

MS Synthetic Library

BC. PETRONELA BELKOVÁ

BC. JOZEF VARGAN

BC. MICHAL HOLUB

2022/2023

Problem Definition

- Mass spectrometry:
 - Analytical method
 - Unrivalled sensitivity
 - Low detection limits
 - Huge amount of data
- Synthetic spectral libraries:
 - Cover blank spaces in database
 - Better availability and retrievability of compounds

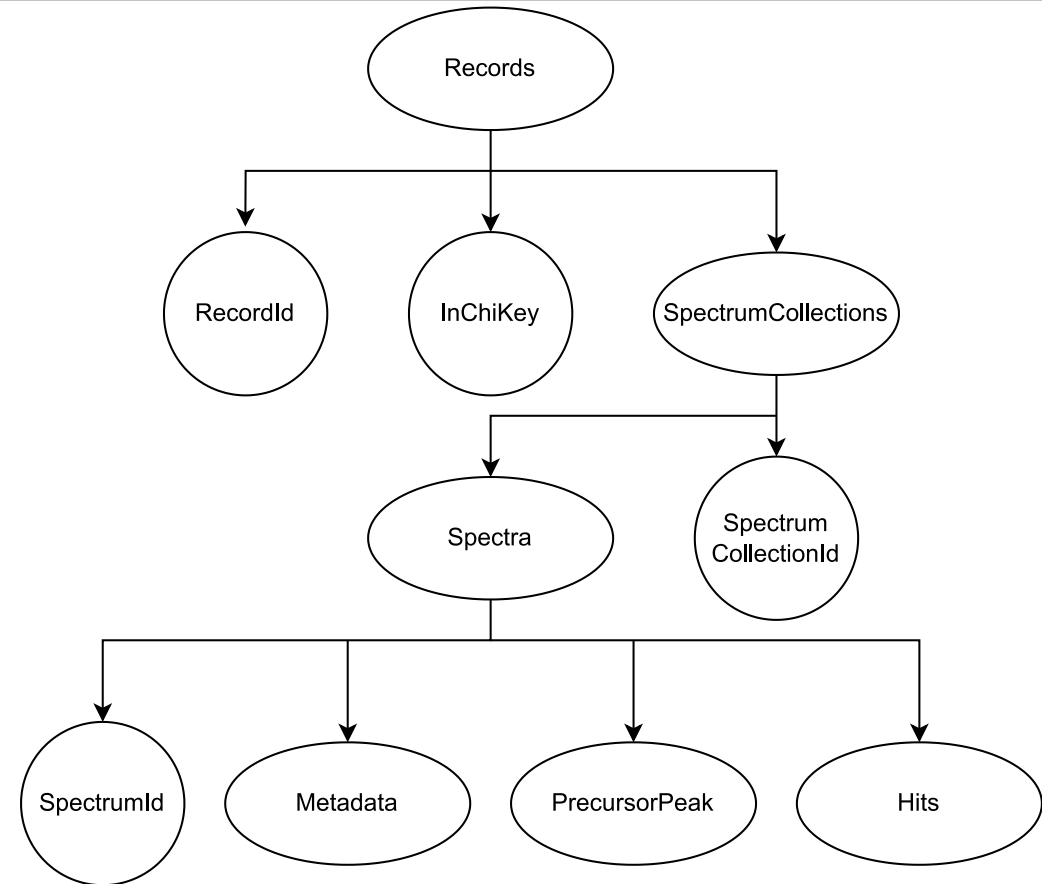
Aims of the Project

- Load mass spectrometry data from two different in-silico spectral libraries
- Evaluate defined criteria on both libraries
- Compare performance of both libraries

Results

Data Extraction

- Format: JSON
- More than 90 records for each library
- Each record has metadata
- Each record can have multiple hits



Results

Create a binary column of whether the hit is true-positive or false-positive

SYNTHETIC SEARCH

Hit	InChIKey	InChIKey_hits	Positivity
0	33	33	True
1	33	27	False
2	33	27	False
...
299	34	34	True

BARCODE SEARCH

Hit	InChIKey	InChIKey_hits	Positivity
0	33	33	True
1	33	27	False
2	24	24	True
...
266	34	69	False

Results

Compute the difference between experimental conditions

SYNTHETIC SEARCH

Hit	Polarity	ScanRange Max	ScanRange Min	NCE	Ion Activation
0	1.0	0.194655	0.0	0.15790	1.0
1	1.0	0.194655	0.0	0.15790	1.0
2	1.0	0.194655	0.0	0.02632	1.0
...
299	1.0	0.113322	0.0	0.10526	0.0

BARCODE SEARCH

Hit	Polarity	ScanRange Max	ScanRange Min	NCE	Ion Activation
0	0.0	0.194655	0.0	0.05263	1.0
1	0.0	0.194655	0.0	0.05263	1.0
2	0.0	0.194821	0.0	0.05263	0.0
...
266	0.0	0.113322	0.0	0.05263	0.0

Results

For each compound compute the number of correct hits and incorrect hits

SYNTHETIC SEARCH

recordId	InChIKey	Correct hits	Incorrect hits
0	33	1	2
1	24	1	0
2	44	2	1
...
94	34	2	0

BARCODE SEARCH

recordId	InChIKey	Correct hits	Incorrect hits
0	33	1	1
1	24	1	0
2	44	1	1
...
93	34	1	1

Results

Relate the ranking of candidates based on their score to the number of potential hits

SYNTHETIC SEARCH

Hit	Cosine relative rank	Denver relative rank	NIST relative rank
0	0.0	0.0	0.0
1	0.5	0.5	0.5
2	1.0	1.0	1.0
...
299	0.0	1.0	0.0

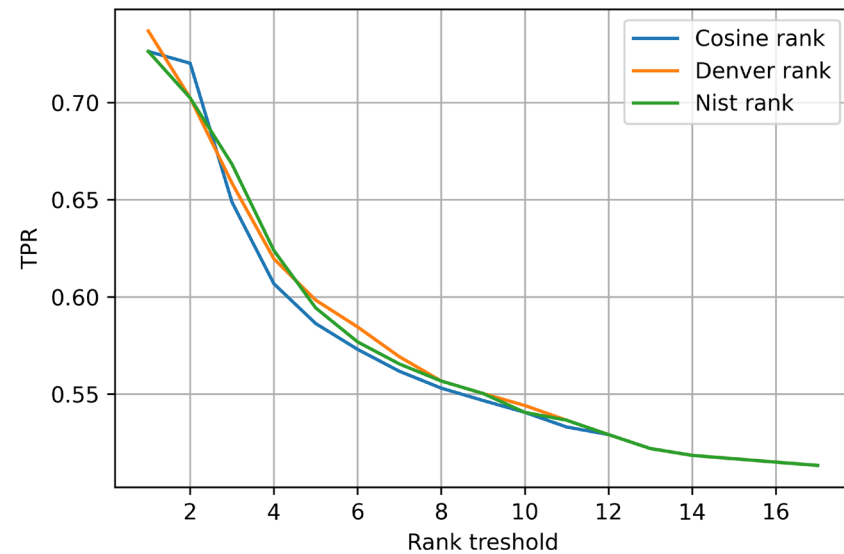
BARCODE SEARCH

Hit	Cosine relative rank	Denver relative rank	NIST relative rank
0	0.0	0.0	0.0
1	1.0	1.0	1.0
2	0.0	0.0	0.0
...
266	1.0	1.0	1.0

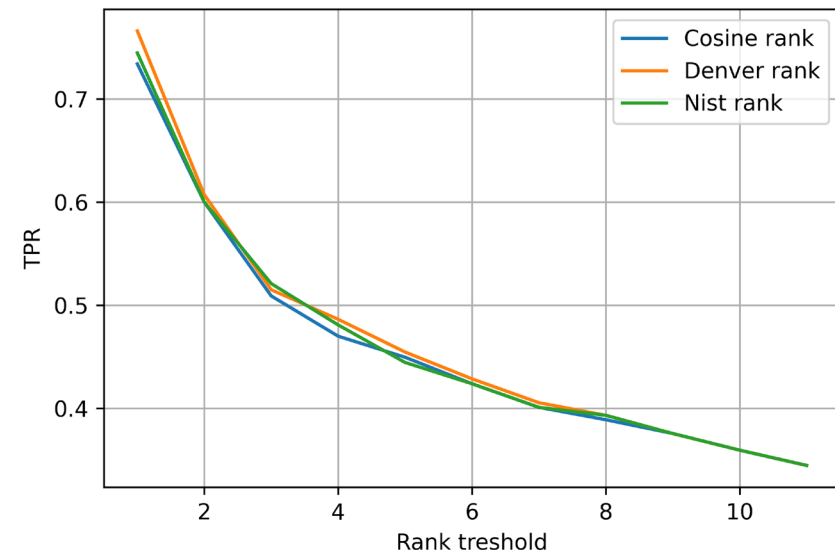
Results

Plot true-positive rate vs. rank threshold

SYNTHETIC SEARCH



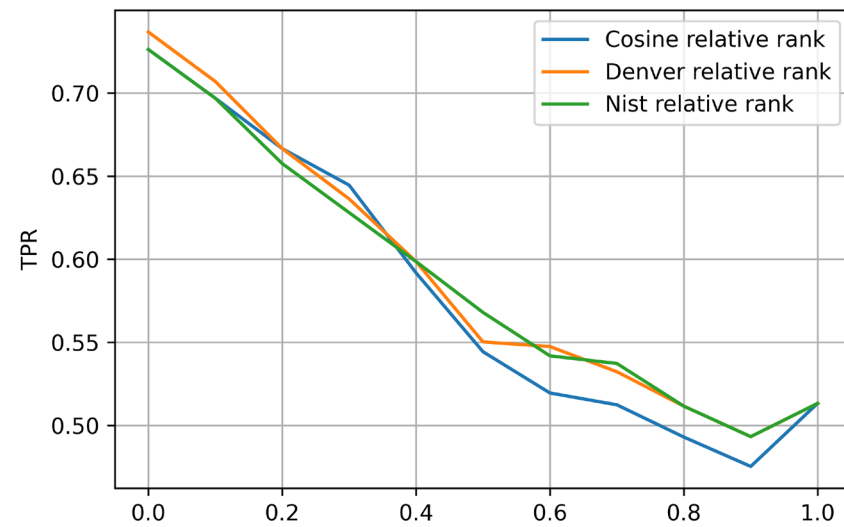
BARCODE SEARCH



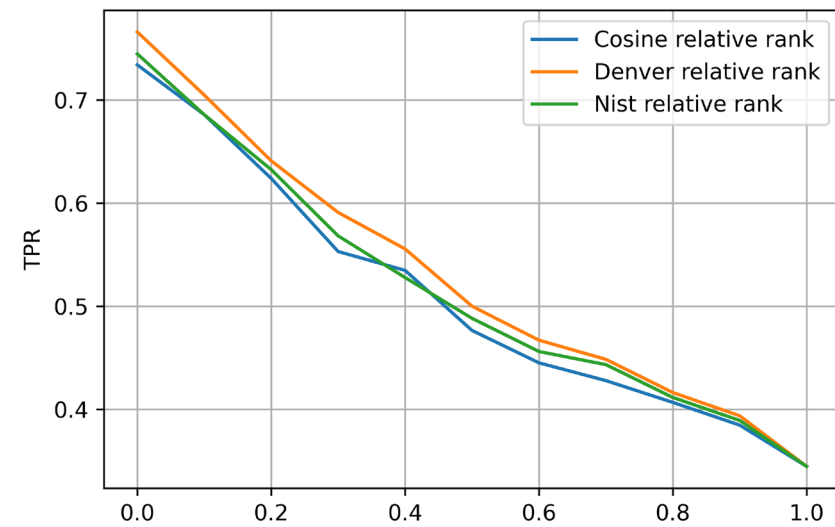
Results

Plot true-positive rate vs. relative rank threshold

SYNTHETIC SEARCH



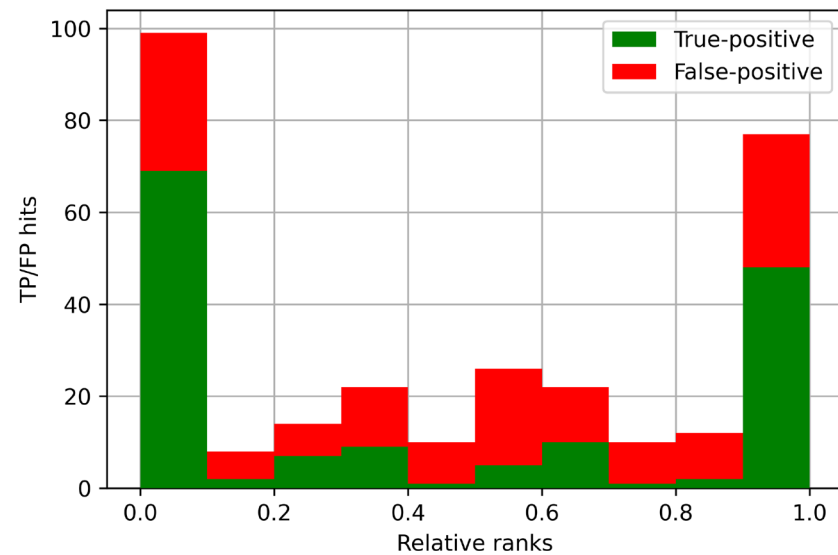
BARCODE SEARCH



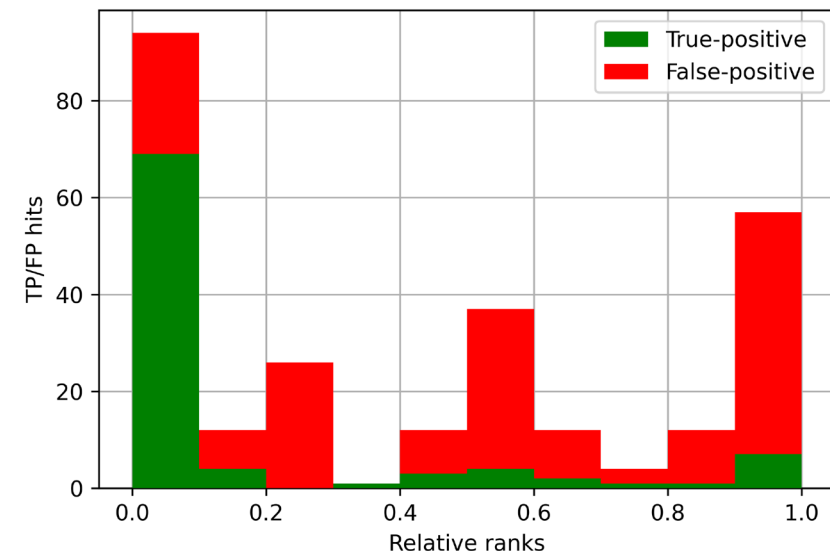
Results

Plot histogram for each metric distinguishing true-positive hits and false-positive hits

SYNTHETIC SEARCH – COSINE RANK



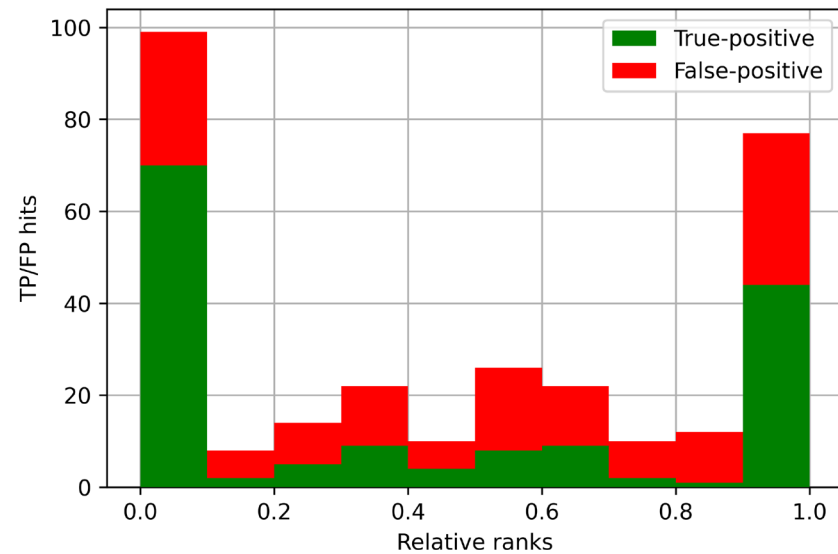
BARCODE SEARCH – COSINE RANK



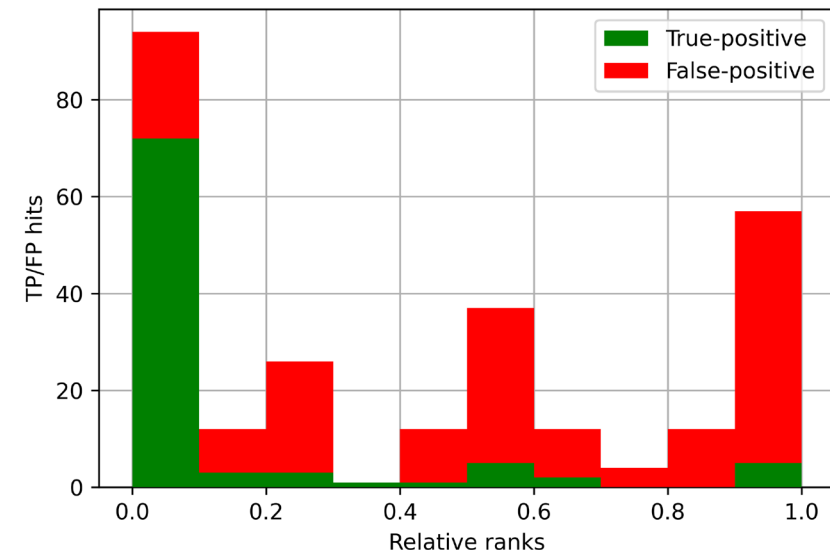
Results

Plot histogram for each metric distinguishing true-positive hits and false-positive hits

SYNTHETIC SEARCH – DENVER RANK



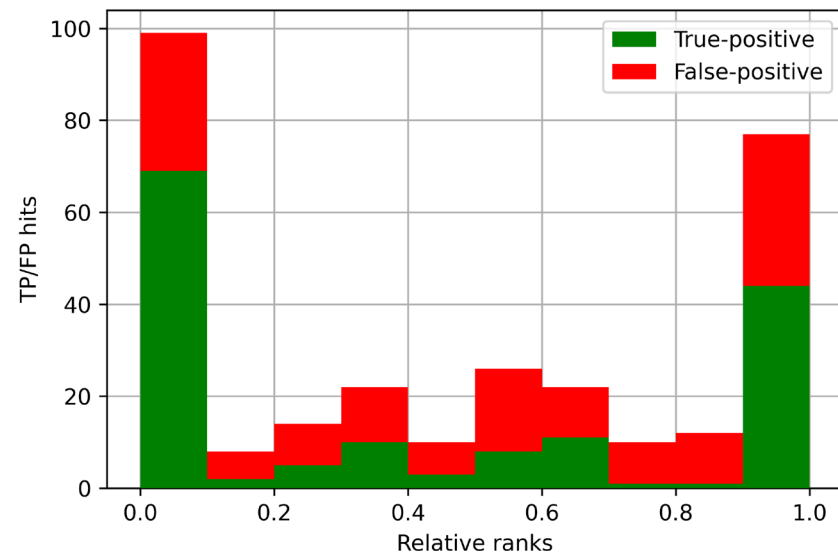
BARCODE SEARCH – DENVER RANK



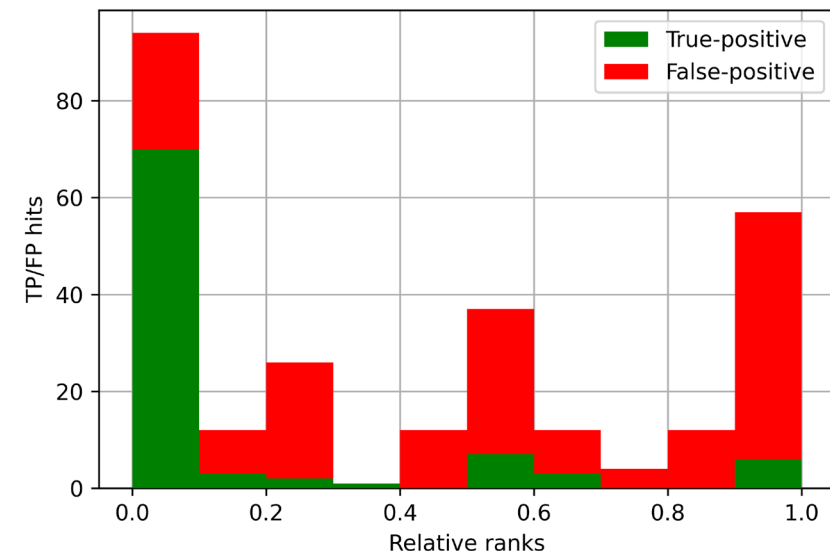
Results

Plot histogram for each metric distinguishing true-positive hits and false-positive hits

SYNTHETIC SEARCH – NIST RANK



BARCODE SEARCH – NIST RANK



Results

Render correlation between deviation in experimental conditions and TP/FP hits

SYNTHETIC SEARCH

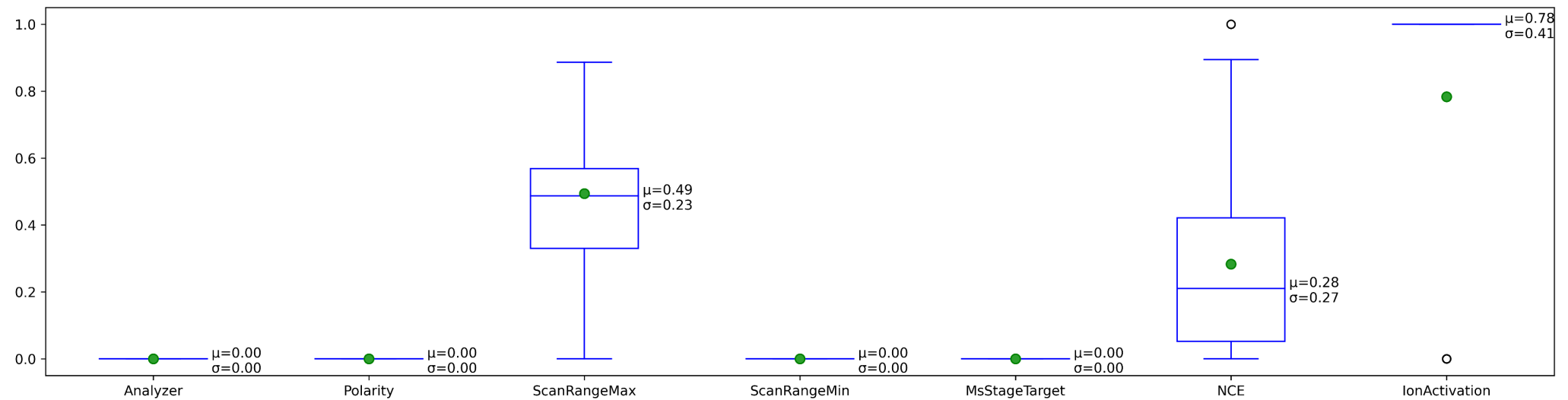
	Pearson	Spearman
Analyzer	-	-
Polarity	-	-
ScanRangeMax	0.1276	0.1248
ScanRangeMin	-	-
MsStageTarget	-	-
NCE	0.0904	0.0811
IonActivation	-0.0707	-0.0707

BARCODE SEARCH

	Pearson	Spearman
Analyzer	-	-
Polarity	-	-
ScanRangeMax	0.2063	0.1743
ScanRangeMin	0.1197	0.1198
MsStageTarget	-	-
NCE	0.2161	0.2710
IonActivation	0.1451	0.1451

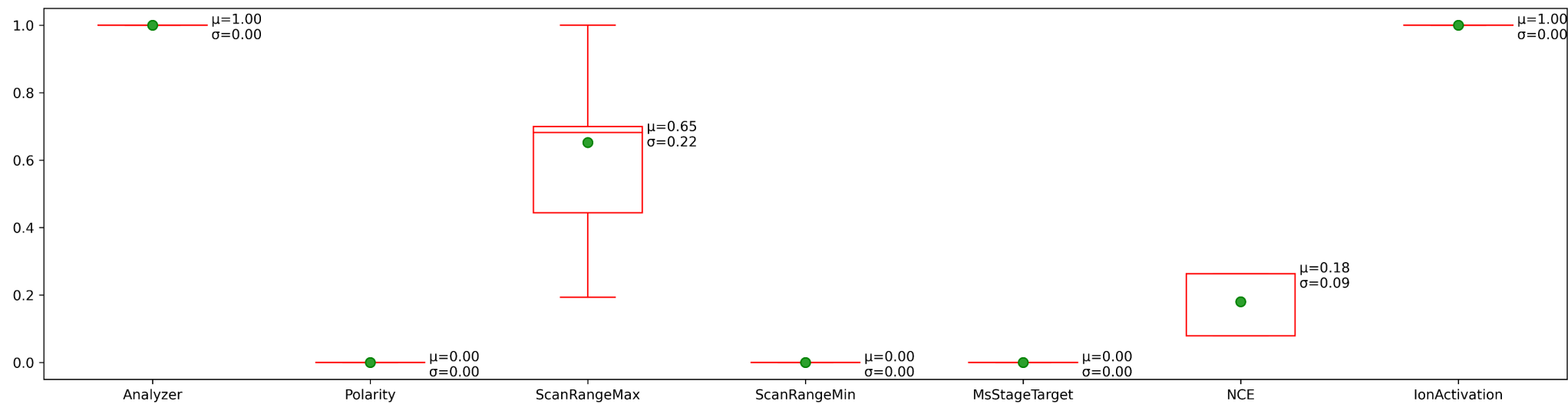
Results

Synthetic search – metadata



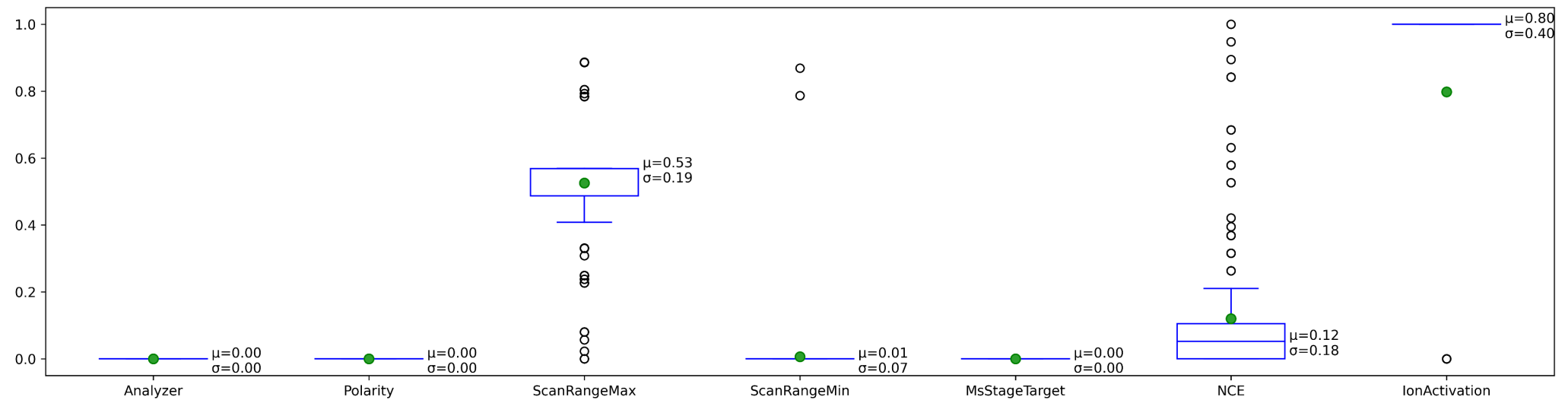
Results

Synthetic search – metadata hits



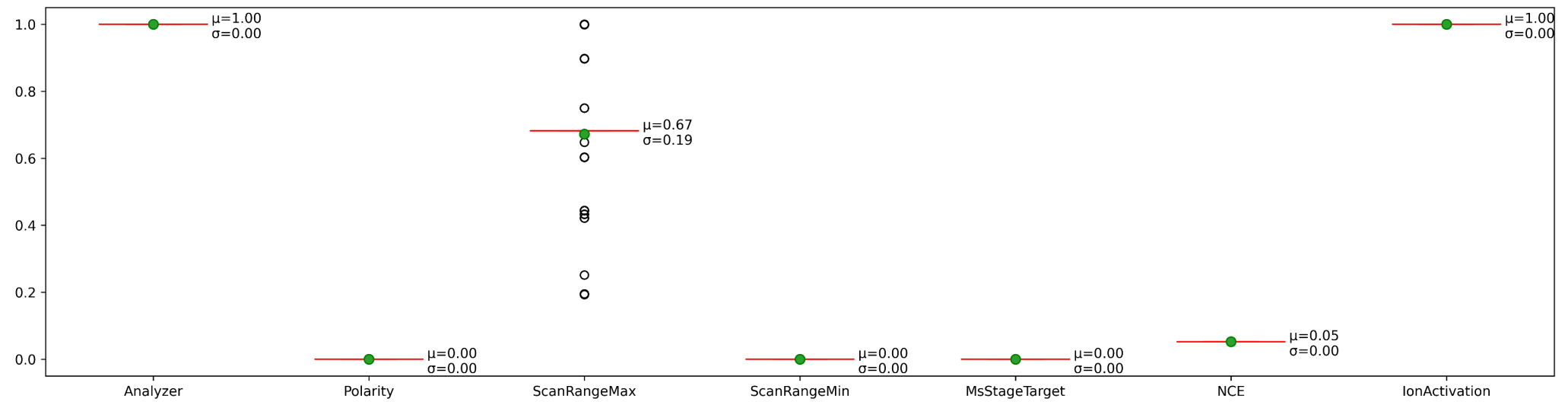
Results

Barcode search – metadata



Results

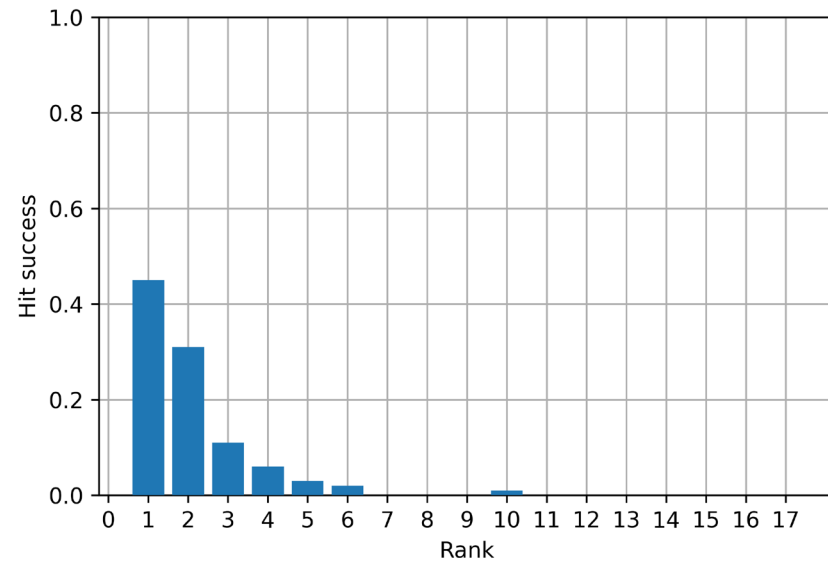
Barcode search – metadata hits



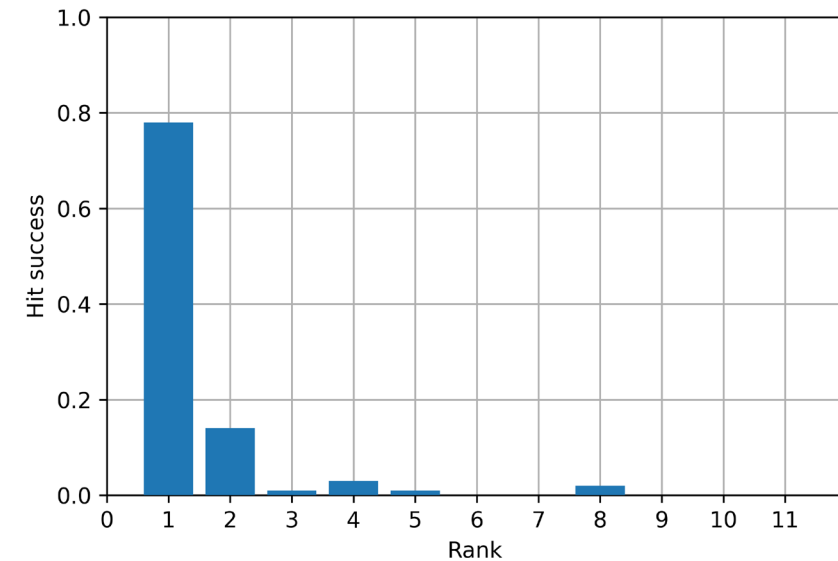
Results

Find a reasonable way to compare the search performance of two libraries

SYNTHETIC SEARCH – SUCCESS RATE



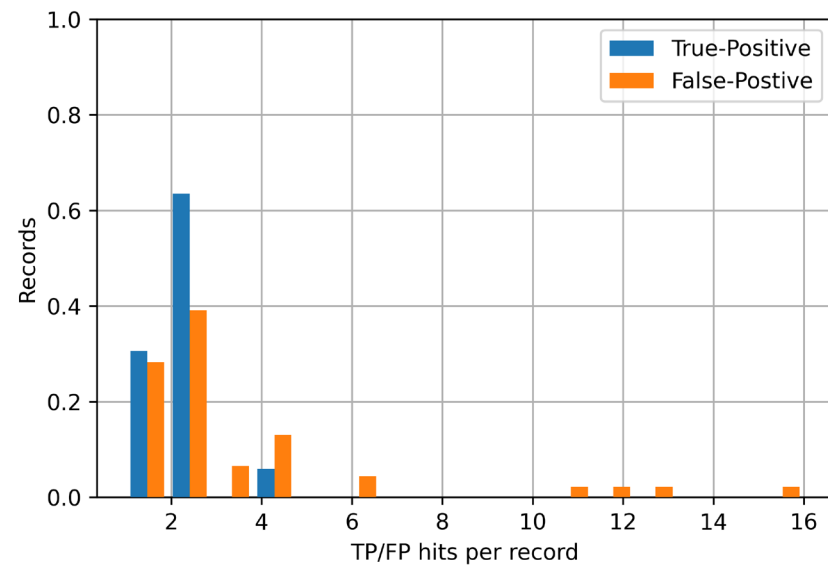
BARCODE SEARCH – SUCCESS RATE



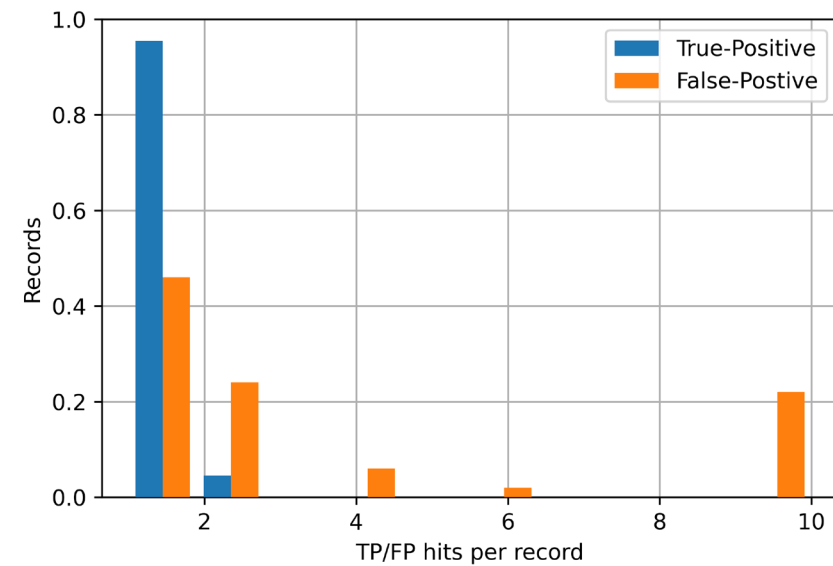
Results

Find a reasonable way to compare the search performance of two libraries

SYNTHETIC SEARCH



BARCODE SEARCH



Results

Find a reasonable way to compare the search performance of two libraries

SYNTHETIC SEARCH

Synthetic	
Total records	95
At least one correct hit	89.47%
Only incorrect hits	10.53%

BARCODE SEARCH

Barcode	
Total records	94
At least one correct hit	93.62%
Only incorrect hits	6.38%