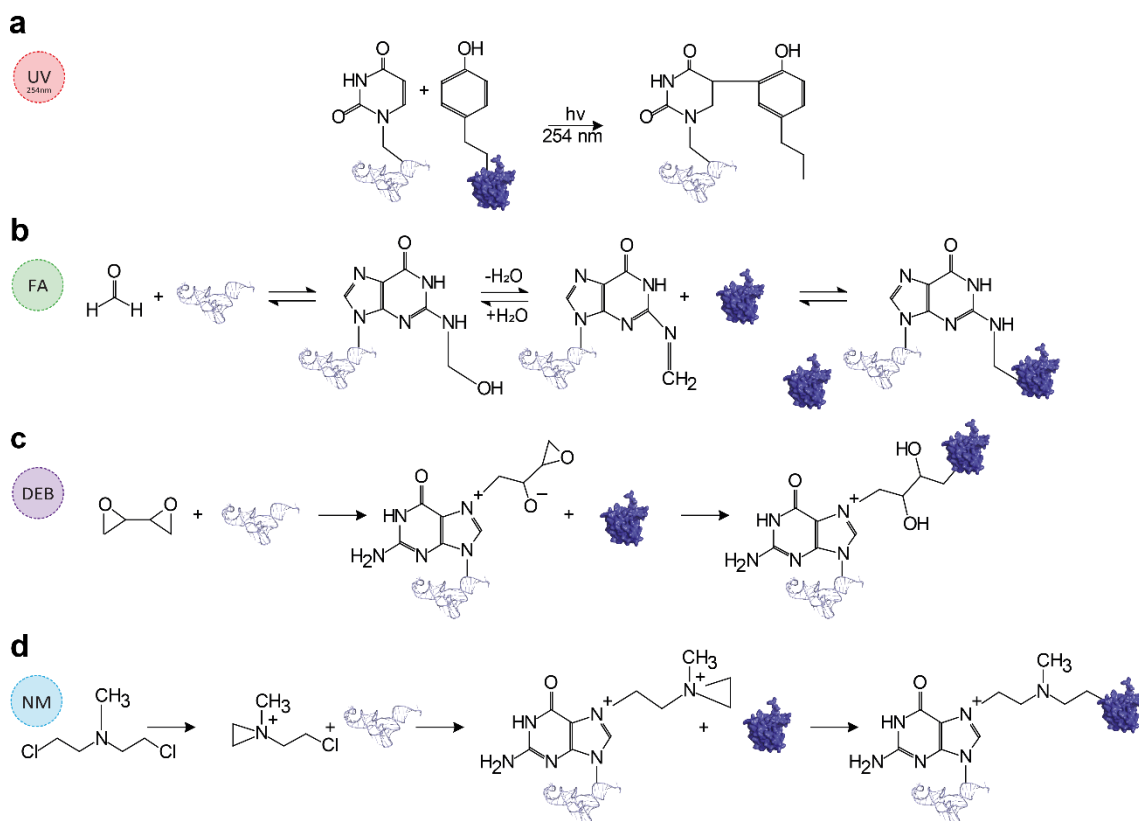
<sup>11</sup> Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA

<sup>12</sup> 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

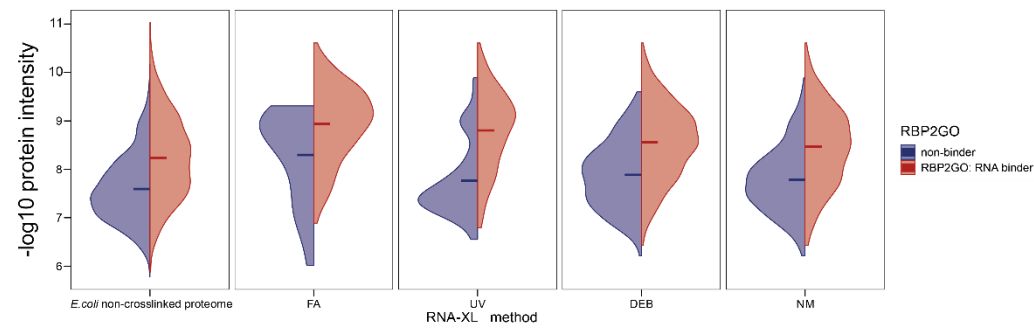
## SUPPLEMENTARY FIGURES



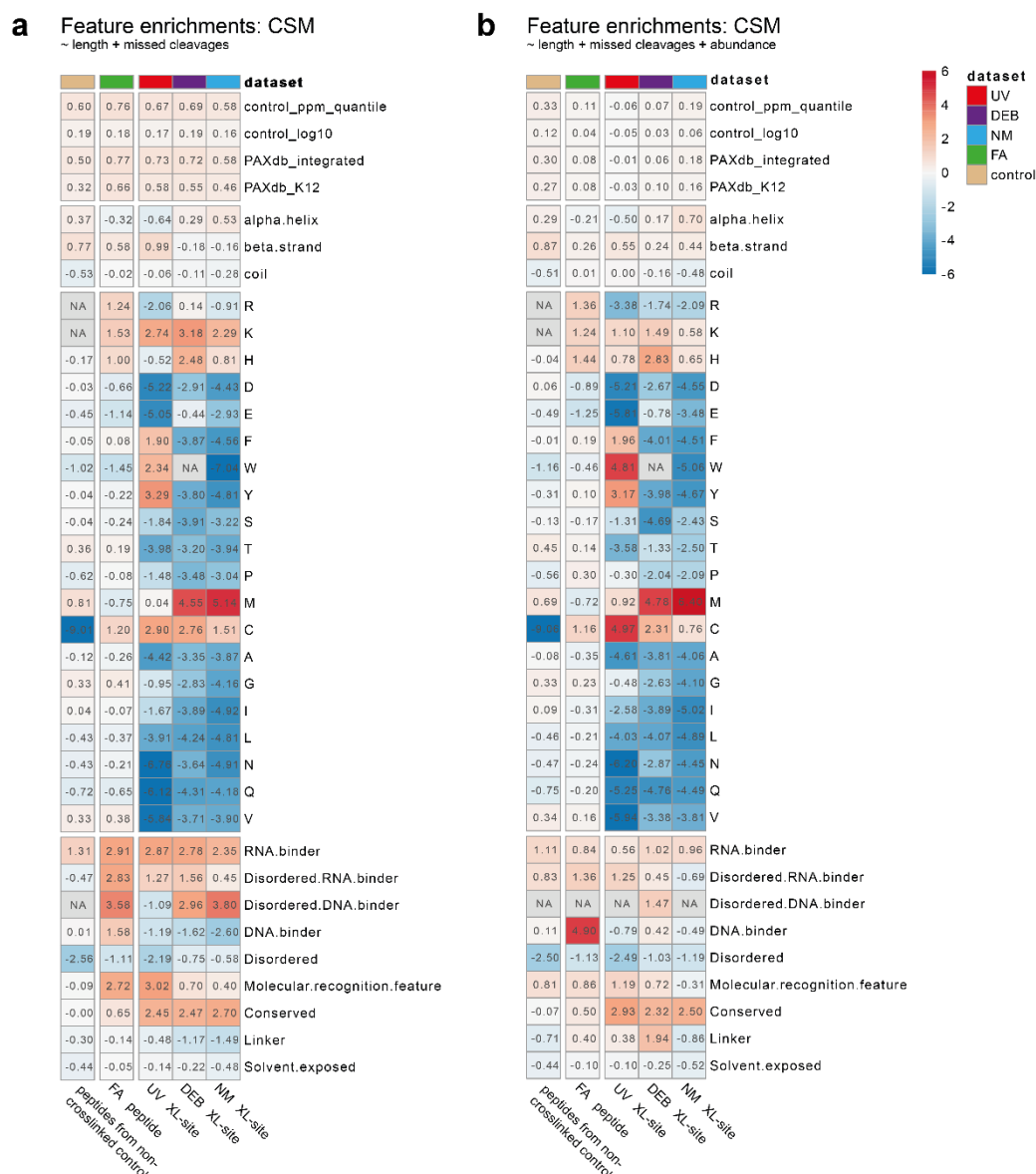
Supplementary Figure S1. Proposed chemical protein-RNA crosslinking reaction mechanisms. *Reaction mechanisms for (a)* UV light at  $\lambda=254$  nm (adapted from Kramer *et al.*, 2011(1)), *(b)* formaldehyde (FA) (adapted from Hoffman *et al.*, 2015(2)), *(c)* 1,2:3,4-diepoxybutane (DEB) (adapted from Gherezghiher *et al.*, 2013(3)), *(d)* mechlorethamine (nitrogen mustard, NM) (adapted from Tretyakova *et al.*, 2015(4)).

### Detection of RNA-interacting proteins across RNA-XL methods

RNA-related proteins are typically high-abundant in *E. coli*

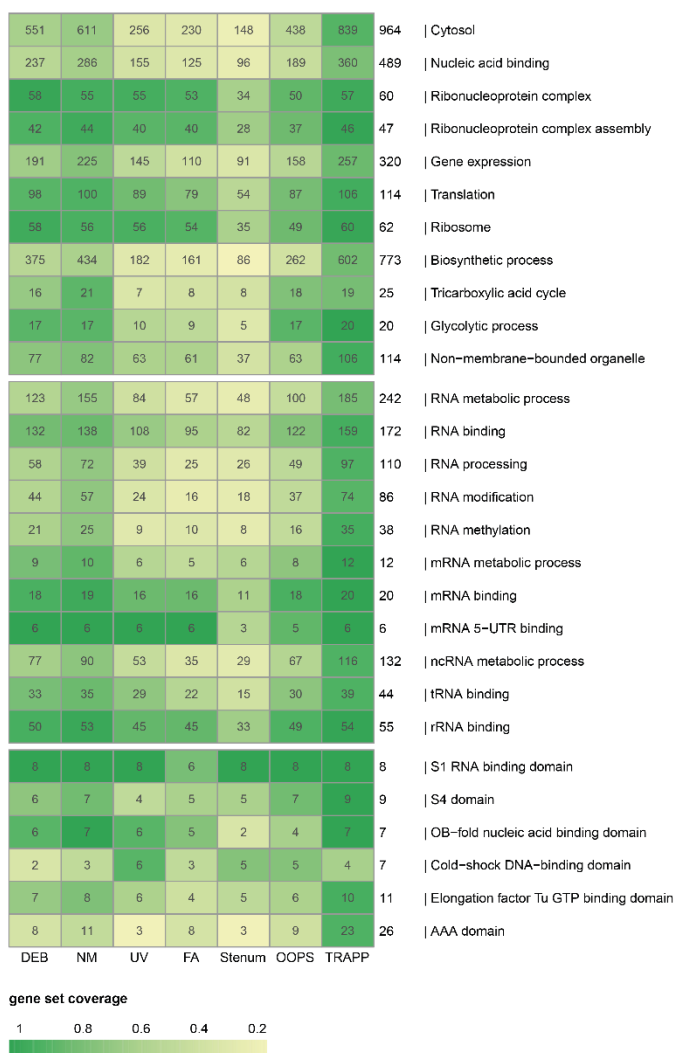


Supplementary Figure S2. Crosslinked protein abundances. Distributions of the *E. coli* protein abundance measured by label-free MS1 intensity and stratified by the annotation in Caudron-Herger et al., 2021 and detection across the crosslinking methods(5). Highlighted are the distribution means.

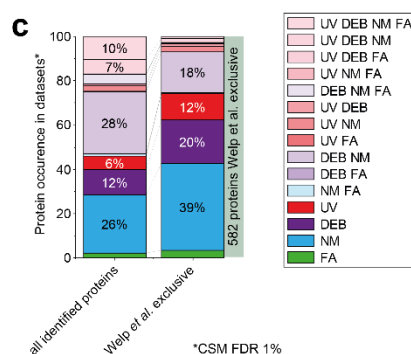
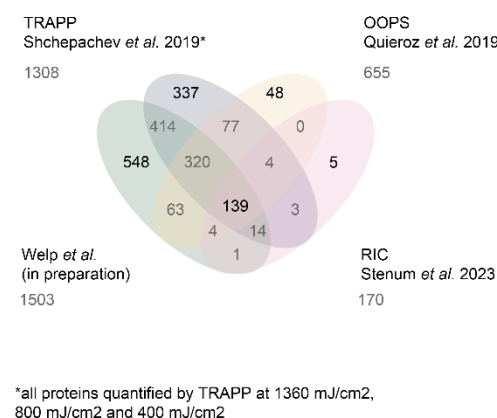


Supplementary Figure S3. The log2FoldChanges between the experimental and an *in-silico* background distribution mean feature counts per residue. The background sampling was performed with regards to: **(a)** the experimentally observed peptide length and missed cleavage distribution and **(b)** the experimentally observed peptide length, missed cleavage distribution and source protein abundance. The tryptic and FA datasets were enriched on the peptide level, whilst the UV, NM and DEB enrichment was done 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. Crosslinker specific feature enrichment. Control, non-crosslinked *E. coli* proteome.

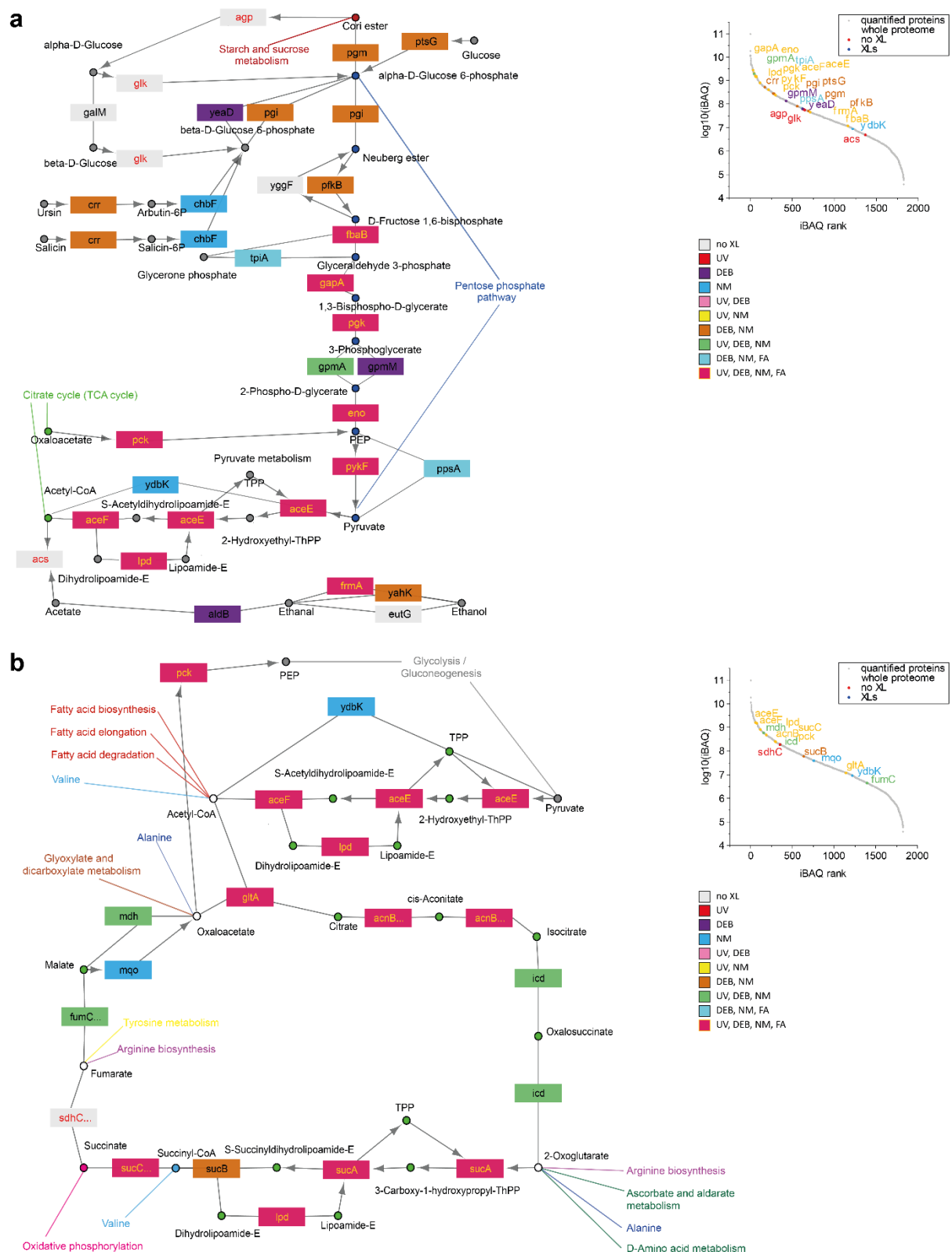
## a Significant sets



## b No. of crosslinked proteins

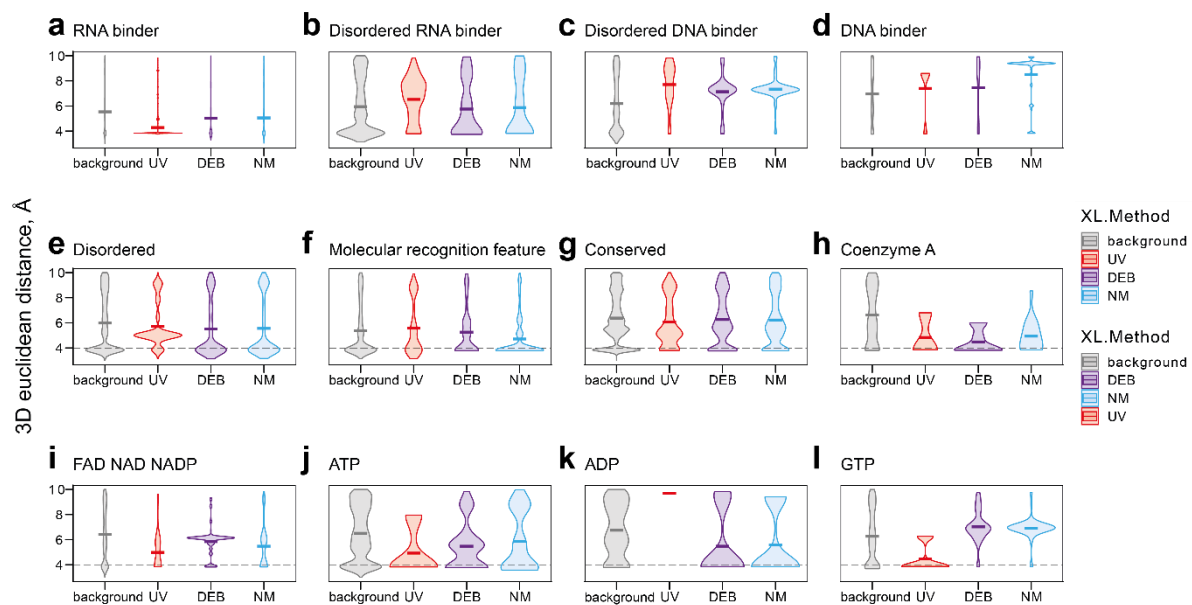


Supplementary Figure S4. Protein-RNA XL-MS dataset comparison. **(a)** GO term and domain gene set coverage by *E. coli* UV, DEB, NM and FA XL-MS datasets presented in this study and *E. coli* XL-MS datasets from (6-8). The terms are selected from those significantly enriched in at least one of the UV, DEB, NM or FA XL-MS datasets ( $p_{adj} \leq 0.01$  for GO terms;  $p_{adj} \leq 0.05$  for domains). **(b)** Venn diagram of *E. coli* RNA-crosslinked proteins identified in this study compared to proteins identified as RNA-binding in Stenum *et al.* 2023 (RIC, protocol used)(8), Quieroz *et al.* 2019 (OOPS)(7) and/or Shchepachev *et al.* 2029 (TRAPP)(6). **(c)** 100% stacked bar plot of all RNA-crosslinked proteins identified in this study and proteins found crosslinked to RNA but not identified as interaction partners in OOPS or TRAPP (Welp *et al.* exclusive). Relative numbers of proteins occurring in UV, DEB, NM and FA are indicated.



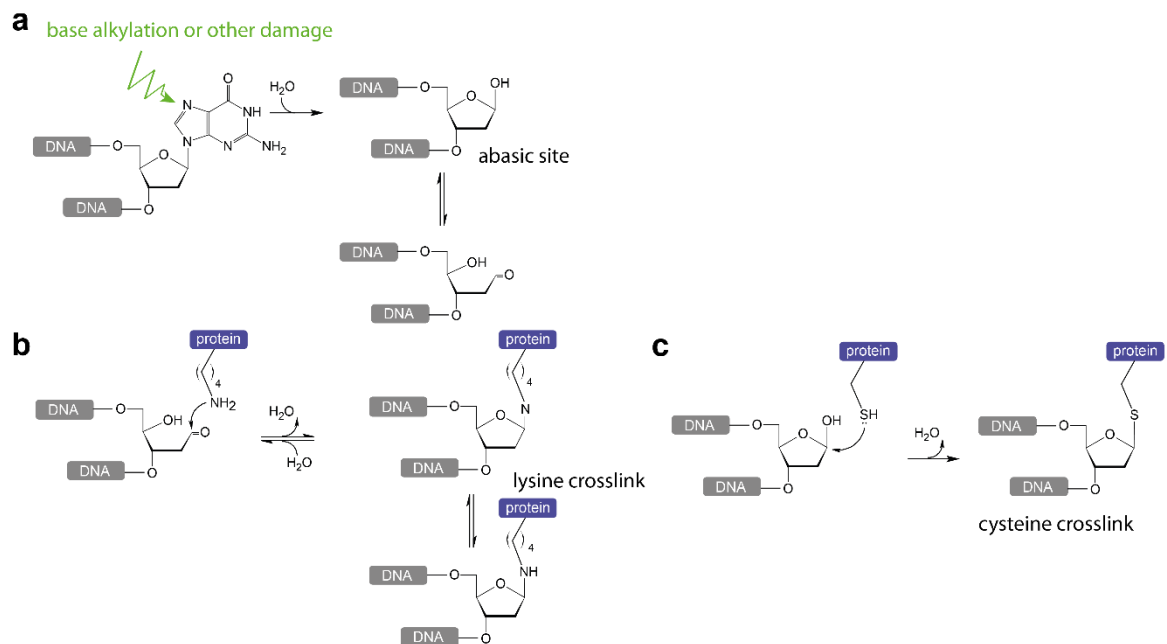
Grey boxes indicate proteins not found as crosslinked. Genes highlighted by red font colour were not identified as crosslinked but were quantified in the proteomics dataset. Right panel: Ranked plot of proteins quantified in the *E. coli* proteomics control experiment. Font colour code matches the rectangle colour code in KEGG pathway. Glycolysis-involved proteins not found as crosslinked are indicated by red font color. **(b)** KEGG pathway of the tricarboxylic acid cycle (TCA-cycle) as in a. Right panel: Ranked plot as in a with TCA-cycle-involved proteins highlighted as in a.

# Euclidean distances: global *in silico* background and RNA-XL sites (CSM)

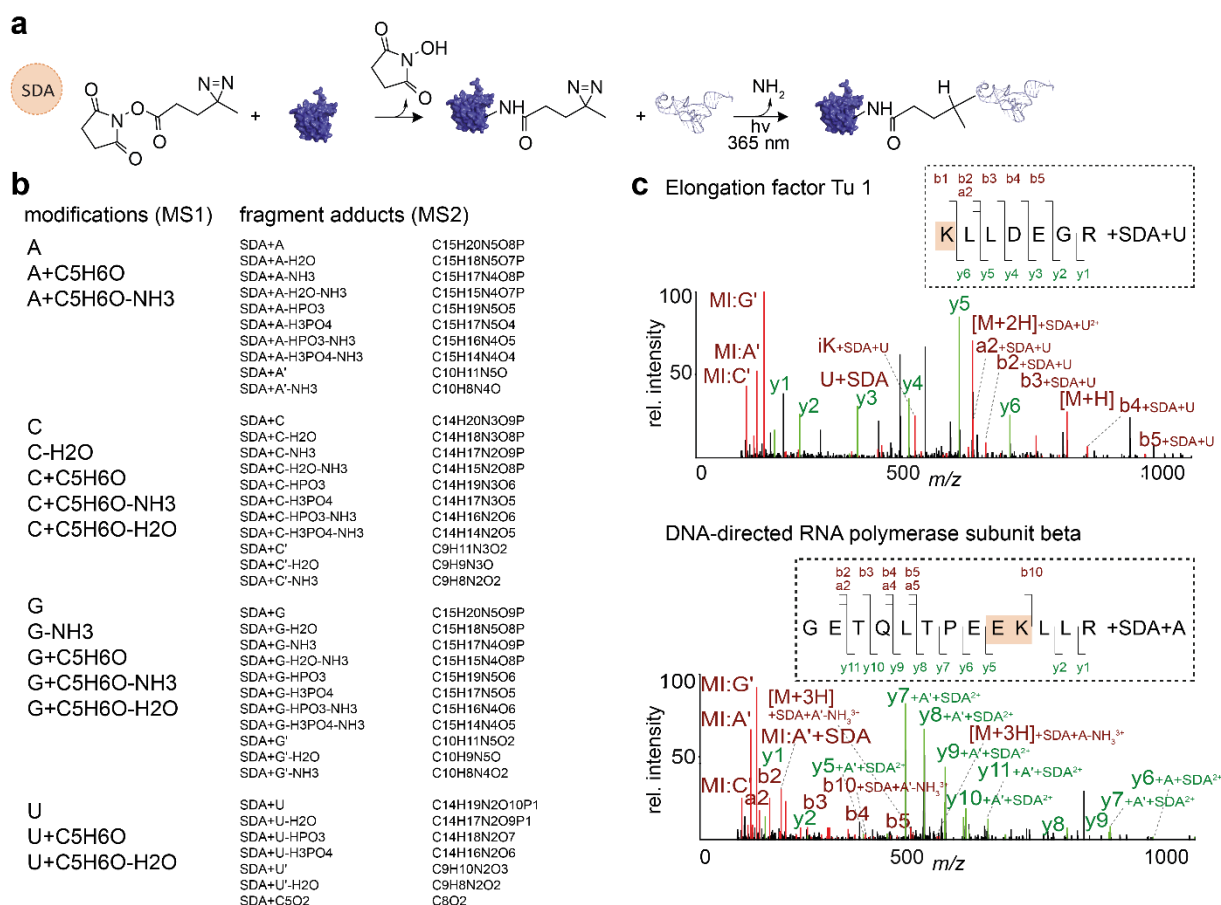


Supplementary Figure S6. 3D distances of crosslinked amino acids to protein sequence features. The distributions of 3D Euclidean distances between the C $\alpha$  atom pairs of crosslinked residues and nearest feature residue C $\alpha$  annotated in DESCRIBEPROT (**a-g**) and InterPro (**h-l**) are shown. Highlighted are the group means. The 4 Å range is marked, corresponding to the typical distance to the nearest residue across the proteome.





Supplementary Figure S7. Depurination and abasic site reactions in DNA. **(a)** Guanosine base alkylation or other forms of base damage weakening N-glycosidic bond, leading to base hydrolysis. Abasic sites exist in ring-opened aldehyde and the closed hemi-acetal form. **(b)** The acetal form of an abasic site can be attacked at the C1' position by the nucleophilic  $\epsilon$ -amino group of lysines to create a crosslink via a reversible Schiff-base intermediate(9). **(c)** The acetal form of an abasic site can be attacked at the C1' position by the nucleophilic thiol group in cysteines to form a stable covalent linkage(10).



Supplementary Figure S8. Protein-nucleic acid XL-MS using SDA. **(a)** Proposed chemical protein-RNA crosslinking reaction mechanism for SDA. **(b)** Sum formula of SDA RNA-crosslink adducts for NuXL analysis. Expected modifications on MS1 level and fragment adducts in MS2 scans. **(c)** MS2 crosslink spectra for SDA-crosslinked peptide-RNA-(oligo)nucleotides from *E. coli* ribosomes resulting from SDA XL-MS and NuXL analysis using settings listed in b. Upper panel: MS2 spectrum of elongation factor Tu 1 peptide KLLDEGR crosslinked to uridine-monophosphate by SDA. Lower panel: MS2 spectrum of DNA-directed RNA polymerase subunit beta peptide GETQLTPEEKLLR crosslinked to adenosine-monophosphate by SDA. Annotated peaks are highlighted in green (y-ions) or red (marker ions and a-/b-ions). Crosslinked amino acids are highlighted in orange.

## SUPPLEMENTARY TABLES

Supplementary Table S1. Crosslinker-specific presets in NuXL. Table including all sum formula for RNA/DNA adducts defined as presets in NuXL. Sheets contain different presets. Columns include: target\_nucleotides, definitions (sum formula) of monophosphate-nucleotides/deoxyribose; modifications, definitions (sum formula) of RNA/DNA adducts on peptide precursor (MS1); fragment\_adducts, definitions (sum formula) of RNA/DNA adducts on peptide fragments (MS2).

Supplementary Table S2. Protein-RNA crosslinks in *E. coli* ribosomes. Table including RNA-crosslink spectrum matches (CSMs) from UV, DEB, NM and FA XL-MS data from purified *E. coli* ribosomes. CSMs are filtered for 1% FDR and rescored by Percolator. Among other values, the table includes: idxml, source file; index, spectrum index number; RT, retention time; precursor *m/z*; score, main score calibrated by percolator algorithm using subscores; PeakAnnotations, MS2 peaks *m/z* values, annotations and intensities; NuXL:NA, RNA adduct on peptide precursor (MS1); NuXL:NT, crosslinked nucleobase.

Supplementary Table S3. List of all subscores used in Percolator-based rescoring. List of all subscores used in Percolator-based rescoring.

Supplementary Table S4. Protein-RNA crosslinks in *in vitro* protein-RNA complexes. Table including RNA-crosslink spectrum matches (CSMs) from UV, DEB, NM and FA XL-MS data from *in vitro* analysed protein-RNA complexes. Sheets include results from: Hsh49-Cus1-U2RNA; Dnmt2-tRNA<sup>Asp</sup>, NELF-TAR complexes separately. CSMs are filtered for 1% FDR and rescored by Percolator. Among other values, the table includes: idxml, source file; index, spectrum index number; RT, retention time; precursor *m/z*; score, main score calibrated by percolator algorithm using subscores; PeakAnnotations, MS2 peaks *m/z* values, annotations and intensities; NuXL:NA, RNA adduct on peptide precursor (MS1); NuXL:NT, crosslinked nucleobase.

Supplementary Table S5. Protein-RNA crosslinks from *E. coli*. Table including RNA-crosslink spectrum matches (CSMs) from UV, DEB, NM and FA XL-MS data from *E. coli* cells. CSMs are filtered for 1% FDR and rescored by Percolator. Among other values, the table includes: fraction, S30/S100 fraction of lysate; idxml, source file; index, spectrum index number; RT,

retention time; precursor  $m/z$ ; score, main score calibrated by percolator algorithm using subscores; PeakAnnotations, MS2 peaks  $m/z$  values, annotations and intensities; NuXL:NA, RNA adduct on peptide precursor (MS1); NuXL:NT, crosslinked nucleobase.

Supplementary Table S6. Label-free quantitative *E. coli* proteomics results. Table including modified proteingroups.txt output from MaxQuant search of *E. coli* label-free quantitative proteomics dataset. Among others, columns include iBAQ values for identified proteins.

Supplementary Table S7. Amino acid / sequence feature tests. Table with enrichment analysis of RNA-crosslinks against sequence features of *E. coli* proteome. Secondary structure counts from S4PRED, amino acid counts, feature counts from DESCRIBEPROT and InterPro are included together with the distances to the next feature in 10 Å radius (1D, 3D-euclidean and 3D-SAS distances). The analysis is performed on the detected peptide and crosslinked residue levels, with and without consideration of crosslink-spectrum match counts in a variety of background assumptions. Residue-level enrichment only considers the CSMs with the NuXL best localization threshold  $\geq 1$ . Provided are the average values of the detected crosslinks, expected background values, t-test outputs, nonparametric goodness-of-fit test outputs, adjusted p-values for both tests (BH procedure) and sample sizes.

Supplementary Table S8. Gene set overrepresentation analysis. Table with a hypergeometric test overrepresentation analysis of identified *E. coli* proteins across crosslinking methods. The proteins were filtered for the detection of at least 1 unique peptide sequence. Provided are the gene set names together with the overlap size, significance of the overlap and the estimates of the gene set redundancy.

Supplementary Table S9. Proximal feature enrichment. Table with significance of co-localization of crosslinked residues and annotated protein sequence features. Reported are the outcomes of binomial tests with feature frequency weights: set sizes of foreground and background within and outside the 4 Å radius, the p-value of enrichment, adjusted p-value (BH method) and 0.95 confidence interval.

Supplementary Table S10. Protein-DNA crosslinks in HeLa nucleosomes. Table including DNA-crosslink spectrum matches (CSMs) from UV, DEB, NM and FA XL-MS data from purified HeLa nucleosomes. CSMs are filtered for 1% FDR and rescored by Percolator. Among other values,

the table includes: idxml, source file; index, spectrum index number; RT, retention time; precursor  $m/z$ ; score, main score calibrated by percolator algorithm using subscores; PeakAnnotations, MS2 peaks  $m/z$  values, annotations and intensities; NuXL:NA, DNA adduct on peptide precursor (MS1); NuXL:NT, crosslinked nucleobase/deoxyribose.

Supplementary Table S11. Protein-DNA crosslinks from *E. coli*. Table including DNA-crosslink spectrum matches (CSMs) from UV, DEB, NM and FA XL-MS data from *E. coli* cells. CSMs are filtered for 1% FDR and rescored by Percolator. Among other values, the table includes: fraction, S30/S100 fraction of lysate; idxml, source file; index, spectrum index number; RT, retention time; precursor  $m/z$ ; score, main score calibrated by percolator algorithm using subscores; PeakAnnotations, MS2 peaks  $m/z$  values, annotations and intensities; NuXL:NA, DNA adduct on peptide precursor (MS1); NuXL:NT, crosslinked nucleobase.

Supplementary Table S12. Ambiguous RNA/DNA mass adducts in XL-MS. Table containing ambiguous RNA/DNA mass adducts in XL-MS. Sum formula corresponding to ambiguous mass adducts are listed with respective monoisotopic masses and possible RNA/DNA adducts.

## **SUPPLEMENTARY DATA**

Supplementary Data S1. Visualisation of RNA-crosslinks in *E. coli*. PDF file displaying 1D Sequence representations of all crosslinked proteins from UV, DEB, NM and FA XL-MS *E. coli* *in vivo* datasets containing at least one crosslink site with a localisation score  $\geq 1$ . The header provides information on: protein accession, protein name, protein abundance (log10 iBAQ values) from label-free quantitative proteomics dataset presented in this study ("tryptic"), PAXdb K12 strain protein abundance in ppm, PAXdb *E. coli* protein abundance in ppm. For all abundance values, the respective abundance quantile the values correspond to is given. The plot includes the protein sequence plotted schematically in 1D, including positions of UV, DEB and NM crosslink sites (localisation score  $\geq 1$ ) and indicated CSM counts (y axis), UV, DEB, NM and FA crosslinked peptide positions, potential tryptic cleavages.

1. Kramer, K., Hummel, P., Hsiao, H.-H., Luo, X., Wahl, M. and Urlaub, H. (2011) Mass-spectrometric analysis of proteins cross-linked to 4-thio-uracil- and 5-bromo-uracil-substituted RNA. *International Journal of Mass Spectrometry*, **304**, 184-194.
2. Hoffman, E.A., Frey, B.L., Smith, L.M. and Auble, D.T. (2015) Formaldehyde crosslinking: a tool for the study of chromatin complexes. *J Biol Chem*, **290**, 26404-26411.
3. 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.
4. 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.
5. 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 Res*, **49**, D425-d436.
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. 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.
9. Sczepanski, J.T., Wong, R.S., McKnight, J.N., Bowman, G.D. and Greenberg, M.M. (2010) Rapid DNA-protein cross-linking and strand scission by an abasic site in a nucleosome core particle. *Proc Natl Acad Sci U S A*, **107**, 22475-22480.
10. Chan, W., Ham, Y.-H., Jin, L., Chan, H.W., Wong, Y.-L., Chan, C.-K. and Chung, P.-Y. (2019) Quantification of a Novel DNA-Protein Cross-Link Product Formed by Reacting Apurinic/Apyrimidinic Sites in DNA with Cysteine Residues in Protein by Liquid Chromatography-Tandem Mass Spectrometry Coupled with the Stable Isotope-Dilution Method. *Analytical Chemistry*, **91**, 4987-4994.