User's Manual of IMP-X-FDR

Contents

Aim

The main aim of IMP-X-FDR is to provide an easy to use and automated way for calculation of the real false discovery rate (FDR) of crosslink search results (on sequence match or unique residue pair level) obtained from any peptide-library system. This makes it a valuable tool for testing novel crosslinkers, to compare acquisition strategies or to benchmark different (novel) crosslink search algorithms.

Additional functionalities are the comparison of crosslink IDs found from different searches (also from different search algorithms) using Venn diagrams as well as to investigate intrinsic properties (mass, pI value, amino-acid sequences analyses, …) of the crosslinks dependent e.g., on retention time or acquisition strategy. The tool thereby allows a comparison of the physicochemical properties of identified crosslinks vs those theoretically formed within the peptide library used.

1. Installation and quick guide

In order to run the application, you need the following software installed:

- .NET 6 Desktop Runtime
- Python 3.9 (or higher)
    - Missing packages will be installed automatically

Installation

Download the latest release, unpack the archive, and run IMP_X_FDR.exe

In case of missing supporting programs -like Python- on your system, you will be automatically redirected to the respective installation website. Install the suitable version for your computer.

Quick guide

After opening IMP-X-FDR exe a GUI as shown in Figure 1 allows to load (filtered) crosslink sequence match (CSM) or unique crosslink (XL) files as input. Set a corresponding library file that fits to the used dataset. Pre-set libraries can be found under resources/libraries in the installation folder. If CSM containing files are used as input, they can be grouped to XLs by IMP-X-FDR by ticking “Group crosslink spectrum matches to unique crosslinks”. In case XL lists are used as input, this tick-box can be ignored and will have no effect.

Please refer to the following sections for more details on how to use IMP-X-FDR.

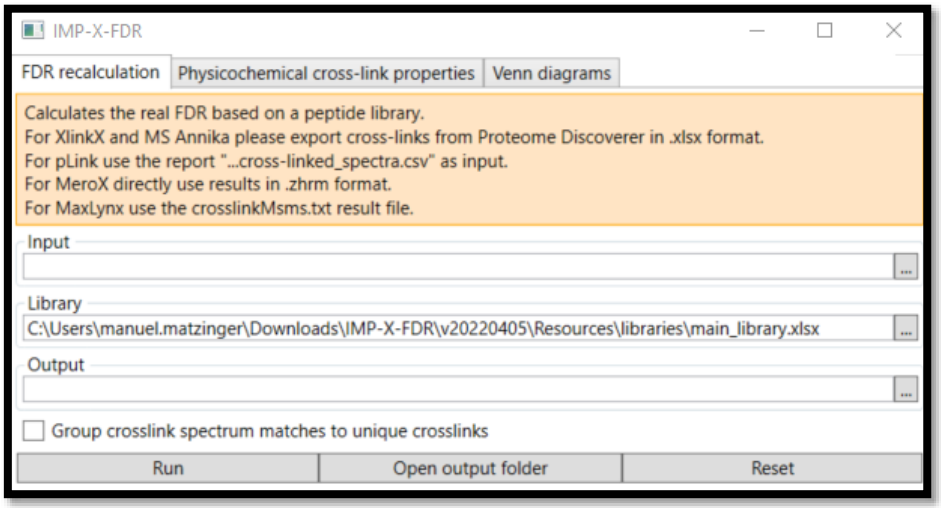

Figure 1: GUI of IMP-X-FDR

2. FDR-recalculation

2.1. Input files

2.1.1. Annika and XlinkX Input

If you have chosen “MS Annika” or “XlinkX”, the input file you need for the “FDR-recalculation” can be generated from Proteome Discoverer by exporting the (filtered) crosslink or CSM list to Microsoft Excel.

2.1.2. Merox Input

As input file a “.zhrm” result file will be used. The software will temporarily convert it in a “.zip” data, extract the necessary information and then bring the file to its initial form. The script was created for the version 2.0.1.4 of MeroX and it is recommended that input files come from this version of the software, otherwise the results may be incomplete or not analysable.

This script an also be run from the command prompt, where you can give two optional arguments:

```
optional arguments:
  -h, --help            show this help message and exit
  -sintra SCORE_INTRALINK, --score_intralink SCORE_INTRALINK
                        introduce the intraprotein XL cut-off score for the selected FDR.
 It can be found in -Show decoy analysis-
  -sinter SCORE_INTERLINK, --score_interlink SCORE_INTERLINK
                        introduce the interprotein XL cut-off score for the selected FDR.
 It can be found in -Show decoy analysis-
```

Figure 2: Optional arguments, MeroX – refer to Decoy analysis in MeroX to see scores for specific FDR (Figure 3)

For example, if the number of the crosslinks delivered by the IMP-X-FDR doesn’t match the number shown by the software (if it happens, it normally differs by 1 or 2 crosslinks), this is because MeroX performs an extra filtering. The extra filtering can be also performed by “merox_master_score.py” script when used in via command prompt and adding the optional arguments.

The advantage of this feature is that one can run multiple FDR-Check searches by just changing the intra- and interscore cut-off value without having to perform multiple MeroX searches by changing the FDR value.

The image below shows, where to find the cut-off values of a MeroX search with a defined FDR of 1%.

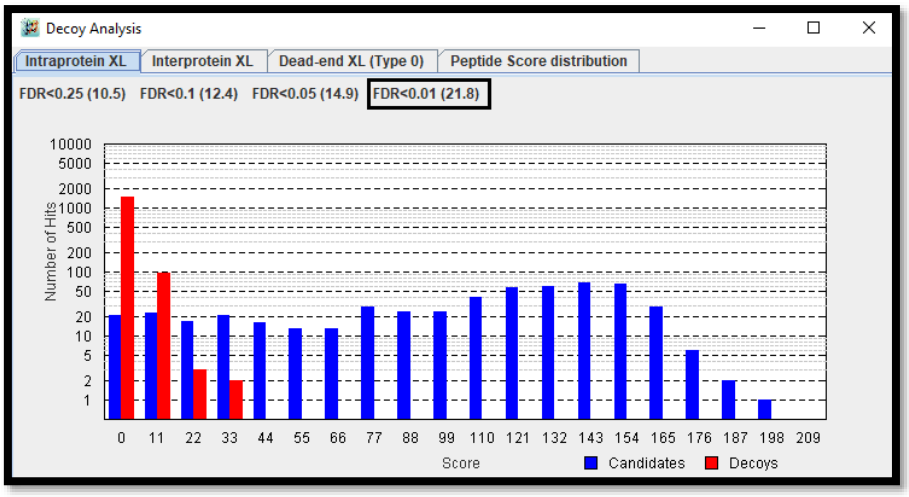

Figure 3: Example, scores for specific FDR, from MeroX (Iacobucci, C. et al. Nat. Protoc., 2018)

2.1.3. pLink input

The needed file of pLink 2 searches for the FDR-Check script is found in the pLink 2 output subdirectory "reports" with the default extension "**filtered_crosslinked_spectra.csv**".

2.1.4 MaxLynx input

For MaxLynx/MaxQuant searches use the "crosslinkMsms.txt" output as input for IMP-X-FDR.

2.1.5 XiSearch input

For xiSearch results export a crosslink table as csv from xiFDR to be used as input for IMP-X-FDR.

2.2. Support libraries

The support library files for the published datasets based on ribosomal peptides (peptide library 1-3) as well as our previously published Cas9 data (Beveridge, R. et al., Nat Commun, 2020) are provided within the resources/libraries folder. They are needed to support IMP-X-FDR in grouping crosslink IDs to correct (within the same peptide group) and incorrect crosslinks (connected peptides within different peptide groups).

In case another peptide library system will be utilized, the structure and the extension of the document (.xlsx) must be kept consistent, as indicated below.

	A	B
1	Group1	x
2	VALVAKIGENINIR	5
3	EIAEKMVEGR	4
4	KAGNVAADGVIK	0
5	EFIAKLQANPAK	4
6	MAALMKQR	5
7	ILKCGFR	2
8	VKDLPGVR	1
9	MAHIEKQAGELQEK	5
10	QAGELQEKLIAVNR	7
11	EVPAAIQKAMEK	7
12	Group2	x
13	VKGGFTVELNGIR	1
14	ITDVEVLKAQFEEER	7
15	TGAGMMDCKK	8
16	KAGFVTR	0
17	HKATLLGLGLR	1
18	MAKTIK	2
19	VKHPSEIVNVGDEITVK	1
20	ITLNMGVGEAIADKK	13
21	QCKANPWQQFAETHNK	2
22	ANPWQQFAETHNKGDR	12

Figure 4: Structure of library support files

The sequences contained in the first group are followed directly (without any empty cells) by the next groups.

The numbers in column B are only mandatory if a physiochemical investigation is performed and indicate the position of the crosslinker-binding amino-acid in the peptide sequence minus one. To define the respective groups a ‘x’ is listed in column B.

2.3. Output files

1. "**[name]_output.csv**"
   The main output data is a csv document, which returns a detailed list of the crosslinks containing information about their sequences, theoretical origin-proteins, position in protein, their score and if the crosslinks are formed within the same group (considered as true) or not (considered as false).

Sequence A	Sequence B	Accession A	Accession B	Position in protein A	Position in protein B	Score crosslink	Within same group
KSSAAR	TSGEKHLR	P0A7X3	P0A7N4	13	37	189.06	TRUE
EKLQER	KQQGHR	P0A6F5	P0AG48	364	85	184.14	TRUE
KDIHPK	KSSAAR	P0A7M9	P0A7X3	3	13	173.44	TRUE
KQQGHR	SKYGVK	P0AG48	P0A7S3	85	116	224.58	TRUE
KQQGHR	MAVQQNKPT	P0AG48	P0A7N4	85	7	193.84	TRUE
AMEKAR	TSGEKHLR	P0A7W1	P0A7N4	66	37	216.67	TRUE
KSSAAR	MAKTIK	P0A7X3	P0AG51	13	3	160.36	TRUE
HIGGGHKQA'	QPHAKGR	P60422	P25888	59	71	194.88	TRUE
AMEKAR	SGAIKAAK	P0A7W1	P0A6P1	66	48	204.26	TRUE
GEVKGK	HIGGGHKQAY	P0ABB0	P60422	318	59	206.21	TRUE
AMEKAR	ELKPHDR	P0A7W1	P0A9Q1	66	195	148.43	TRUE
TSGEKHLR	TSGEKHLR	P0A7N4	P0A7N4	37	37	168.81	TRUE

Figure 5:Structure of main "output" file

2. "[**name]_Number-XL.svg**" (Fig. 6A)

   Case 1: Real FDR is above 5%:
   First bar shows the type of crosslinks in the results delivered by the XL-search engines together with the real FDR. The second and third bars are obtained by filtering the results until the FDR reaches or less than or equal to 5% and 1%.

   Case 2: Real FDR is less than 5%, but greater than 1%:
   First bar shows the type of crosslinks in the results delivered by the XL-search engines together with the real FDR. The third bar is obtained by filtering the results until the FDR reaches less than or equal to 1%.

   Case 3: Real FDR is less than 1%:
   In this case there will be only one bar displaying the real FDR without any filtering

3. **"[name]_FDR-at-Score.svg"** (Fig. 6B)

   Case 1: real FDR is above 5%:
   First point: lowest accepted score, real FDR
   Second point: the lowest score at which the real FDR is less or equal to 5%
   Third point: the lowest score at which the real FDR is less or equal to 1%

   Case 2: real FDR is under 5% and above 1%
   First point: lowest accepted score, real FDR
   Second point: the lowest score at which the real FDR is less or equal to 1%

   Case 3: real FDR is under 1%
   First and only point: lowest accepted score, real FDR

4. **“[name]_Score-vs-Number.svg” (Fig. 6C)** shows the distribution of the crosslinks depending on the score.

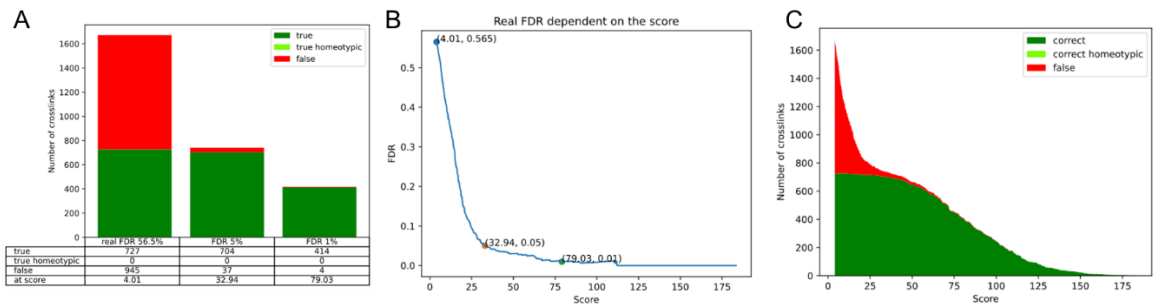

*Figure 6: **A**: # of unique XL sites as given by search engine and # of XL sites post-score cut-off to achieve a real FDR of 5 or 1% respectively, **B**: real FDR at specific score, **C**: XL sites at specific score*

5. “**[name]_output_venn_input.xlsx**”
   Is used as input for the Venn diagram option (as described in) and overcomes the different XL output formats by distinct search engines.

3. Venn diagrams

To generate a uniform data structure of identified cross links across different search engines, we have implemented the "[name]_venn_input.xlsx" file which is an output of "FDR-recalculation". The script selects common features of crosslinks, that are present in each software (i.e. position of XL in protein A and B, sequence A and B and protein name A and B) and sorts them alphabetically. The score is not considered by the Venn script, as each software has different methods of scoring a crosslink.

When comparing more than one result, Venn diagrams for all, "true" or "false" XLs together with more detailed information is saved in a ".xlsx". For a maximum of three result files weighted diagrams are generated, when comparing 4 results an unweighted Venn diagram will be displayed.

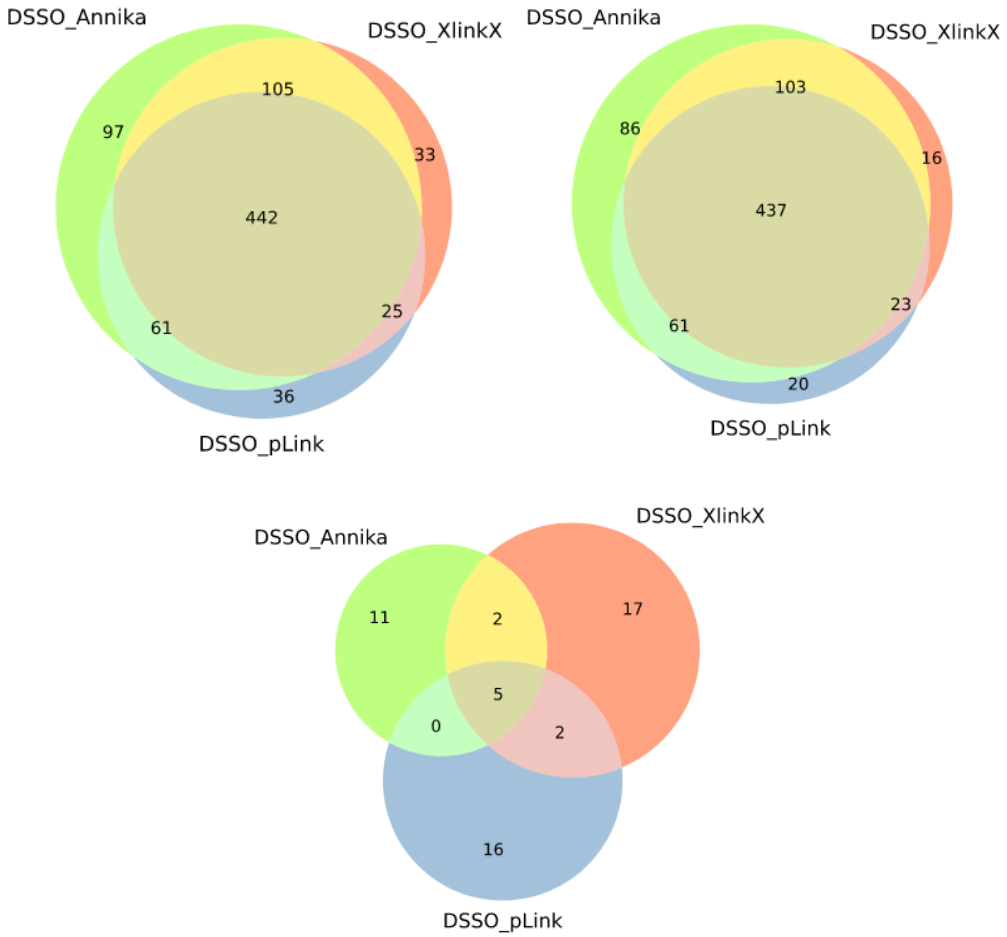

Figure 7: All, "true" or "false" XL-Venn diagrams

Furthermore, a document with three different sheets for all, "true" and "false" crosslinks will be created. If the overlap of specific crosslinks should be investigated or common features of false crosslinks identified, the user can find lists of the crosslinks present in each area within said document.

4. Physicochemical properties

All properties were calculated by using the Bio.SeqUtils package from Biopython (except for the frequency of the neighbours). The code was adapted for crosslinks but except for the mass determination the crosslinker itself was not considered.

Source:https://biopython.org/docs/1.75/api/Bio.SeqUtils.html.

The frequency of the nearest neighbours to the binding amino acid was calculated by using the package "seqlogo". Source: https://pypi.org/project/seqlogo/

4.1. Input file

This tool is only compatible with Proteome Discoverer results (i.e. “MS Annika” or “XlinkX”), which can be generated by exporting the (filtered) CSM list to Microsoft Excel.

4.2. Support library

There are two types of support libraries: the digested form and the undigested form. In case the user wants to follow MS-detection related properties, when the peptides were already digested, by trypsin for example, it is recommended to use the digested version of the library. In case, there is the thought, that the undigested version of the peptide can have a higher influence regarding the crosslinking process, the undigested version is recommended.

In any case, for the three existent libraries (Library 1, 2 and 3) the correct results with regard of the “WGGGGR” sequence will be delivered in the function “most frequent neighbours of the binding amino acid”, even though the digested form of the peptide library was chosen. In case, the user wants to utilize another library, an undigested form should be supplied for the function of the “most frequent neighbours of the binding amino acid” to work correctly. For all the other properties and functions the undigested form would be advantageous.

Additional information can be found in Chapter 2.2

4.3. Normalisation and comparison to theoretical crosslinks

We generate comparisons between “reality” and “theory”, which are calculated by assuming that each peptide from each group builds a crosslink (with only one CSM) with itself and with all the other peptides from the same group. Homeotypic duplicates are be filtered out of the pool.

Normalisation is performed as follows:

densitybool, default: False

> If True, draw and return a probability density: each bin will display the bins raw count divided by the total number of counts *and the bin width* (density = counts / (sum(counts) * np.diff(bins))), so that the area under the histogram integrates to 1 (np.sum(density * np.diff(bins)) == 1)
>
> If *stacked* is also True, the sum of the histograms is normalized to 1.

Source: https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.hist.html

Through the normalisation process, the area of all the bins in histograms summed up will be equal to 1.

For normalization, both “theory” and “reality” results are normalized to 1 (“density=True”). Due to that, some bins from the real results can be higher than the bins from the theoretically maximum reachable numbers of crosslinks (e.g., Peptide library 1 = 1018). The normalization is done on the CSM level, consequently a crosslink with 10 CSMs, can influence the result 10 times more than the same crosslink with just one CSM.

In the unnormalized version, the area of all bins will be not equal, therefore neither the “reality”, nor the “theory” will be equal to one. In this case, the analysis is done on the crosslink level, hence duplicate CSMs will be eliminated, and two crosslinks have the same weight in the results, regardless of their CSM counts

4.4. Output Files

4.4.1. Isoelectric Point

The pI value is calculated for each peptide of the identified crosslink and is weighted by the number of K, R, H, D, E present.

AHHKEGGR Peptide A
|
MTKEEGGR Peptide B

$$pI\ value(Crosslink) = \frac{pI\ value(Peptide\ A) * 5\ +\ pI\ value(Peptide\ B) * 4}{5 + 4}$$

In the case of pI calculation, the lysins connected by the crosslinker are also included in the algorithm to focus on the isoelectric properties of the initial peptides influencing the formation of crosslinks.

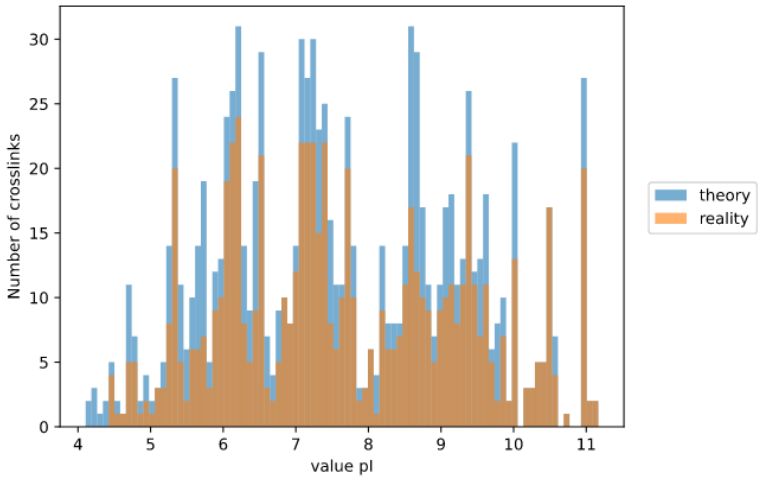

Figure 8: Isoelectric point; XL level; areas not normalised; weighted to frequency of KRHDE

4.4.2. Hydrophobicity

The gravy index of the crosslink is calculated based on Kyte and Doolittle scale (Kyte J., Doolittle R. F., J Mol Biol., 1982). Every amino acid possesses an empirically determined gravy value, which is used to calculate a weighted average considering the length of both peptides. Of note even though the crosslinker itself is not included the analysis allows for a comparative overview to the theoretical results.

$Gravy\ value(Crosslink)$ =

$$= \frac{[Gravy\ value(Peptide\ A) *\ Length(Peptide\ A\ ) +\ \ Gravy\ value(Peptide\ B) *\ Length(Peptide\ B)]}{Length(Peptide\ A) + Length(Peptide\ B)}$$

Aromaticity- Frequency of aromatic amino acids

R	K	N	D	Q	E	H	P	Y	W	S	T	G	A	M	C	F	L	V	I
-4.5	-3.9	-3.5	-3.5	-3.5	-3.5	-3.2	-1.6	-1.3	-0.9	-0.8	-0.7	-0.4	1.8	1.9	2.5	2.8	3.8	4.2	4.5

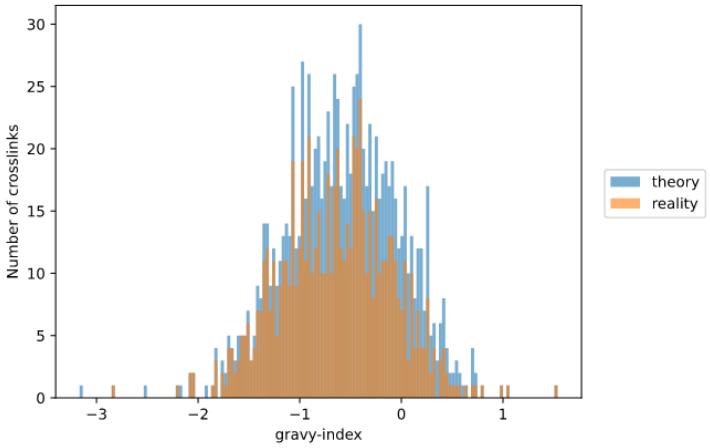

Figure 9:Hydrophobicity; XL level; areas not normalised

4.4.3. Frequency of amino acids in crosslinks

The image below shows the frequency (# specific amino acid/ total number of amino acids per CSM) of each amino acid regardless of their position in the peptide across all CSMs. The crosslinker reactive amino acids are also included in the calculation.

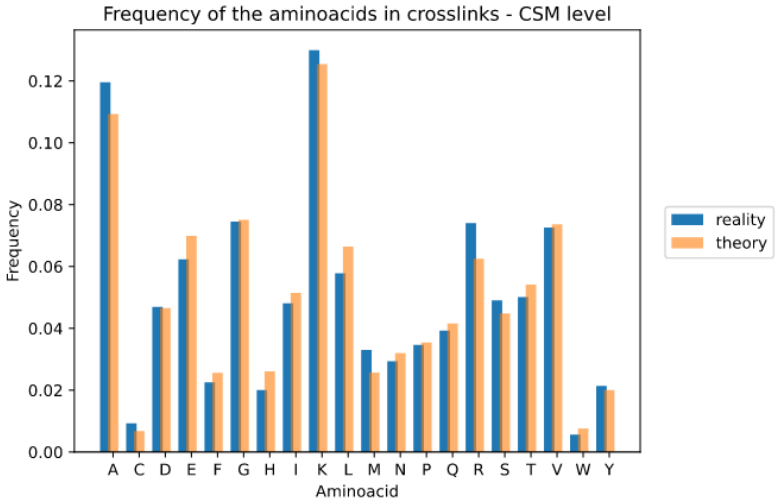

Figure 10: Frequency of each amino acid within all identified CSMs of a file

4.4.4. Frequency of aromatic amino acids (F, W, Y)

The acids phenylalanine (F), tryptophan (W) and tyrosine (Y) are considered aromatic. As these amino acids might alter the reactivity of neighbouring functional groups and as they are sterically demanding their frequency might give information if there is any influence on the formation/ ability for detection of a crosslink.

Calculation is explained based on the following example:

AA**Y**MTKEMPR Peptide A (length = 10, number of aromatic amino acids = 1)
|
ST**W**PQNKPQ**F**R Peptide B (length = 11, number of aromatic amino acids = 2)

$$Frequency\ of\ aromaticity\ (Crosslink) =$$

$$= \frac{[Frequency\ of\ aromaticity(Peptide\ A) * Length(Peptide\ A) + Frequency\ of\ aromaticity(Peptide\ B) * Length(Peptide\ B)]}{Length(Peptide\ A) + Length(Peptide\ B)}$$

$$Frequency\ of\ aromaticity\ (Crosslink) = \frac{\left[\left(\frac{1}{10}\right) * 10\ +\ \left(\frac{2}{11}\right) * 11\right]}{10 + 11}$$

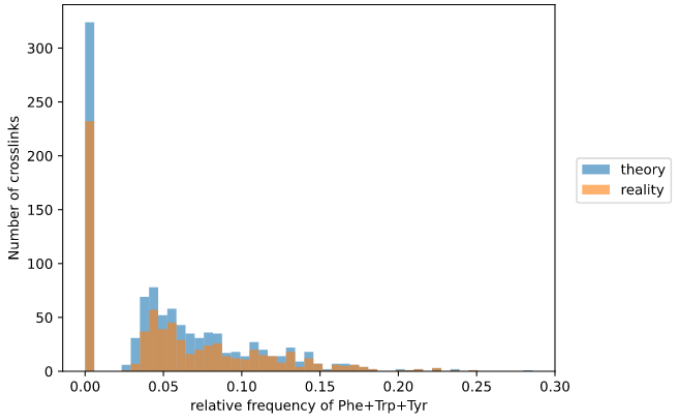

Figure 11: Aromaticity XL level areas not normalised

4.4.5. Molecular weight

For the molecular weight the mass of the connected peptides and the crosslinker are summed. Importantly, the program ignores all other chemical modifications, which might alter the calculated mass vs real mass.

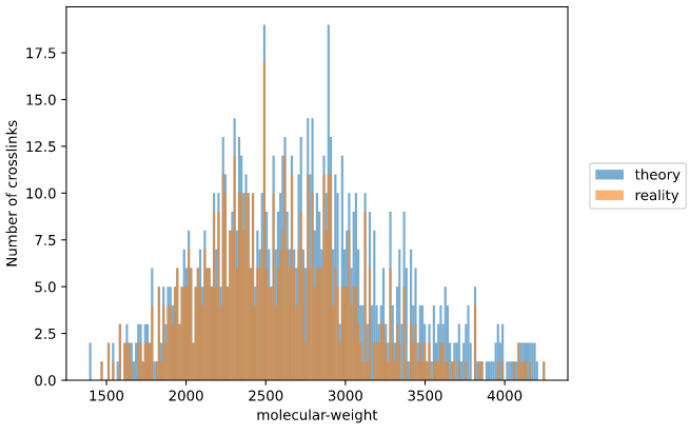

Figure 12:Molecular weight XL level areas not normalised

4.4.6. Neighbours of the crosslinkers binding amino acid

To address if the direct chemical environment of crosslinker reactive amino acids influences its reactivity we report three amino acids to the left and to right of the crosslinker bound amino acid. Below we show the position frequency matrix structure from the analysed structures, where each row represents the following characteristics:

- first row - third neighbour from the left (farthest left)
- second row - second left neighbour (middle left located)
- third row - first left neighbour (closest left)
- fourth row - first neighbour on the right (closest right)
- fifth row - second neighbour from the right (middle located right)
- sixth row - third neighbour on the right (farthest right)

The most frequent amino acid residues neighboring the binding amino acid are:

GGA ~binding amino acid~ A--

Position frequency matrix of the closest neighbours to the binding amino acid from left (N-Terminus-direction) to right (C-Terminus-direction)
Each row of the matrix represents a distinct position reported to the binding amino acid.

A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	X	*	-
0.125000	0.013057	0.016053	0.046233	0.012414	0.251284	0.012200	0.029966	0.000000	0.040026	0.046447	0.046447	0.003853	0.030822	0.128425	0.071490	0.032962	0.081122	0.00000	0.012200	0.0	0.0	0.000000
0.093108	0.000000	0.046447	0.027825	0.040454	0.183861	0.012628	0.045591	0.000000	0.058219	0.033818	0.035103	0.033176	0.054580	0.072132	0.072988	0.084760	0.069349	0.00899	0.026969	0.0	0.0	0.000000
0.148330	0.044092	0.082192	0.108091	0.008990	0.064212	0.017123	0.085616	0.000000	0.048587	0.061858	0.030394	0.031036	0.036387	0.112158	0.034461	0.032320	0.054152	0.00000	0.000000	0.0	0.0	0.000000
0.205051	0.010702	0.028896	0.065068	0.031892	0.070848	0.037671	0.061430	0.036601	0.097817	0.016053	0.000000	0.041738	0.088399	0.067423	0.031464	0.042594	0.029324	0.00000	0.037029	0.0	0.0	0.000000
0.101884	0.000000	0.067637	0.014983	0.011772	0.069563	0.019264	0.074272	0.041310	0.067851	0.025471	0.029324	0.011772	0.063570	0.063142	0.048159	0.093964	0.075985	0.00000	0.016053	0.0	0.0	0.104024
0.063142	0.000000	0.031678	0.054152	0.037243	0.032106	0.012842	0.032320	0.047517	0.073630	0.014341	0.082620	0.029966	0.015197	0.134204	0.052654	0.012842	0.049229	0.00000	0.019692	0.0	0.0	0.204623

Figure 13: neighbours of the crosslinker's binding position CSM

We report in the one-letter amino acid code and the special characters "*", "X" and "-"

X	Any
*	Translation stop
-	empty space (beyond the end of peptide sequence)

4.4.7. Retention time vs physicochemical properties

We here display the retention time versus above-described physiochemical properties of each CSM.

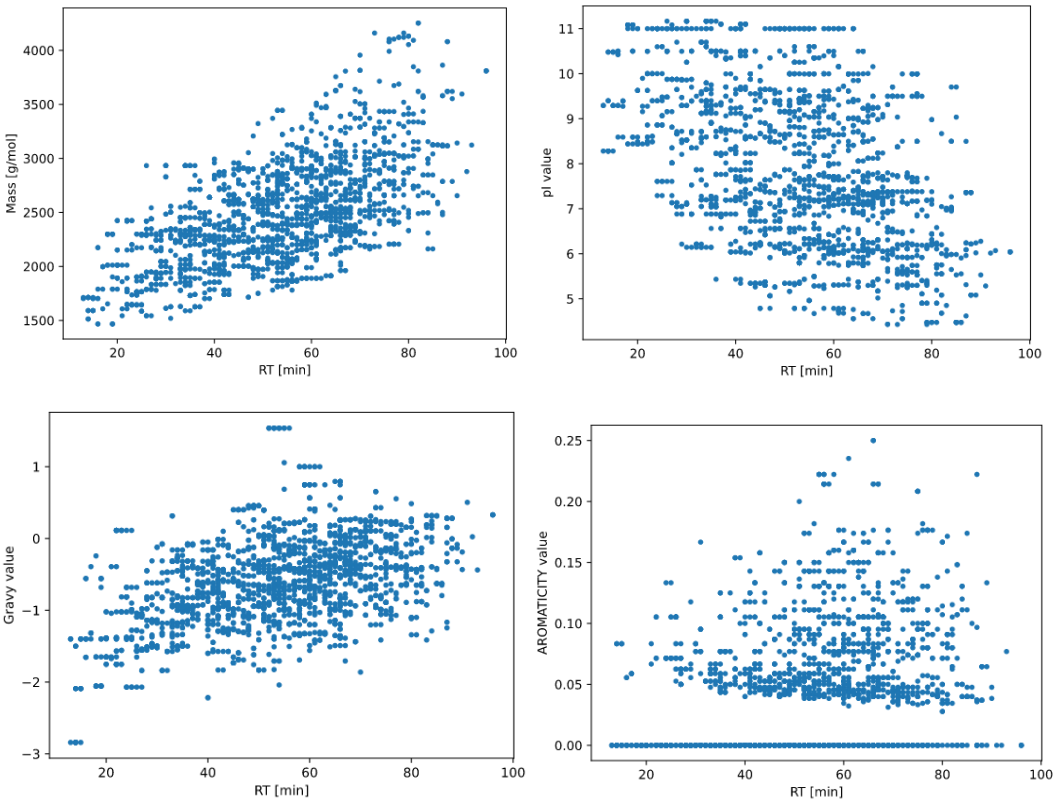

Figure 14: physicochemical (mass-top left; isoelectric point-top right; gravy value-bottom left; aromaticity fraction-bottom right) properties over retention time