Supporting Information

Intensity and retention time prediction improves the rescoring of protein-nucleic acid cross-links.

Arslan Siraj[1,6], Robbin Bouwmeester[2,3], Arthur Declercq[2,3], Luisa Welp[4,5], Aleksandar Chernev[4], Alexander Wulf[4], Henning Urlaub[4,5], Lennart Martens[2,3], Sven Degroeve[2,3], Oliver Kohlbacher[1,6], Timo Sachsenberg[1,6,*]

[1] Department of Computer Science, Applied Bioinformatics, University of Tübingen, Tübingen, 72076, Germany; [2] Department of Biomolecular Medicine, Ghent University, Belgium; [3] VIB-UGent Center for Medical Biotechnology, VIB, Belgium; [4] Bioanalytical Mass Spectrometry, Max Planck Institute for Multidisciplinary Sciences, Göttingen, 37077, Germany; [5] Bioanalytics, Institute of Clinical Chemistry, University Medical Center Göttingen, Göttingen, 37075, Germany; [6] Institute for Biological and Medical Informatics, University of Tübingen, Tübingen, 72076, Germany

* corresponding author: timo.sachsenberg@uni-tuebingen.de

Table S1: NuXL search parameters for the individual protocols; The default parameter of NuXL search engine parameters is provided in NuXL.ini file on GitHub in addition to the search databases in fasta format.

Datasets	**Common search parameters**	**Preset settings**
In-house E. coli dataset	-peptide:max_size 30 -peptide:min_size 05 -NuXL:length 2 -report:top_hits 4 -database Ecoli_SwPr_canon_20220310_4400seq_MQ2030contaminants.fasta	**RNA_UV:** -NuXL:presets "RNA-UV (U)" **RNA_NM:** -NuXL:presets "RNA-NM" **RNA_DEB:** -NuXL:presets "RNA-DEB**"** **RNA_4SU:** -NuXL:presets "RNA-UV (4SU)"
iTRRAP	-modifications:fixed 'Carbamidomethyl (C)' -NuXL:length 4 -report:top_hits 4 -database Swiss_S.cerevisiae_06032022.fasta	-NuXL:presets "RNA-UV Extended (4SU) "
RBSID/ pRBSID	-modifications:variable "Oxidation (M)" "Carbamidomethyl (C)" -report:top_hits 4 -NuXL:length 1 -database Swiss_Human_uniprot_05032022.fasta	**RBSID:** -NuXL:presets "RNA-UV Extended (U)" **pRBSID:** -NuXL:presets 'RNA-UV Extended (4SU)'

Table S2: Train/test split for DeepLC model. Number of CSMs (1% CSM-level FDR). In case multiple CSMs for one cross-link peptide was found, only the most abundant one was considered.

Protocol	**Train**	**Test**
UV	3833	761
DEB	6751	1166
NM	6726	1282
4SU	2676	1158
All	19986	4367

Table S3: Hyperparameter settings for deeplcretrainer to fine tune the DeepLC model; (freeze_after_concat: number of layers after concatenation layer frozen in training).

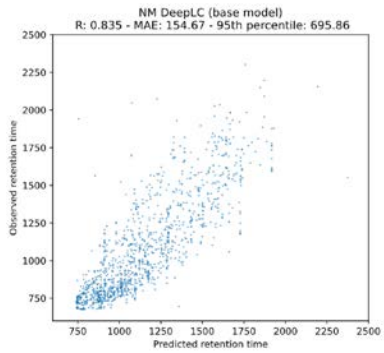
Model	**freeze_after_concat**	**freeze_layers**	**epochs**
Generic	0	True	50
NM	1	True	15
DEB	1	True	15
UV	1	True	15
4SU	1	True	20

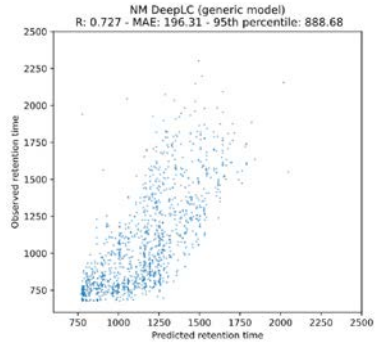
Table S4: Retention time and intensity features added for rescoring with percolator.

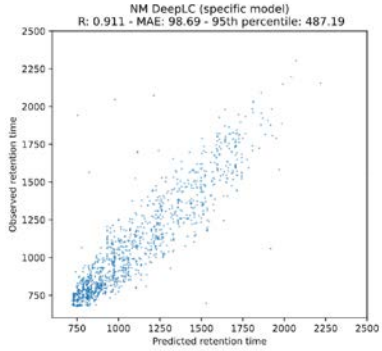
Feature	**Description**
rt_diff	absolute difference between observed and predicted retention time
rt_diff_best	best retention time difference by selecting the minimum rt_diff of the same peptidoform
observed_retention_time_best	best observed retention time by selecting minimum rt_diff of same peptidoform
predicted_retention_time_best	best predicted retention time by selecting minimum rt_diff of same peptidoform
b_y_corr	max value from calculated pearson correlation between predicted vs target intensities of b and y ions

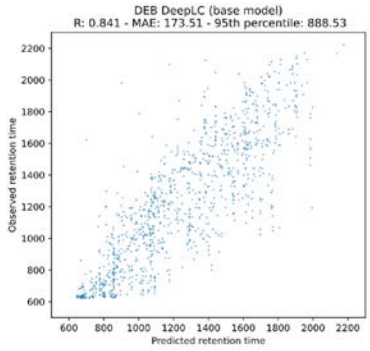
Table S5: Performance of base model (no fine-tuning), generic model, and specific models evaluated on test dataset. R: correlation coefficient, MAE: mean absolute error

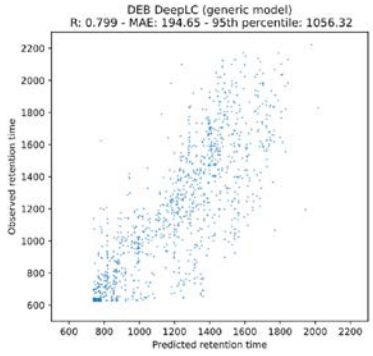
protocol	**base model**		**generic model**		**specific model**	
	R	**MAE**	**R**	**MAE**	**R**	**MAE**
NM	0.835	154.67	0.727	196.31	**0.911**	98.69
DEB	0.841	173.51	0.799	194.65	**0.941**	92.07
UV	0.677	1654.08	**0.740**	1281.7	0.713	1524.9
4SU	0.941	1180.69	0.956	1095.9	**0.964**	1103.87
All	0.647	1078.5	**0.66**	1086.3	**-**	-

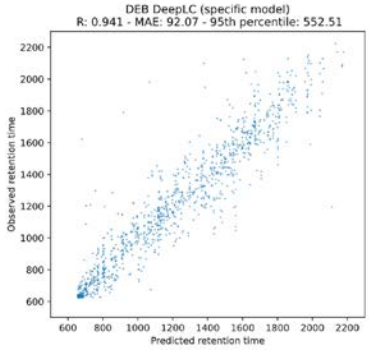

NM DeepLC (base model)
R: 0.835 - MAE: 154.67 - 95th percentile: 695.86
Observed retention time
Predicted retention time

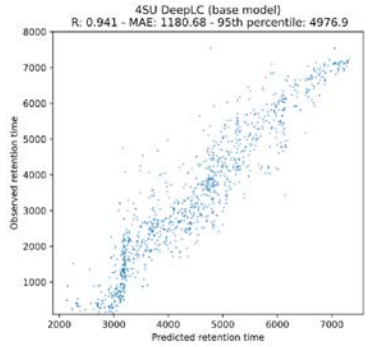

NM DeepLC (generic model)
R: 0.727 - MAE: 196.31 - 95th percentile: 888.68
Observed retention time
Predicted retention time

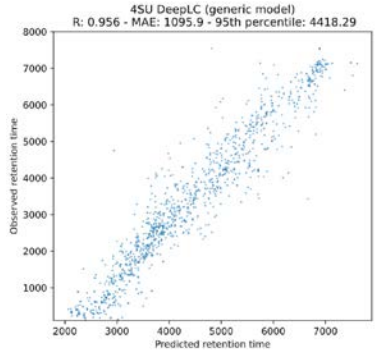

NM DeepLC (specific model)
R: 0.911 - MAE: 98.69 - 95th percentile: 487.19
Observed retention time
Predicted retention time

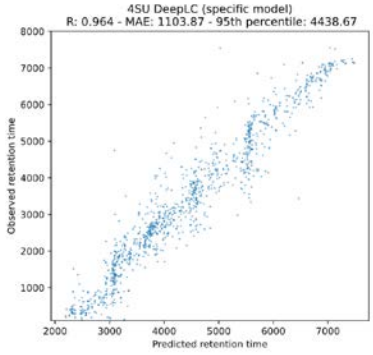

DEB DeepLC (base model)
R: 0.841 - MAE: 173.51 - 95th percentile: 888.53
Observed retention time
Predicted retention time

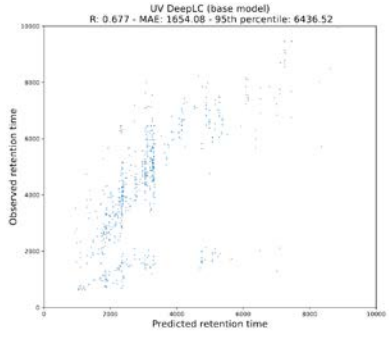

DEB DeepLC (generic model)
R: 0.799 - MAE: 194.65 - 95th percentile: 1056.32
Observed retention time
Predicted retention time

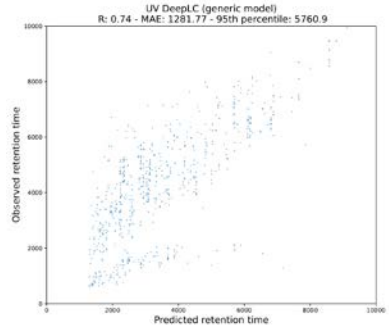

DEB DeepLC (specific model)
R: 0.941 - MAE: 92.07 - 95th percentile: 552.51
Observed retention time
Predicted retention time

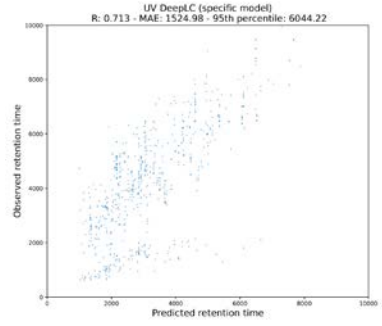

4SU DeepLC (base model)
R: 0.941 - MAE: 1180.68 - 95th percentile: 4976.9
Observed retention time
Predicted retention time


4SU DeepLC (generic model)
R: 0.956 - MAE: 1095.9 - 95th percentile: 4418.29
Observed retention time
Predicted retention time


4SU DeepLC (specific model)
R: 0.964 - MAE: 1103.87 - 95th percentile: 4438.67
Observed retention time
Predicted retention time


UV DeepLC (base model)
R: 0.677 - MAE: 1654.08 - 95th percentile: 6436.52
Observed retention time
Predicted retention time


UV DeepLC (generic model)
R: 0.74 - MAE: 1281.77 - 95th percentile: 5760.9
Observed retention time
Predicted retention time


UV DeepLC (specific model)
R: 0.713 - MAE: 1524.98 - 95th percentile: 6044.22
Observed retention time
Predicted retention time

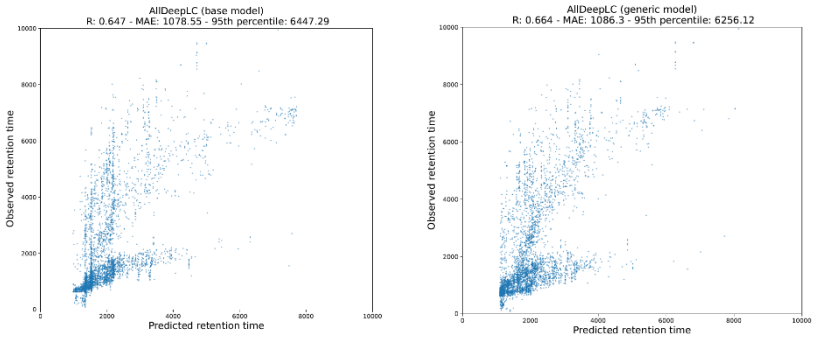
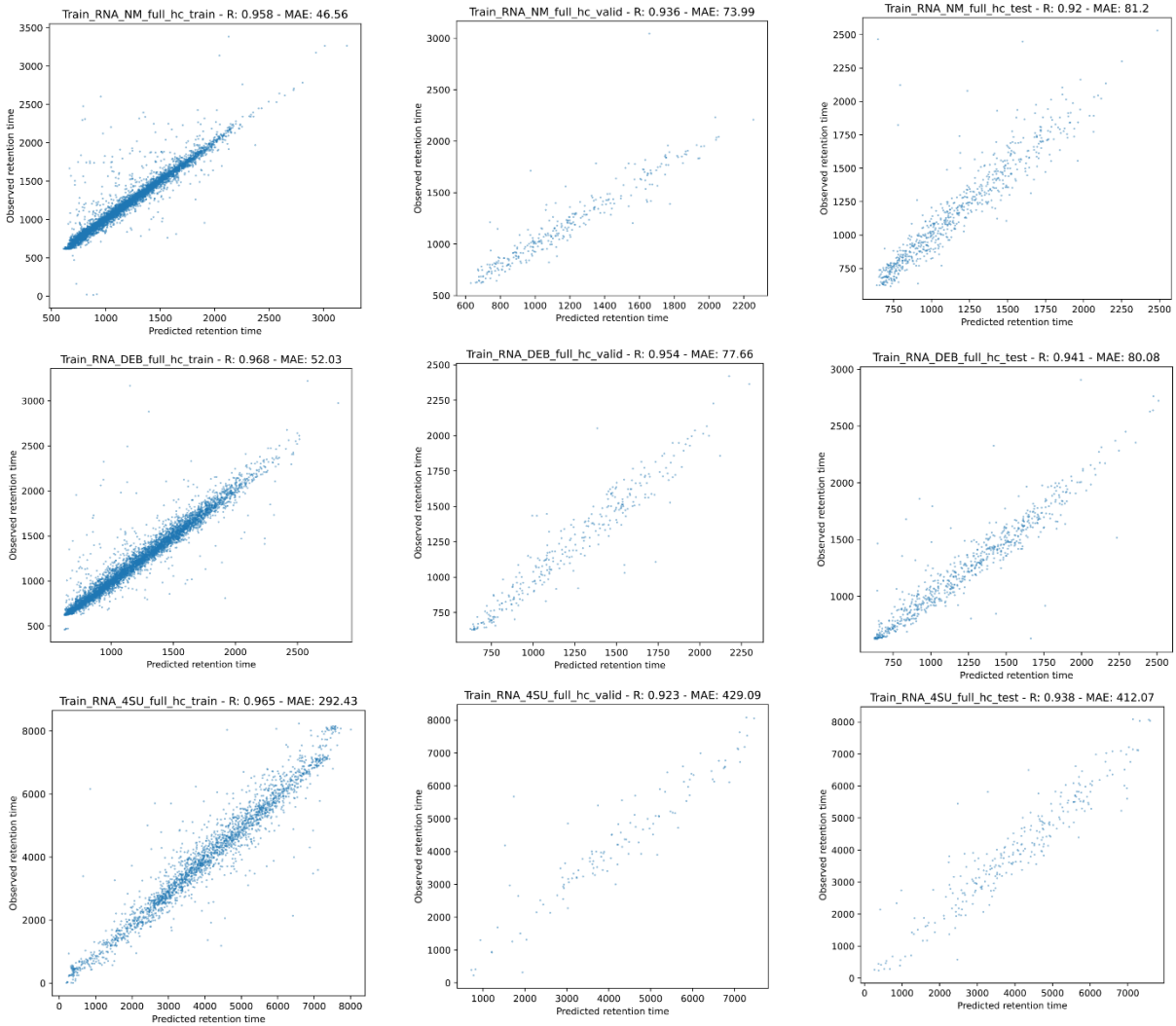


Figure S1: Scatter plots of observed and predicted retention times across four protein-RNA crosslinking protocols. (left column) base model without transfer learning; (center column) transfer learning using the generic model; (right column) transfer learning using the protocol specific models evaluated on the respective test data (see Table S4). Bottom row: (left) base model and (right) generic model evaluated on all test data set.



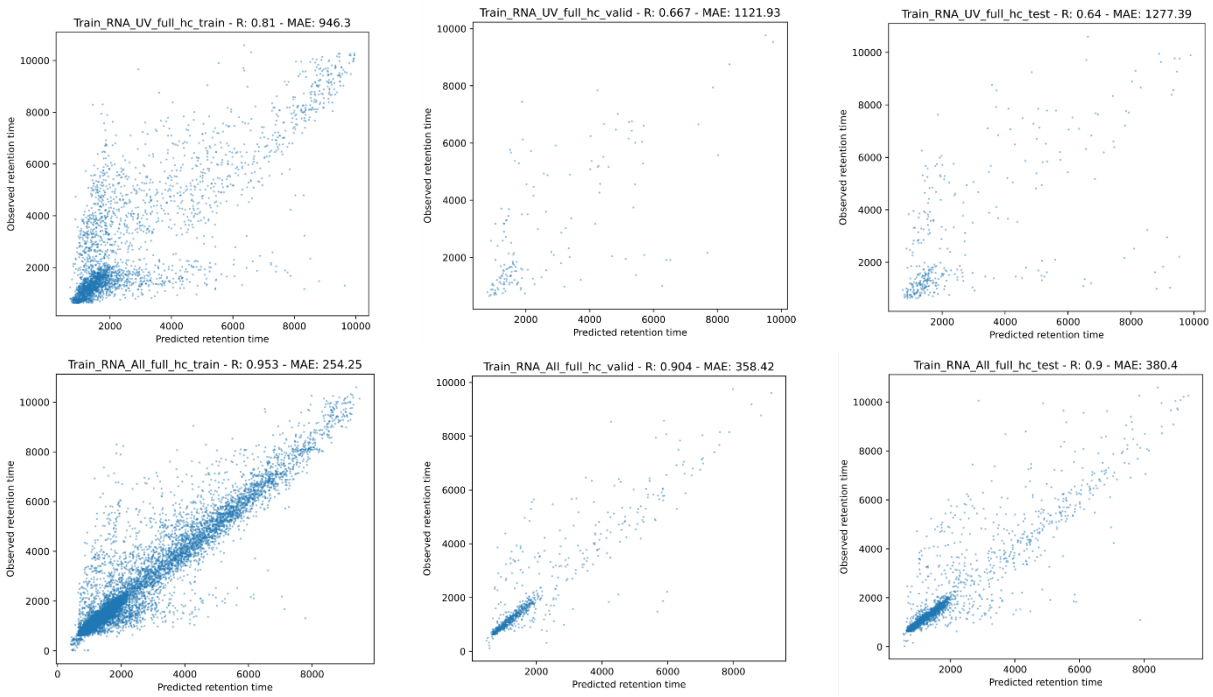
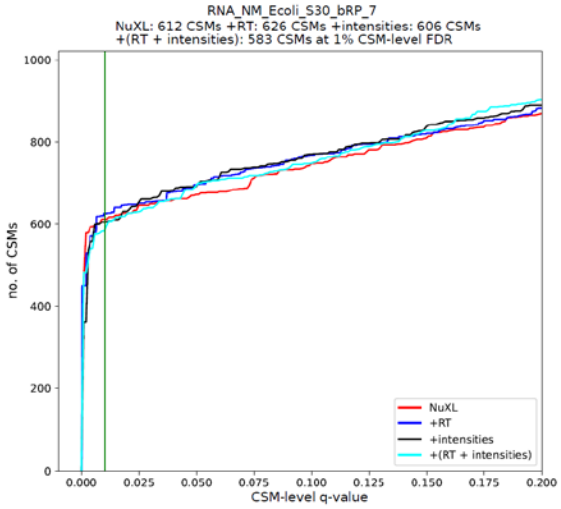
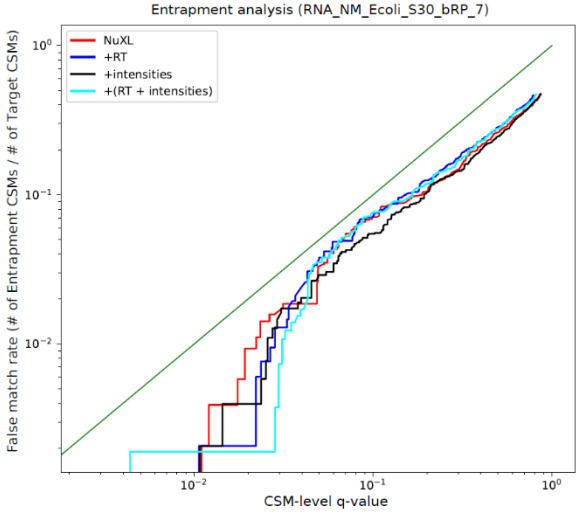
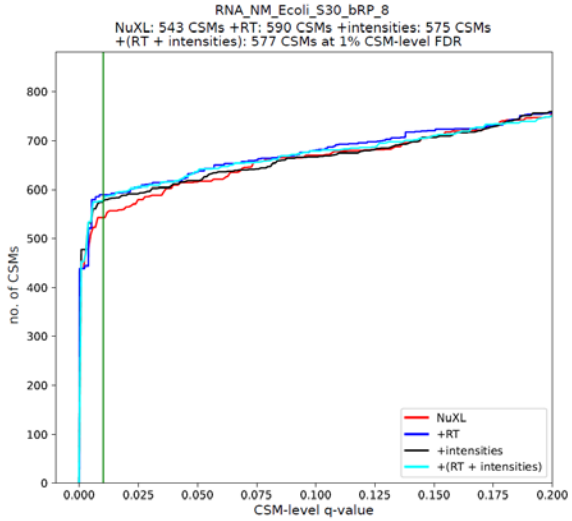
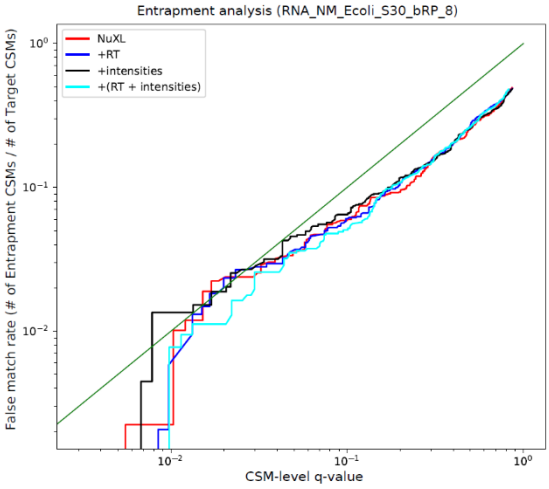


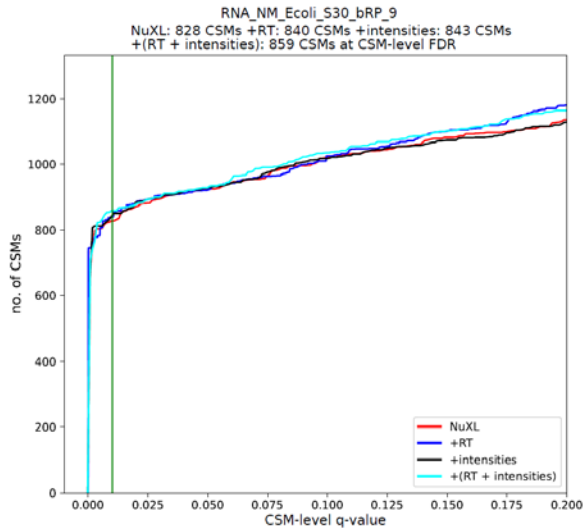
Figure S2: Model training plots of train (left column), validation (center column), and test (right column) dataset splits of model training data with deeplcretrainer.

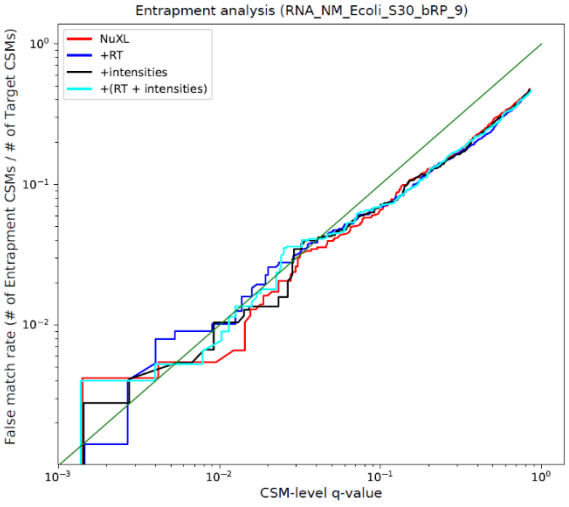

RNA_NM_Ecoli_S30_bRP_7
NuXL: 612 CSMs +RT: 626 CSMs +intensities: 606 CSMs
+(RT + intensities): 583 CSMs at 1% CSM-level FDR
no. of CSMs
CSM-level q-value
NuXL
+RT
+intensities
+(RT + intensities)

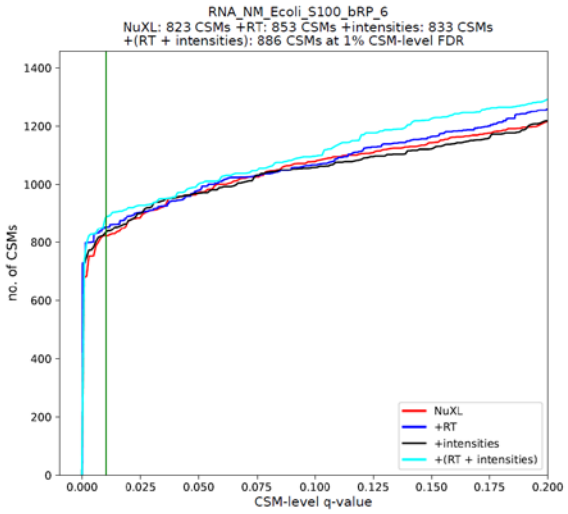

Entrapment analysis (RNA_NM_Ecoli_S30_bRP_7)
False match rate (# of Entrapment CSMs / # of Target CSMs)
CSM-level q-value
NuXL
+RT
+intensities
+(RT + intensities)

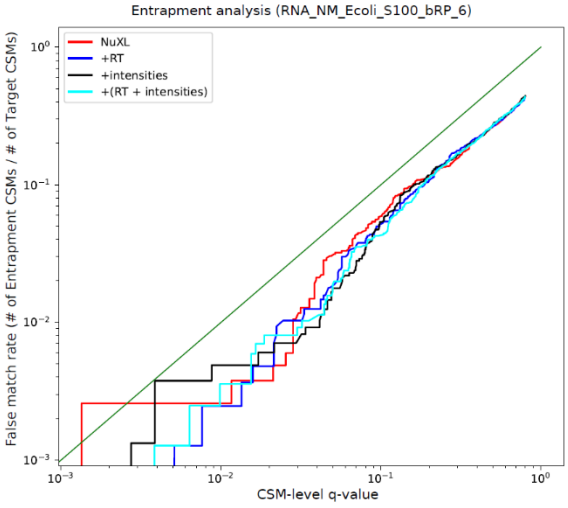

RNA_NM_Ecoli_S30_bRP_8
NuXL: 543 CSMs +RT: 590 CSMs +intensities: 575 CSMs
+(RT + intensities): 577 CSMs at 1% CSM-level FDR
no. of CSMs
CSM-level q-value
NuXL
+RT
+intensities
+(RT + intensities)

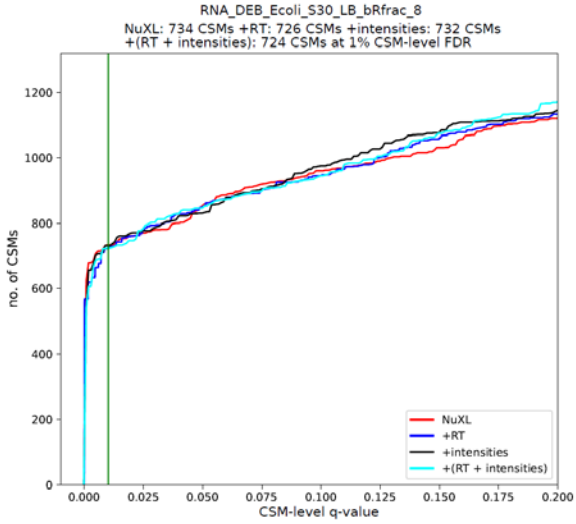

Entrapment analysis (RNA_NM_Ecoli_S30_bRP_8)
False match rate (# of Entrapment CSMs / # of Target CSMs)
CSM-level q-value
NuXL
+RT
+intensities
+(RT + intensities)

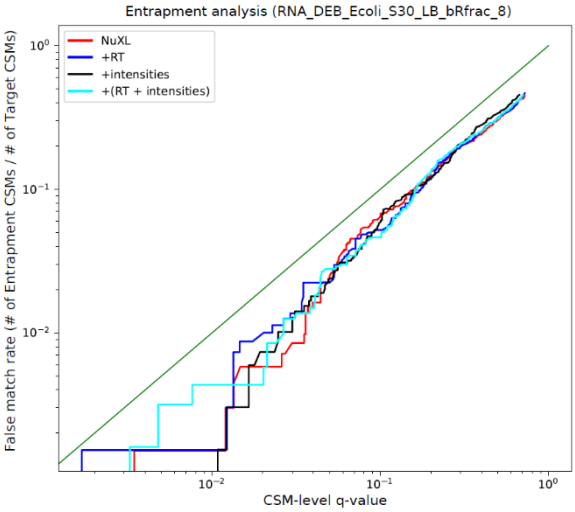

RNA_NM_Ecoli_S30_bRP_9
NuXL: 828 CSMs +RT: 840 CSMs +intensities: 843 CSMs
+(RT + intensities): 859 CSMs at CSM-level FDR
no. of CSMs
CSM-level q-value
NuXL
+RT
+intensities
+(RT + intensities)

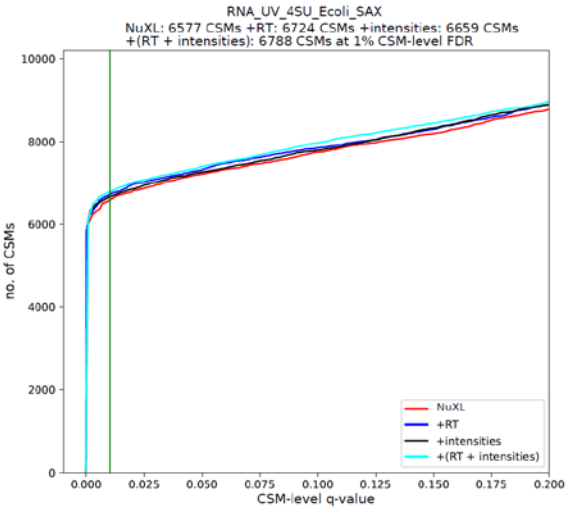

Entrapment analysis (RNA_NM_Ecoli_S30_bRP_9)
False match rate (# of Entrapment CSMs / # of Target CSMs)
CSM-level q-value
NuXL
+RT
+intensities
+(RT + intensities)

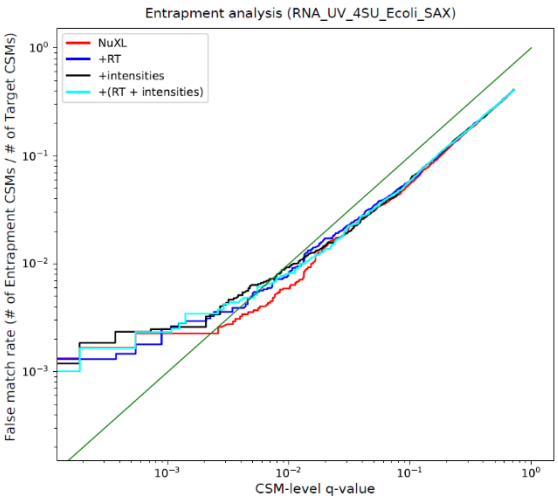

RNA_NM_Ecoli_S100_bRP_6
NuXL: 823 CSMs +RT: 853 CSMs +intensities: 833 CSMs
+(RT + intensities): 886 CSMs at 1% CSM-level FDR
no. of CSMs
CSM-level q-value
NuXL
+RT
+intensities
+(RT + intensities)

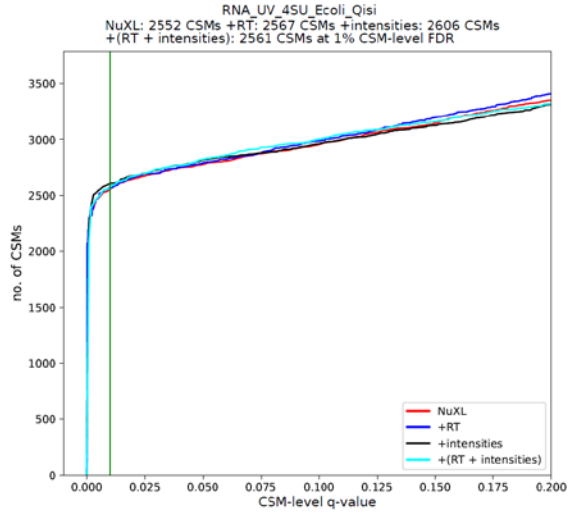

Entrapment analysis (RNA_NM_Ecoli_S100_bRP_6)
False match rate (# of Entrapment CSMs / # of Target CSMs)
CSM-level q-value
NuXL
+RT
+intensities
+(RT + intensities)

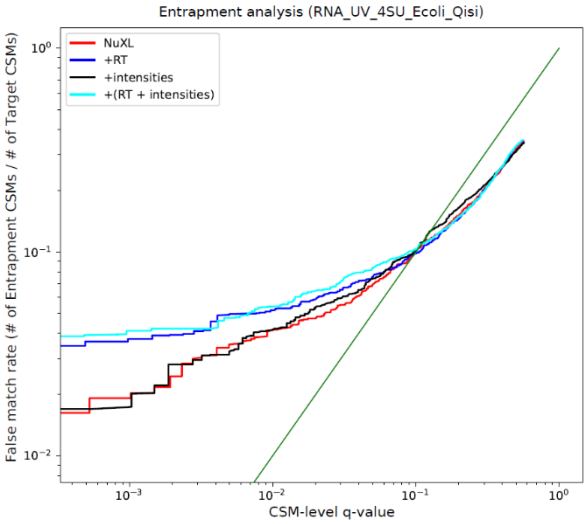

RNA_DEB_Ecoli_S30_LB_bRfrac_8
NuXL: 734 CSMs +RT: 726 CSMs +intensities: 732 CSMs
+(RT + intensities): 724 CSMs at 1% CSM-level FDR
no. of CSMs
CSM-level q-value
NuXL
+RT
+intensities
+(RT + intensities)

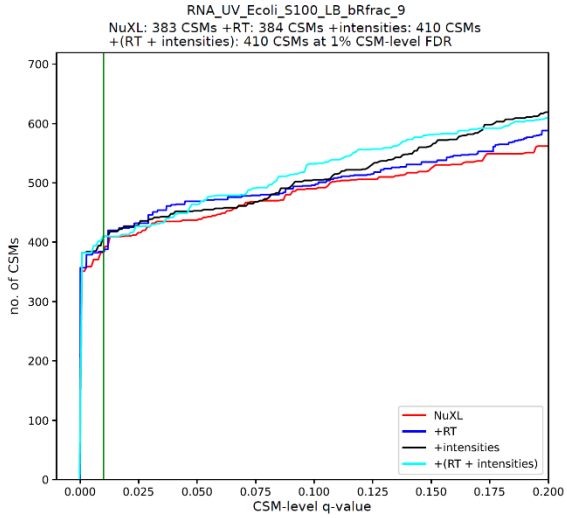

Entrapment analysis (RNA_DEB_Ecoli_S30_LB_bRfrac_8)
False match rate (# of Entrapment CSMs / # of Target CSMs)
CSM-level q-value
NuXL
+RT
+intensities
+(RT + intensities)

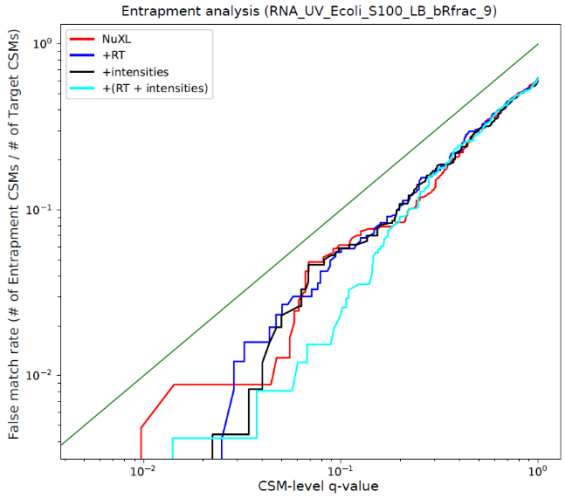

RNA_UV_4SU_Ecoli_SAX
NuXL: 6577 CSMs +RT: 6724 CSMs +intensities: 6659 CSMs
+(RT + intensities): 6788 CSMs at 1% CSM-level FDR
no. of CSMs
CSM-level q-value
NuXL
+RT
+intensities
+(RT + intensities)

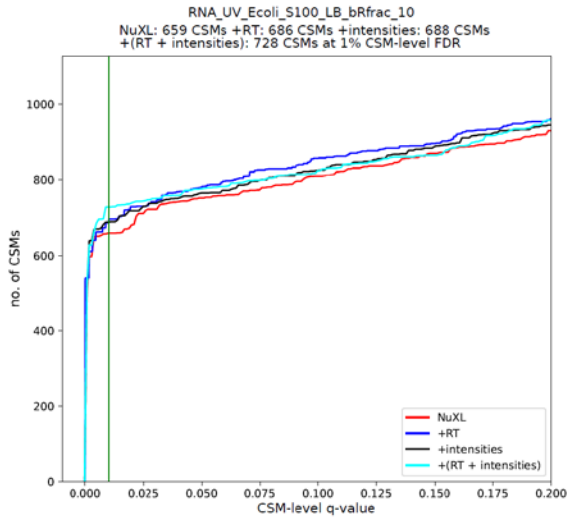

Entrapment analysis (RNA_UV_4SU_Ecoli_SAX)
False match rate (# of Entrapment CSMs / # of Target CSMs)
CSM-level q-value
NuXL
+RT
+intensities
+(RT + intensities)

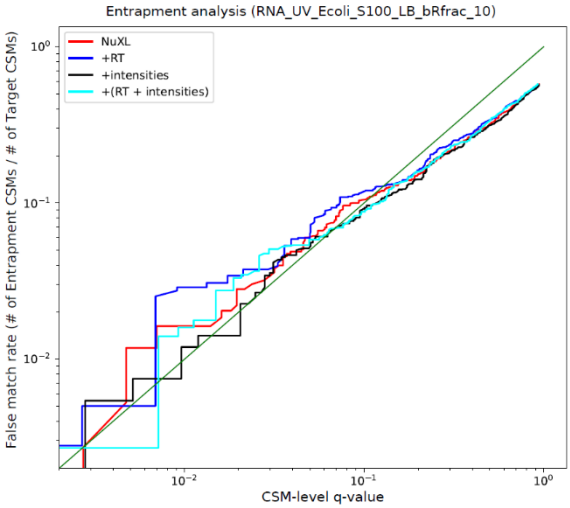

RNA_UV_4SU_Ecoli_Qisi
NuXL: 2552 CSMs +RT: 2567 CSMs +intensities: 2606 CSMs
+(RT + intensities): 2561 CSMs at 1% CSM-level FDR
no. of CSMs
CSM-level q-value
NuXL
+RT
+intensities
+(RT + intensities)


Entrapment analysis (RNA_UV_4SU_Ecoli_Qisi)
False match rate (# of Entrapment CSMs / # of Target CSMs)
CSM-level q-value
NuXL
+RT
+intensities
+(RT + intensities)


RNA_UV_Ecoli_S100_LB_bRfrac_9
NuXL: 383 CSMs +RT: 384 CSMs +intensities: 410 CSMs
+(RT + intensities): 410 CSMs at 1% CSM-level FDR
no. of CSMs
CSM-level q-value
NuXL
+RT
+intensities
+(RT + intensities)


Entrapment analysis (RNA_UV_Ecoli_S100_LB_bRfrac_9)
False match rate (# of Entrapment CSMs / # of Target CSMs)
CSM-level q-value
NuXL
+RT
+intensities
+(RT + intensities)


RNA_UV_Ecoli_S100_LB_bRfrac_10
NuXL: 659 CSMs +RT: 686 CSMs +intensities: 688 CSMs
+(RT + intensities): 728 CSMs at 1% CSM-level FDR
no. of CSMs
CSM-level q-value
NuXL
+RT
+intensities
+(RT + intensities)


Entrapment analysis (RNA_UV_Ecoli_S100_LB_bRfrac_10)
False match rate (# of Entrapment CSMs / # of Target CSMs)
CSM-level q-value
NuXL
+RT
+intensities
+(RT + intensities)

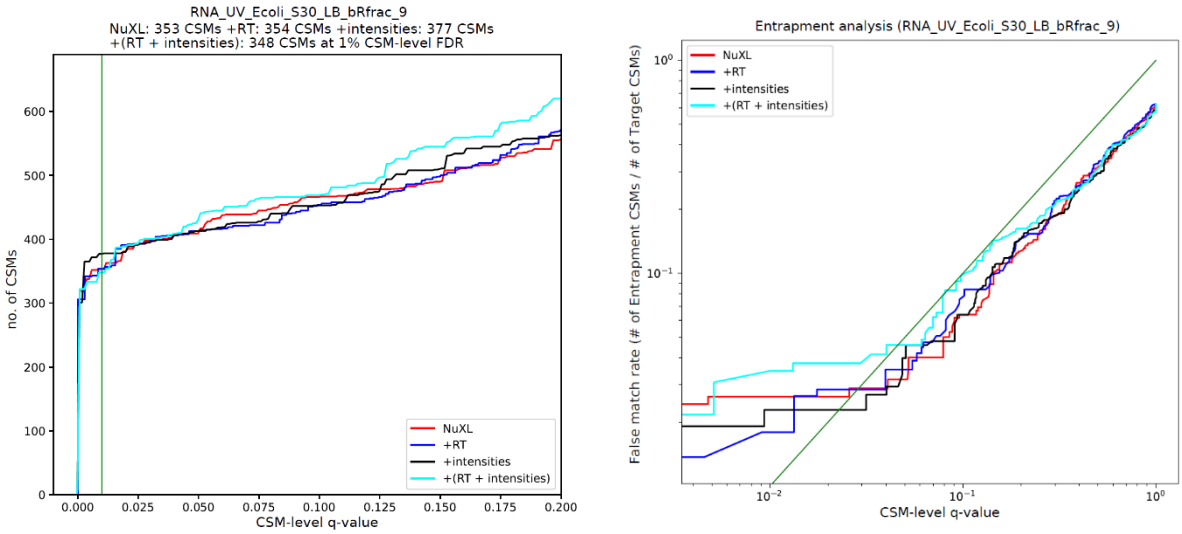


Figure S3: Rescoring performance for the four different protocols (1% CSM-level FDR) when retention time features (+RT), intensity correlation (+intensities), and all (+(RT + intensities)) were used; **left:** pseudo-ROC curves for the different set of features used in rescoring. The y-axis represents the number of Cross-link Spectrum Matches (CSMs) identified at different CSM-level q-value thresholds. The green vertical line indicates the commonly used 0.01 CSM-level q-value threshold. **right:** Comparison between False Match Rate (FMR) q-value (y-axis) and CSM-level q-value (x-axis) for CSM-level entrapment analysis (diagonal reference line indicates matching FMR and empirical CSM-level q-value).

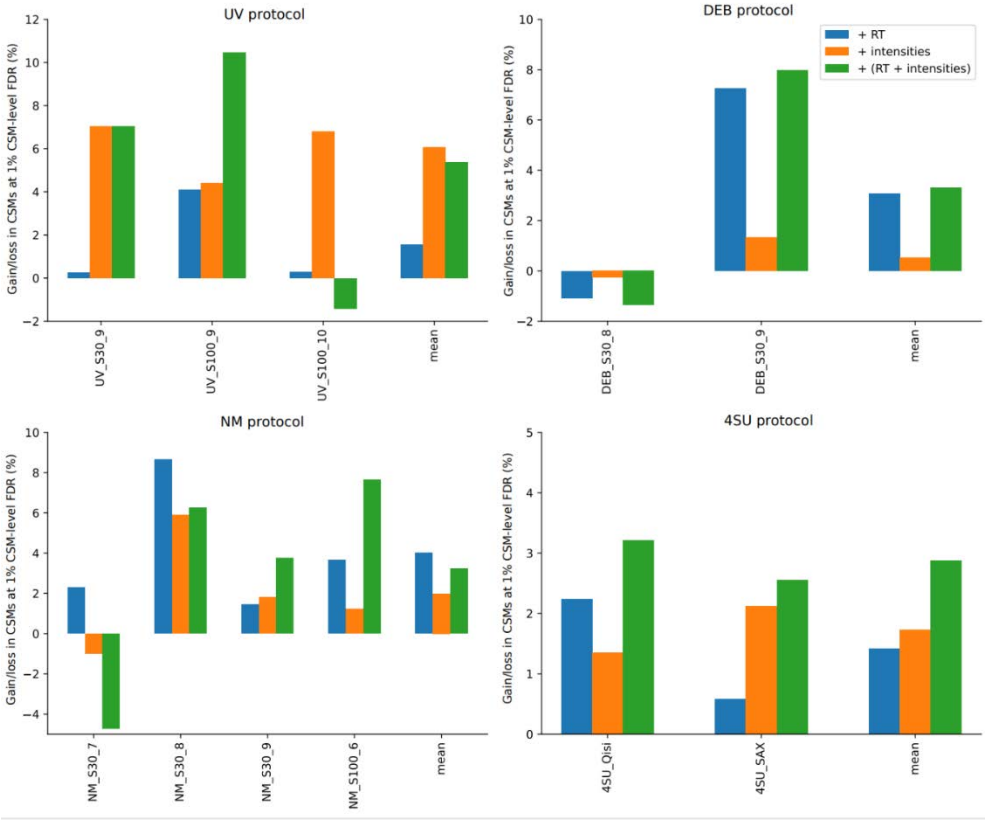
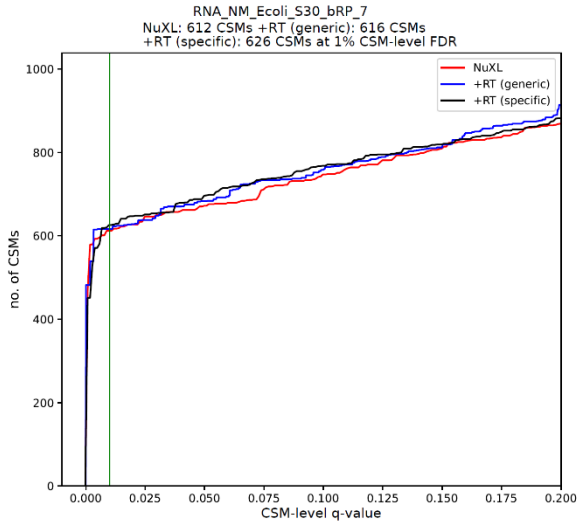
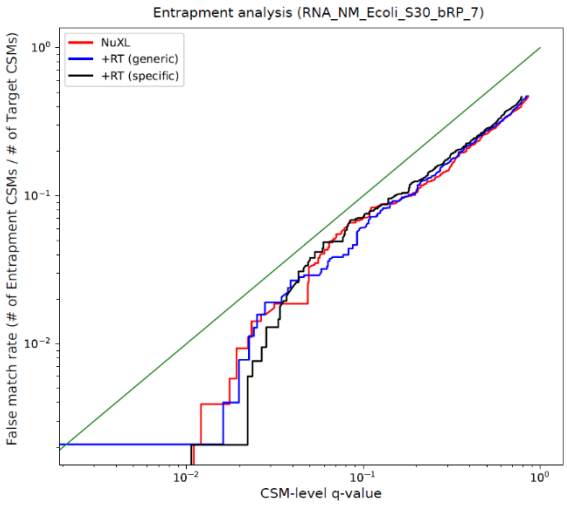
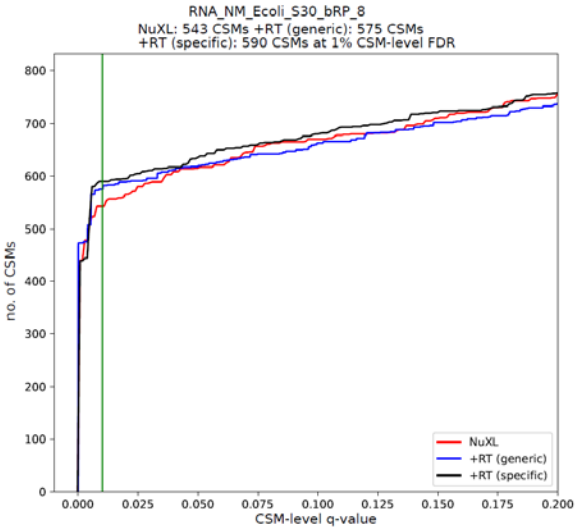
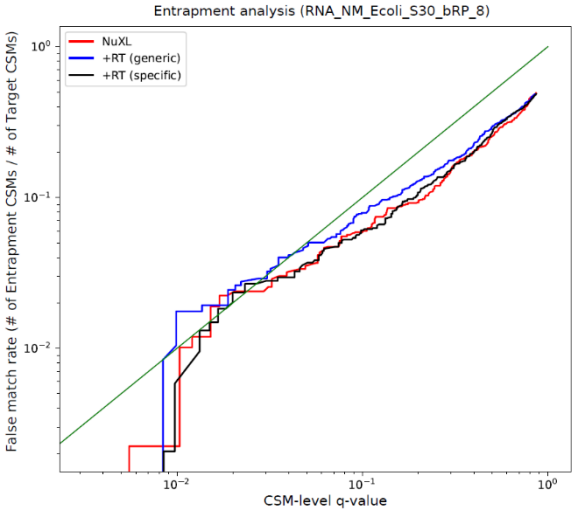


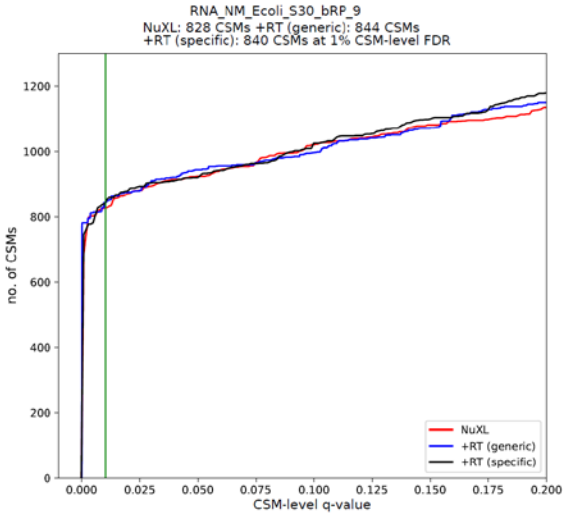
Figure S4: Detail gain/loss plot of rescoring performance of analyze data for the four different protocols (1% CSM-level FDR) if retention time predictions (RT), intensity predictions (Intensities), or both (RT + Intensities) was performed and added as features for percolator rescoring. Results shown of individual runs and mean results across all protocols.

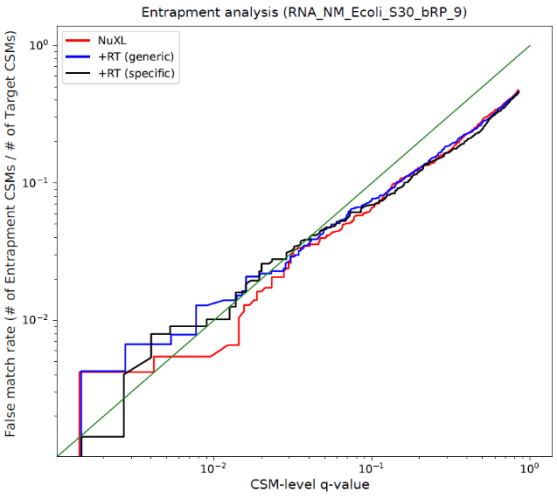

RNA_NM_Ecoli_S30_bRP_7
NuXL: 612 CSMs +RT (generic): 616 CSMs
+RT (specific): 626 CSMs at 1% CSM-level FDR
NuXL
+RT (generic)
+RT (specific)
no. of CSMs
CSM-level q-value

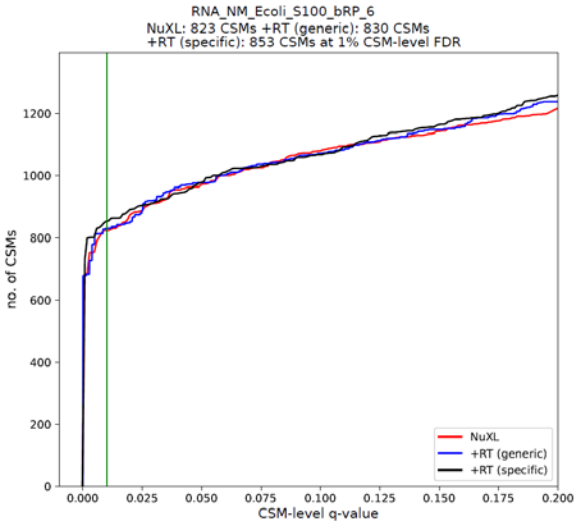

Entrapment analysis (RNA_NM_Ecoli_S30_bRP_7)
NuXL
+RT (generic)
+RT (specific)
False match rate (# of Entrapment CSMs / # of Target CSMs)
CSM-level q-value

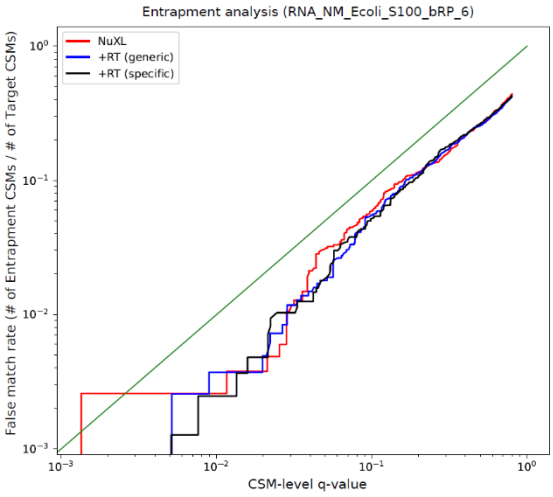

RNA_NM_Ecoli_S30_bRP_8
NuXL: 543 CSMs +RT (generic): 575 CSMs
+RT (specific): 590 CSMs at 1% CSM-level FDR
NuXL
+RT (generic)
+RT (specific)
no. of CSMs
CSM-level q-value

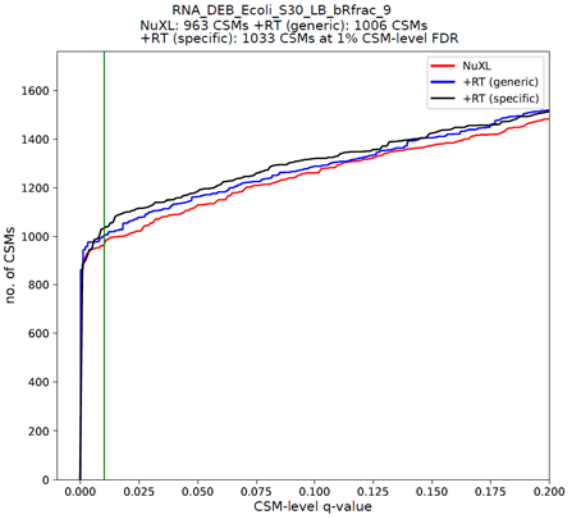

Entrapment analysis (RNA_NM_Ecoli_S30_bRP_8)
NuXL
+RT (generic)
+RT (specific)
False match rate (# of Entrapment CSMs / # of Target CSMs)
CSM-level q-value

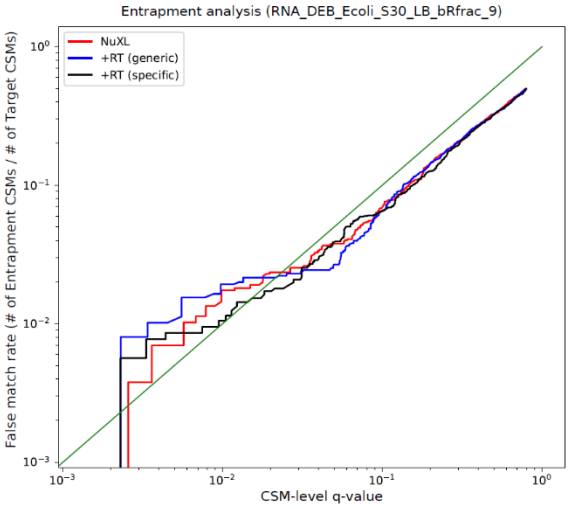

RNA_NM_Ecoli_S30_bRP_9
NuXL: 828 CSMs +RT (generic): 844 CSMs
+RT (specific): 840 CSMs at 1% CSM-level FDR
NuXL
+RT (generic)
+RT (specific)
no. of CSMs
CSM-level q-value

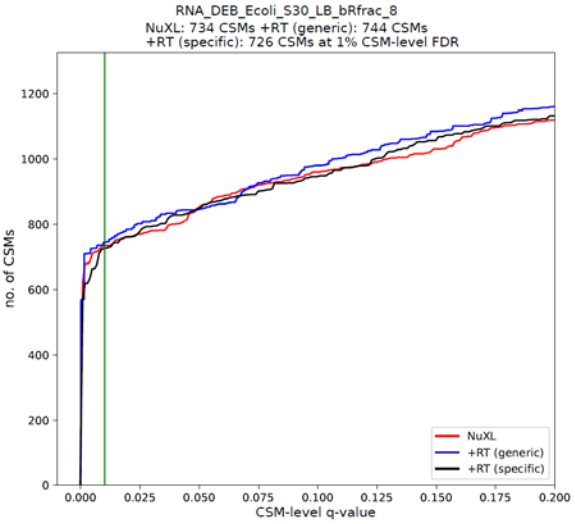

Entrapment analysis (RNA_NM_Ecoli_S30_bRP_9)
NuXL
+RT (generic)
+RT (specific)
False match rate (# of Entrapment CSMs / # of Target CSMs)
CSM-level q-value

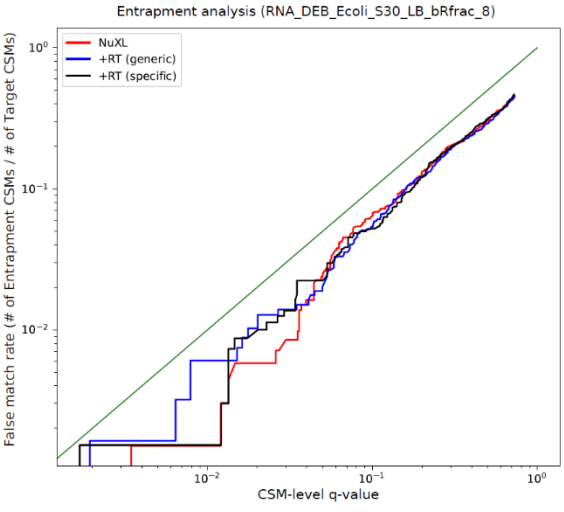

RNA_NM_Ecoli_S100_bRP_6
NuXL: 823 CSMs +RT (generic): 830 CSMs
+RT (specific): 853 CSMs at 1% CSM-level FDR
no. of CSMs
CSM-level q-value
NuXL
+RT (generic)
+RT (specific)

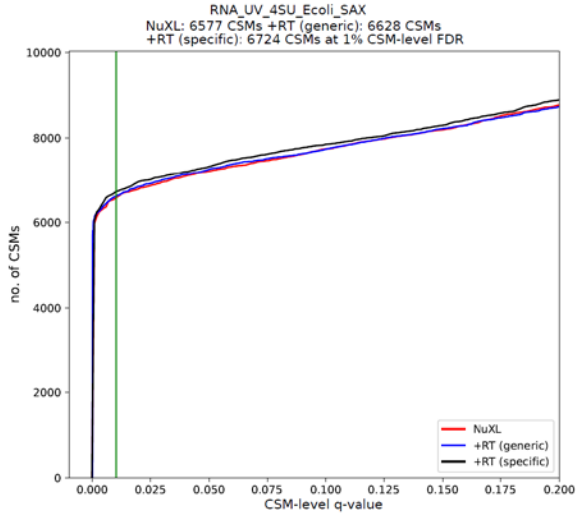

Entrapment analysis (RNA_NM_Ecoli_S100_bRP_6)
False match rate (# of Entrapment CSMs / # of Target CSMs)
CSM-level q-value
NuXL
+RT (generic)
+RT (specific)

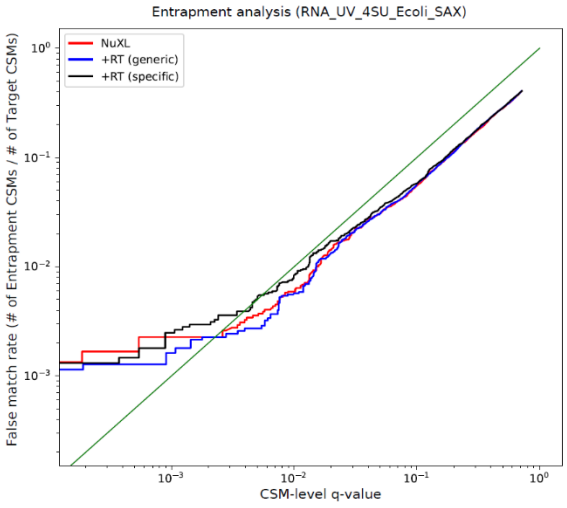

RNA_DEB_Ecoli_S30_LB_bRfrac_9
NuXL: 963 CSMs +RT (generic): 1006 CSMs
+RT (specific): 1033 CSMs at 1% CSM-level FDR
no. of CSMs
CSM-level q-value
NuXL
+RT (generic)
+RT (specific)

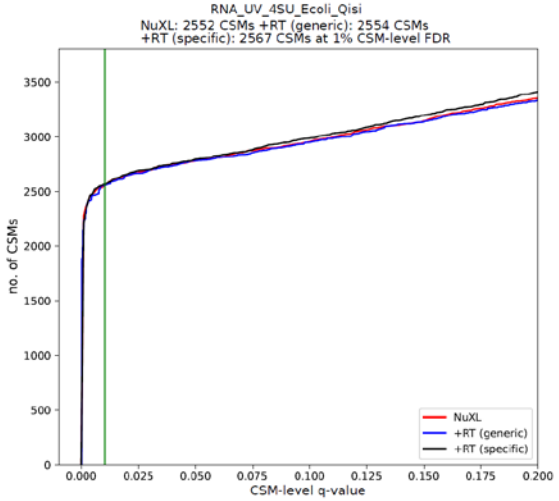

Entrapment analysis (RNA_DEB_Ecoli_S30_LB_bRfrac_9)
False match rate (# of Entrapment CSMs / # of Target CSMs)
CSM-level q-value
NuXL
+RT (generic)
+RT (specific)

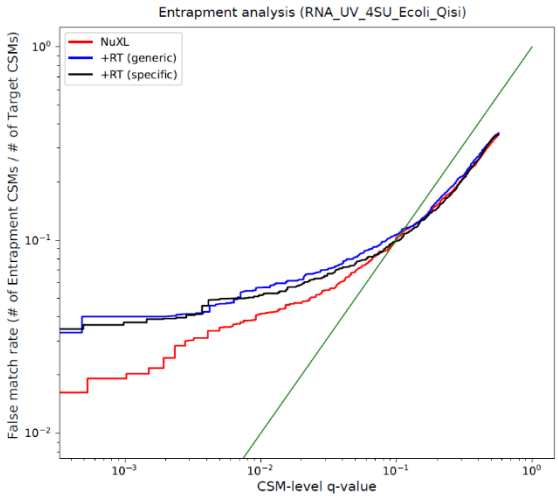

RNA_DEB_Ecoli_S30_LB_bRfrac_8
NuXL: 734 CSMs +RT (generic): 744 CSMs
+RT (specific): 726 CSMs at 1% CSM-level FDR
no. of CSMs
CSM-level q-value
NuXL
+RT (generic)
+RT (specific)

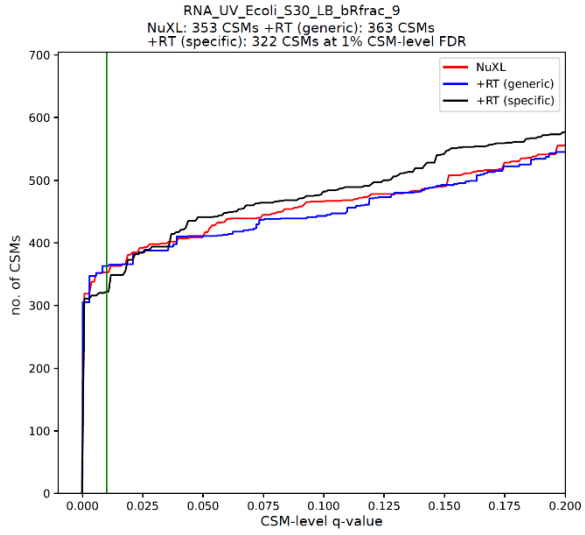

Entrapment analysis (RNA_DEB_Ecoli_S30_LB_bRfrac_8)
False match rate (# of Entrapment CSMs / # of Target CSMs)
CSM-level q-value
NuXL
+RT (generic)
+RT (specific)

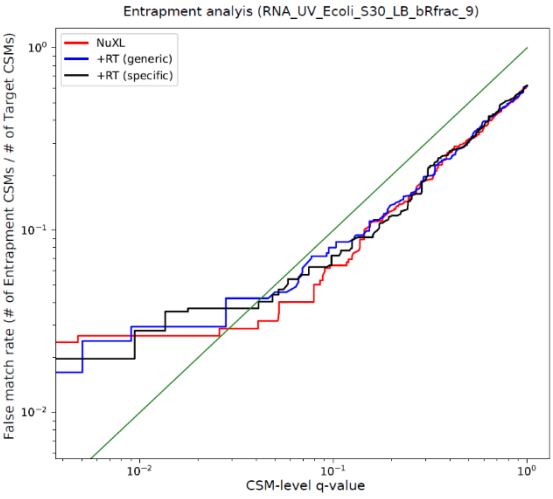

RNA_UV_4SU_Ecoli_SAX
NuXL: 6577 CSMs +RT (generic): 6628 CSMs
+RT (specific): 6724 CSMs at 1% CSM-level FDR
no. of CSMs
CSM-level q-value
NuXL
+RT (generic)
+RT (specific)


Entrapment analysis (RNA_UV_4SU_Ecoli_SAX)
False match rate (# of Entrapment CSMs / # of Target CSMs)
CSM-level q-value
NuXL
+RT (generic)
+RT (specific)


RNA_UV_4SU_Ecoli_Qisi
NuXL: 2552 CSMs +RT (generic): 2554 CSMs
+RT (specific): 2567 CSMs at 1% CSM-level FDR
no. of CSMs
CSM-level q-value
NuXL
+RT (generic)
+RT (specific)


Entrapment analysis (RNA_UV_4SU_Ecoli_Qisi)
False match rate (# of Entrapment CSMs / # of Target CSMs)
CSM-level q-value
NuXL
+RT (generic)
+RT (specific)


RNA_UV_Ecoli_S30_LB_bRfrac_9
NuXL: 353 CSMs +RT (generic): 363 CSMs
+RT (specific): 322 CSMs at 1% CSM-level FDR
no. of CSMs
CSM-level q-value
NuXL
+RT (generic)
+RT (specific)


Entrapment analyis (RNA_UV_Ecoli_S30_LB_bRfrac_9)
False match rate (# of Entrapment CSMs / # of Target CSMs)
CSM-level q-value
NuXL
+RT (generic)
+RT (specific)

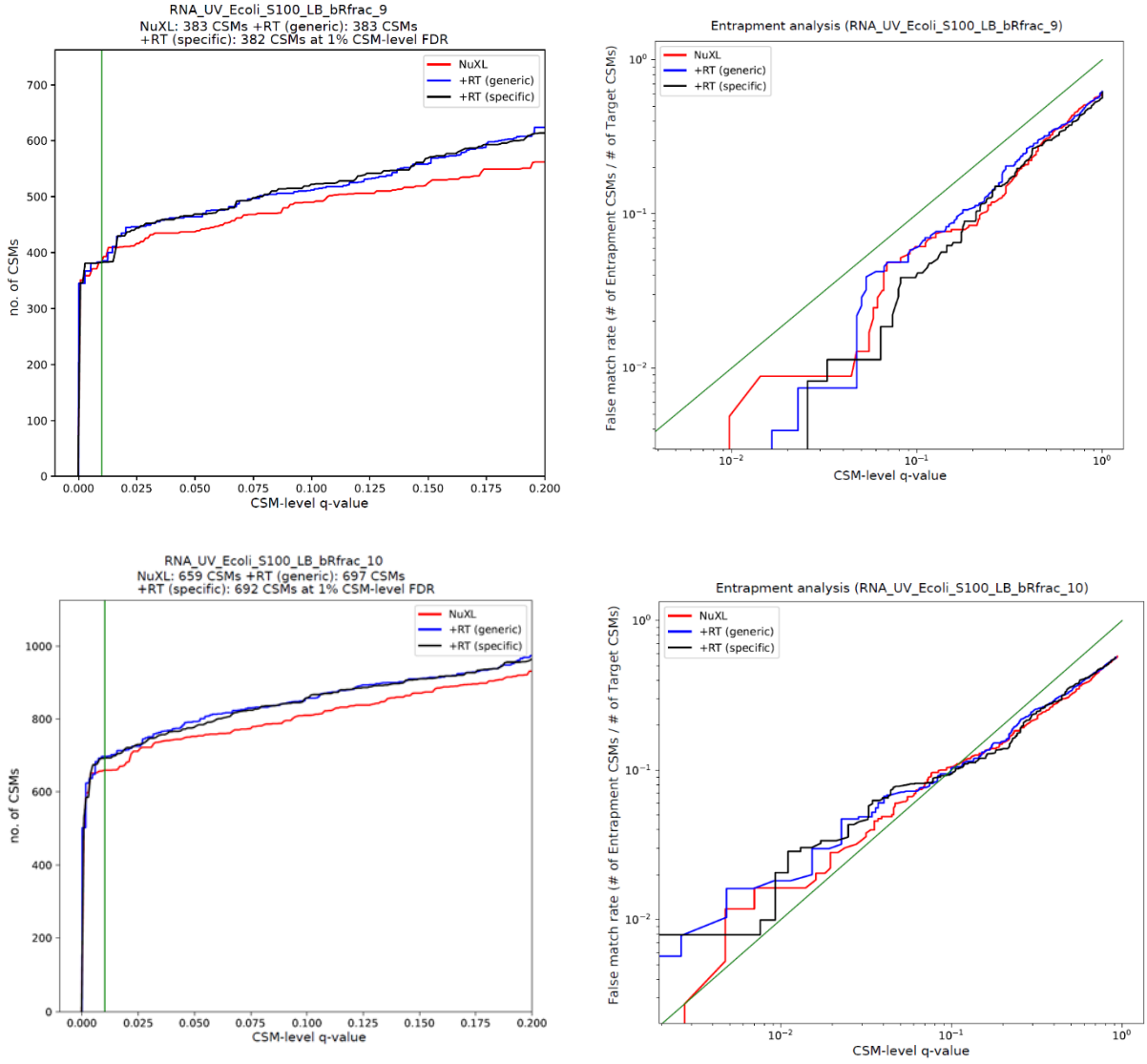


Figure S5: Rescoring performance for the four different protocols (1% CSM-level FDR) when specific protocol models retention time predictions (+RT(specific)), and generic model retention time predictions (+RT(generic)) were used; **left:** pseudo-ROC curves for the different set of features used in rescoring. The y-axis represents the number of Cross-link Spectrum Matches (CSMs) identified at different CSM-level q-value thresholds. The green vertical line indicates the commonly used 0.01 CSM-level q-value threshold. **right:** Comparison between False Match Rate (FMR) q-value (y-axis) and CSM-level q-value (x-axis) for CSM-level entrapment analysis (diagonal reference line indicates matching FMR and empirical CSM-level q-value).

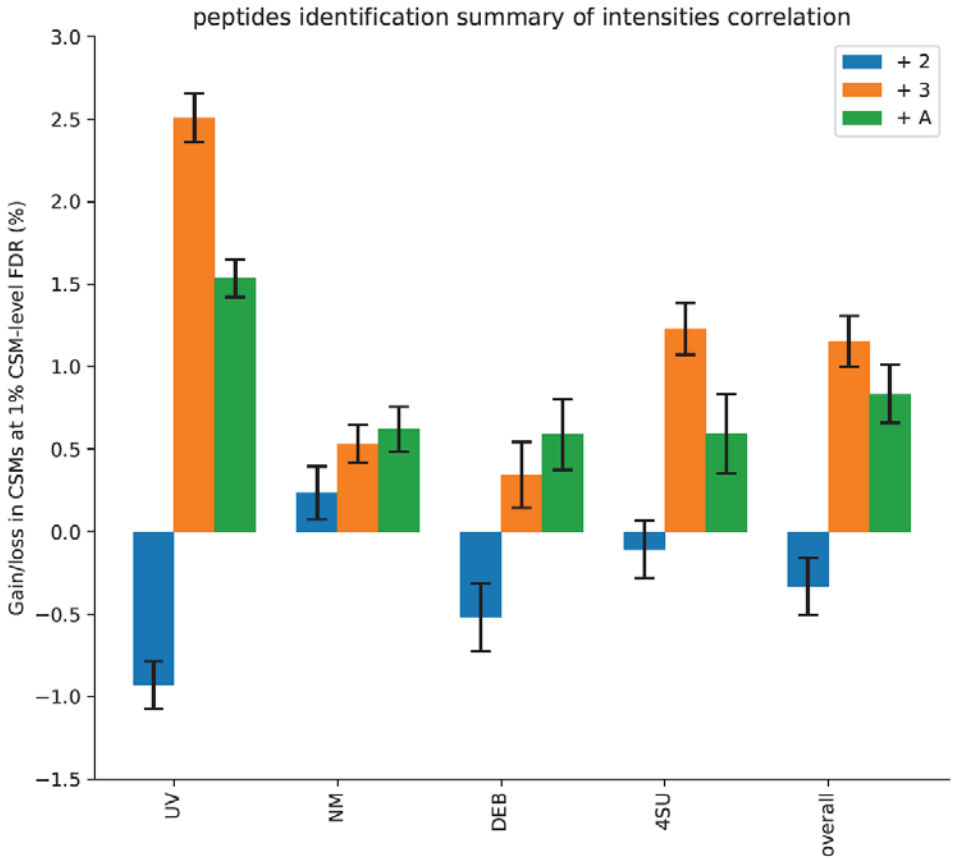


Figure S6: Rescoring performance of the four distinct protocols (1% CSM-level FDR) if the first two (+2), three (+3) and all (+A) prefix/suffix peak ion correlations were used in percolator rescoring. The error bar shows standard deviations for each protocol evaluated at 1% CSM-level FDR, analyzed using 10 random percolator seeds (Figure S4 shows all files for each protocol).

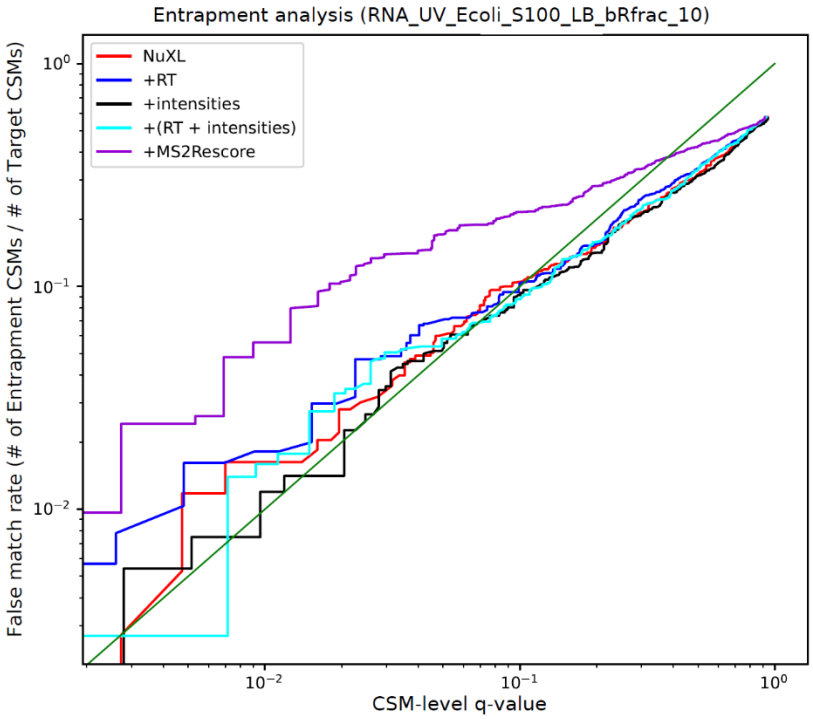


Figure S7: CSM-level entrapment plot for adding additional retention time features (+RT), intensity features (+Intensities), and both (+RT +intensities) compared to adaptation of MS2Rescore features (+MS2Rescore). The diagonal reference line indicates matching False Match Rate (FMR) and empirical CSM-level q-value.