SCIENCE MEETS LIFE

MS²PIP: Predicting peptide spectrum peak intensities to improve proteomics identification

Ralf Gabriels

VIB-UGENT CENTER FOR MEDICAL BIOTECHNOLOGY

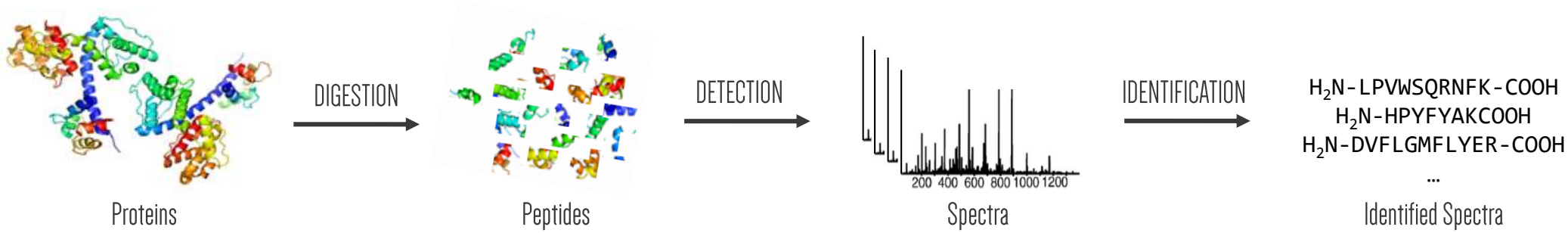
UNIVERSITEIT GENT

CompOmics

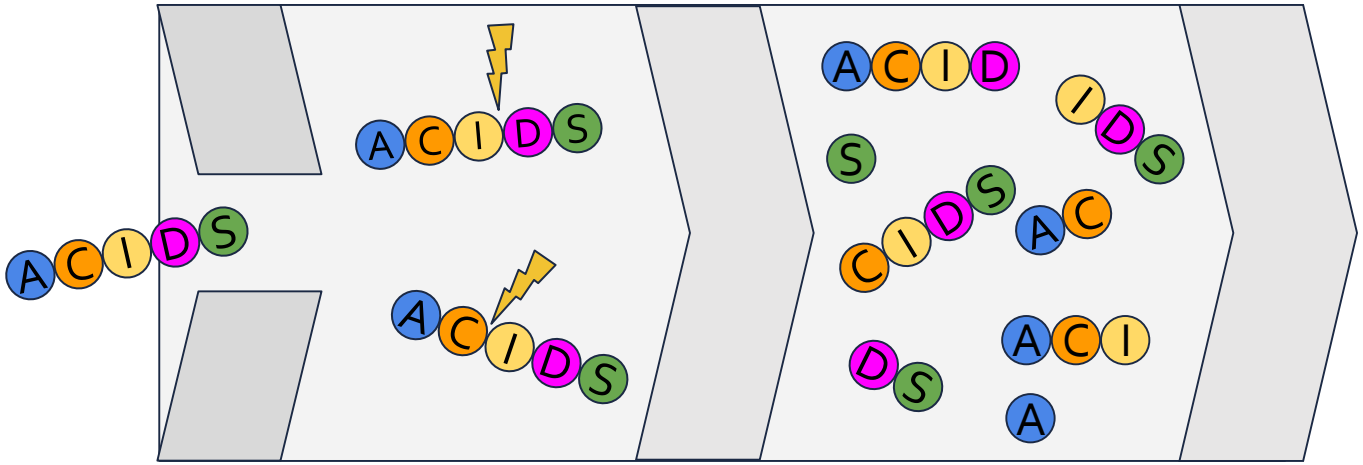
18/07/2019

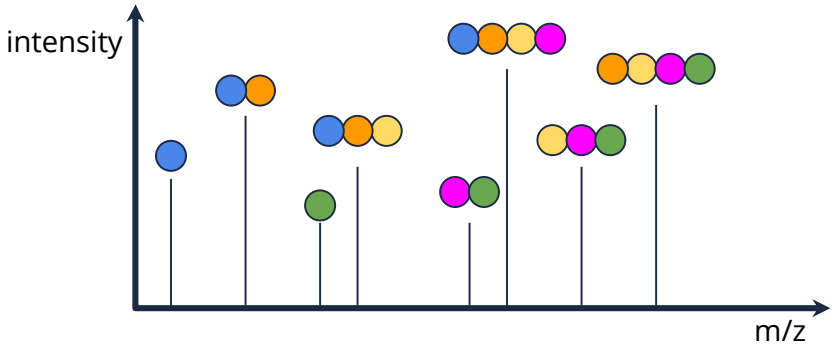
BePA Mini Symposium on Statistics in Proteomics, ISCB2019

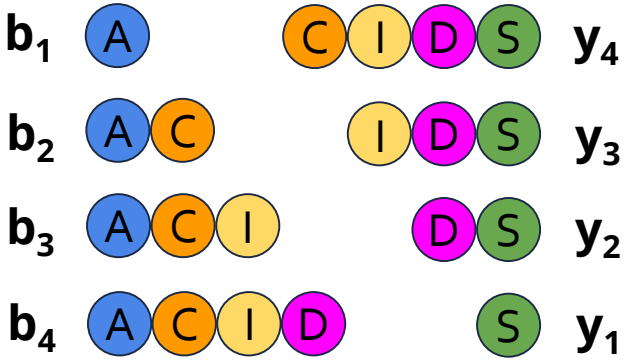
In MS-based proteomics, peptides are identified by their fragmentation spectra



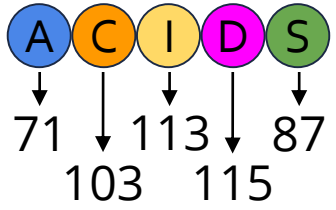
For every peptide, a fragmentation spectrum is generated







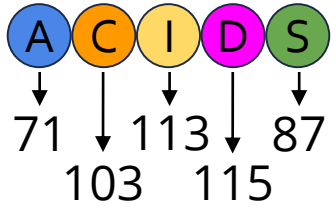
We can easily calculate the m/z values for any given peptide spectrum



Amino Acid	Chemical formula	Molecular mass
A	C_3H_5ON	71.03711
R	$C_6H_{12}ON_4$	156.10111
N	$C_4H_6O_2N_2$	114.04293
D	$C_4H_5O_3N$	115.02694
C	C_3H_5ONS	103.00919
E	$C_5H_7O_3N$	129.04259
Q	$C_5H_8O_2N_2$	128.05858
G	C_2H_3ON	57.02146
H	$C_6H_7ON_3$	137.05891
I	$C_6H_{11}ON$	113.08406
L	$C_6H_{11}ON$	113.08406
K	$C_6H_{12}ON_2$	128.09496
M	C_5H_9ONS	131.04049
F	C_9H_9ON	147.06841
P	C_5H_7ON	97.05276
S	$C_3H_5O_2N$	87.03203
T	$C_4H_7O_2N$	101.04768
W	$C_{11}H_{10}ON_2$	186.07931
Y	$C_9H_9O_2N$	163.06333
V	C_5H_9ON	99.06841

We can easily calculate the m/z values for any given peptide spectrum

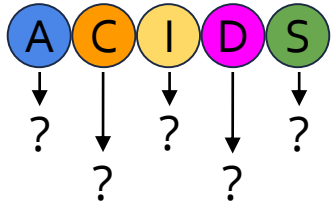
Amino Acid	Chemical formula	Mass
A	C_3H_5ON	71.03711
R	$C_6H_{12}ON_4$	156.10111
N	$C_4H_6O_2N_2$	114.04293
D	$C_4H_5O_3N$	115.02694
C	C_3H_5ONS	103.00919
E	$C_5H_7O_3N$	129.04259
Q	$C_5H_8O_2N_2$	128.05858
G	C_2H_3ON	57.02146
H	$C_6H_7ON_3$	137.05891
I	$C_6H_{11}ON$	113.08406
L	$C_6H_{11}ON$	113.08406
K	$C_6H_{12}ON_2$	128.09496
M	C_5H_9ONS	131.04049
F	C_9H_9ON	147.06841
P	C_5H_7ON	97.05276
S	$C_3H_5O_2N$	87.03203
T	$C_4H_7O_2N$	101.04768
W	$C_{11}H_{10}ON_2$	186.07931
Y	$C_9H_9O_2N$	163.06333
V	C_5H_9ON	99.06841



ion	x-axis: m/z
A	72,04435
A C	175,0535
A C I	288,1376
A C I D	403,1646
C I D S	437,17
I D S	334,1608
D S	221,0768
S	106,0498

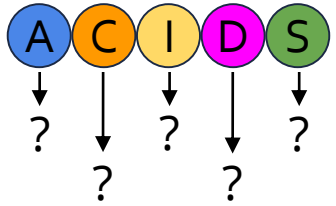
But how do we get the intensity values?

Amino Acid	Chemical formula	Mass
A	C_3H_5ON	71.03711
R	$C_6H_{12}ON_4$	156.10111
N	[illegible]	114.04293
D	[illegible]	115.02694
C	[illegible]	103.00919
E	$C_5H_7O_3N$	129.04259
Q	$C_5H_8O_2N_2$	128.05858
G	C_2H_3ON	57.02146
H	C_6H_7ON	137.05891
I	[illegible]	113.08406
L	[illegible]	113.08406
K	[illegible]	128.09496
M	[illegible]	131.04049
F	[illegible]	147.06841
P	C_5H_7ON	97.05276
S	[illegible]	87.03203
T	[illegible]	101.04768
W	[illegible]	186.07931
Y	$C_9H_9O_2N$	163.06333
V	C_5H_9ON	99.06841



ion	x-axis: m/z	y-axis: intensity
A	72,04435	?
A C	175,0535	?
A C I	288,1376	?
A C I D	403,1646	?
C I D S	437,17	?
I D S	334,1608	?
D S	221,0768	?
S	106,0498	?

Every amino acid has known physicochemical properties



Basicity	Helicity	Hydrophobicity	pI	Molecular mass
37	68	51	32	71.03711
35	23	75	23	156.10111
59	33	25	0	114.04293
129	29	35	4	115.02694
94	70	100	27	103.00919
0	58	16	32	129.04259
210	41	3	48	128.05858
81	73	94	32	57.02146
191	32	0	69	137.05891
106	66	82	29	113.08406
101	38	12	26	113.08406
117	0	0	35	128.09496
115	40	22	28	131.04049
343	39	22	79	147.06841
49	44	21	29	97.05276
90	53	39	28	87.03203
60	71	80	31	101.04768
134	51	98	31	186.07931
104	55	70	28	163.06333

Machine learning enables us to predict intensities based on these properties

Basicity	Helicity	Hydrophobicity	pI	Molecular mass
37	68	51	32	71.03711
35	23	75	23	156.10111
59	33	25	0	114.04293
129	29	35	4	115.02694
94	70	100	27	103.00919
0	58	16	32	129.04259
210	41	3	48	128.05858
81	73	94	32	57.02146
191	32	0	69	137.05891
106	66	82	29	113.08406
101	38	12	26	113.08406
117	0	0	35	128.09496
115	40	22	28	131.04049
343	39	22	79	147.06841
49	44	21	29	97.05276
90	53	39	28	87.03203
60	71	80	31	101.04768
134	51	98	31	186.07931
104	55	70	28	163.06333

A C I D S

ion	x-axis: m/z	y-axis: intensity
A	72,04435	?
A C	175,0535	?
A C I	288,1376	?
A C I D	403,1646	?
C I D S	437,17	?
I D S	334,1608	?
D S	221,0768	?
S	106,0498	?

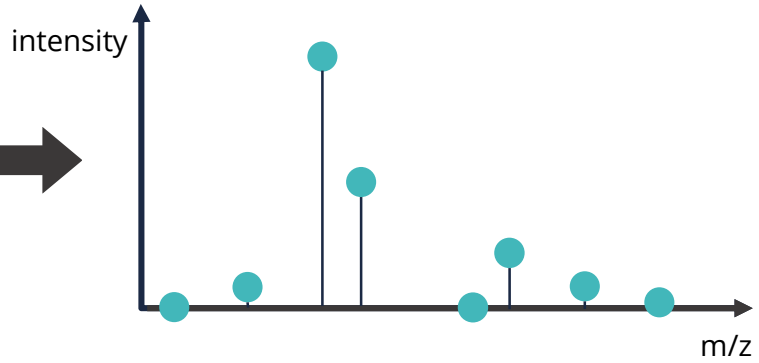
Machine learning enables us to predict intensities based on these properties

Basicity	Helicity	Hydrophobicity	pI	Molecular mass
37	68	51	32	71.03711
35	23	75	23	156.10111
59	33	25	0	114.04293
129	29	35	4	115.02694
94	70	100	27	103.00919
0	58	16	32	129.04259
210	41	3	48	128.05858
81	73	94	32	57.02146
191	32	0	69	137.05891
106	66	82	29	113.08406
101	38	12	26	113.08406
117	0	0	35	128.09496
115	40	22	28	131.04049
343	39	22	79	147.06841
49	44	21	29	97.05276
90	53	39	28	87.03203
60	71	80	31	101.04768
134	51	98	31	186.07931
104	55	70	28	163.06333

A C I D S

MS²PIP

ion	x-axis: m/z	y-axis: intensity
A	72,04435	0.002482
A C	175,0535	0.729443
A C I	288,1376	0.000000
A C I D	403,1646	0.003302
C I D S	437,17	0.004636
I D S	334,1608	0.057674
D S	221,0768	0.194737
S	106,0498	0.007725



Machine learning enables us to predict intensities based on these properties

→ **MS²PIP** →

ion	Target: intensities
A	0.002482
A C	0.729443
A C I	0.000000
A C I D	0.003302
C I D S	0.004636
I D S	0.057674
D S	0.194737
S	0.007725

In MS²PIP, one input leads to multiple targets

Input
Piece of fruit

→

Feature 1	Feature 2	Feature 3	...	Feature n
Yellowness	Acidity	Hardness	...	Weight

→

Target
Banana

Input
Peptide

→

Feature 1	Feature 2	Feature 3	...	Feature n
Basicity?	Hydro.?	Helicity?	...	Ip?

→

Target
int. b_1
int. b_2
int. b_3
...

In MS²PIP, one input leads to multiple targets, leading to multiple feature sets

Input
Piece of fruit

→

Feature 1	Feature 2	Feature 3	...	Feature n
Yellowness	Acidity	Hardness	...	Weight

→

Target
Banana

Input
Peptide

→

Feature 1	Feature 2	Feature 3	...	Feature n
Basicity b_1?	Hydro. b_1?	Helicity b_1?	...	Ip b_1?
Basicity b_2?	Hydro. b_2?	Helicity b_2?	...	Ip b_2?
Basicity b_3?	Hydro. b_3?	Helicity b_3?	...	Ip b_3?
...	...	...	...	...

→

Target
int. b_1
int. b_2
int. b_3
...

Peptide length

Variable input length, requires creative feature engineering

ion	Feature 1	Feature 2	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	...	Target
A	Hydro 1	?	?	?	?	Basicity 1	?	...	int. b_1
A C	Hydro 1	Hydro 2	?	?	?	Basicity 1	Basicity 2	...	int. b_2
A C I	Hydro 1	Hydro 2	Hydro 3	?	?	Basicity 1	Basicity 2	...	int. b_3
A C I D	Hydro 1	Hydro 2	Hydro 3	Hydro 4	?	Basicity 1	Basicity 2	...	int. b_4
C I D S	?	Hydro 2	Hydro 3	Hydro 4	Hydro 5	?	Basicity 2	...	int. y_4
I D S	?	?	Hydro 3	Hydro 4	Hydro 5	?	?	...	int. y_3
D S	?	?	?	Hydro 4	Hydro 5	?	?	...	int. y_2
S	?	?	?	?	Hydro 5	?	?	...	int. y_1

Variable input length, requires creative feature engineering

Hydrophobicity
Basicity
Iso-electric point
Helicity

✖

Full peptide
b-ion
y-ion

✖

Minimum
Q1
Q2
Q3
Maximum

✚

Hydrophobicity
Basicity
Iso-electric point
Helicity

✖

N-term
Fragmentation site - 1
Fragmentation site
Fragmentation site + 1
Fragmentation site + 2
C-term

✚

Peptide length
Peptide charge

Variable input length, requires creative feature engineering

ion	Charge	Length	Hydro min	Hydro Q1	Hydro Q2	Hydro Q3	Hydro max	...	Target
A	2	5	37	37	37	37	37	...	int. b_1
A C	2	5	72	35	35	35	35	...	int. b_2
A C I	2	5	153	35	35	37	37	...	int. b_3
A C I D	2	5	212	35	35	37	59	...	int. b_4
C I D S	2	5	224	35	49	59	81	...	int. y_4
I D S	2	5	189	49	49	59	81	...	int. y_3
D S	2	5	108	49	49	59	59	...	int. y_2
S	2	5	49	49	49	49	49	...	int. y_1

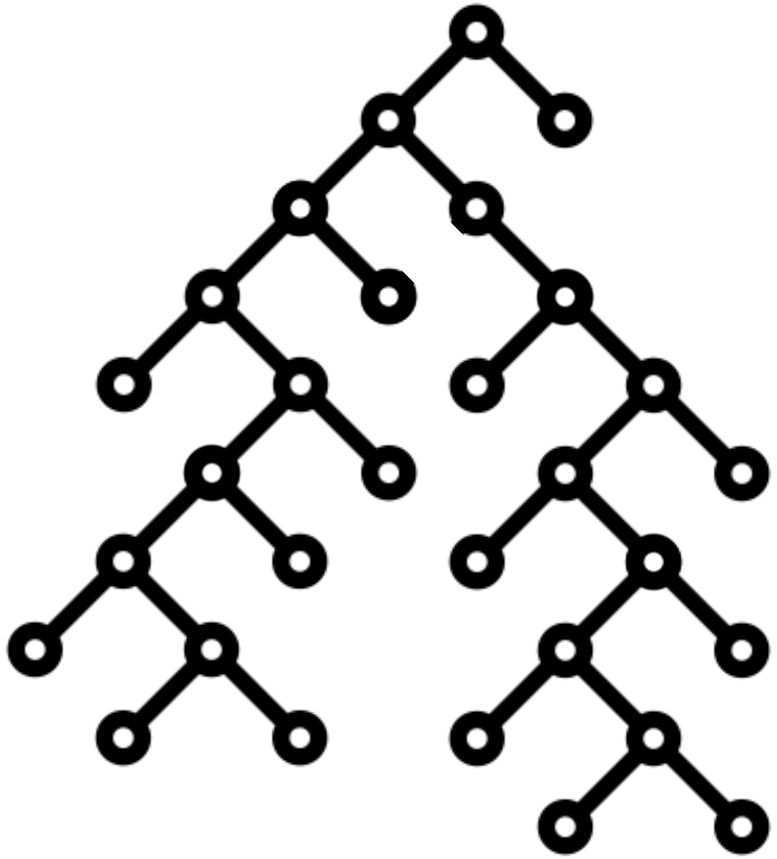
Given their shared fragmentation event, we can combine b- and y-ion features

Ions (b and y)	Charge	Length	Hydro b min	Hydro b Q1	…	Hydro y min	…
A C I D S	2	5	37	37	…	224	…
A C I D S	2	5	72	35	…	189	…
A C I D S	2	5	153	35	…	108	…
A C I D S	2	5	212	35	…	49	…

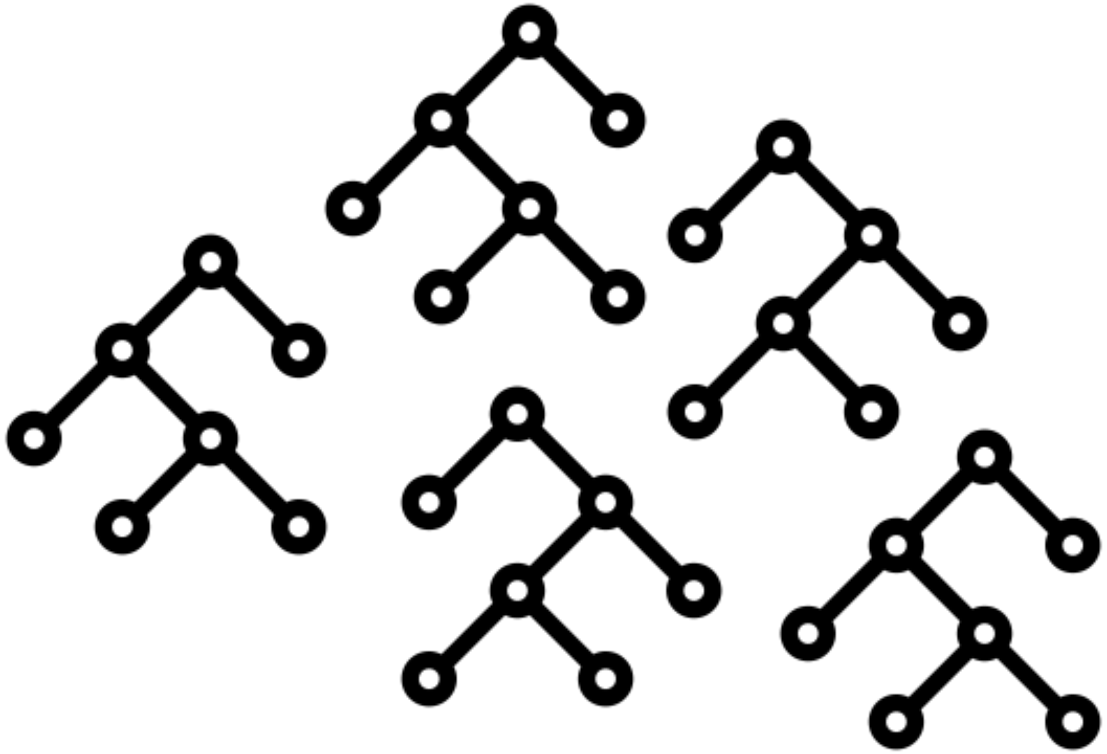
Target b	Target y
int. b_1	int. y_4
int. b_2	int. y_3
int. b_3	int. y_2
int. b_4	int. y_1

MS²PIP employs XGBoost, an ensemble decision tree algorithm

Decision tree

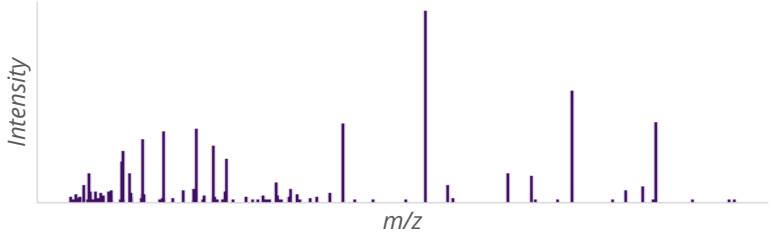


Ensemble of (weak learner) decision trees

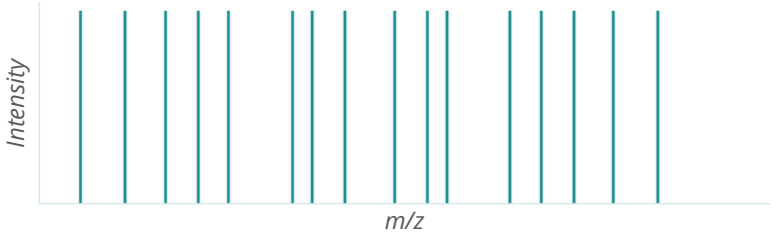


The result is a predicted spectrum that is a much better resemblance of a real spectrum

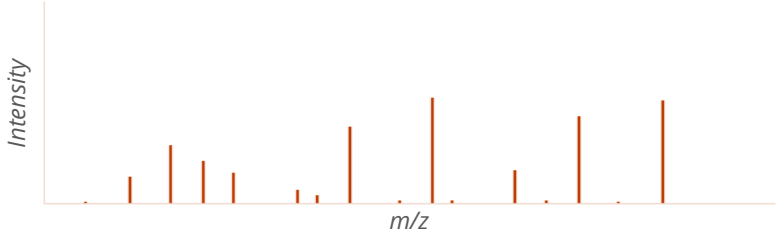
Empirical spectrum



Theoretical spectrum



MS²PIP-predicted spectrum



MS²PIP is available on iomics.ugent.be/ms2pip

MS²PIP SERVER

HOW TO RUN MS²PIP CONTACT

MS²PIP SERVER

MS² Peak Intensity Prediction

MS²PIP is a tool to predict MS² signal peak intensities from peptide sequences. It employs the XGBoost machine learning algorithm and is written in Python.

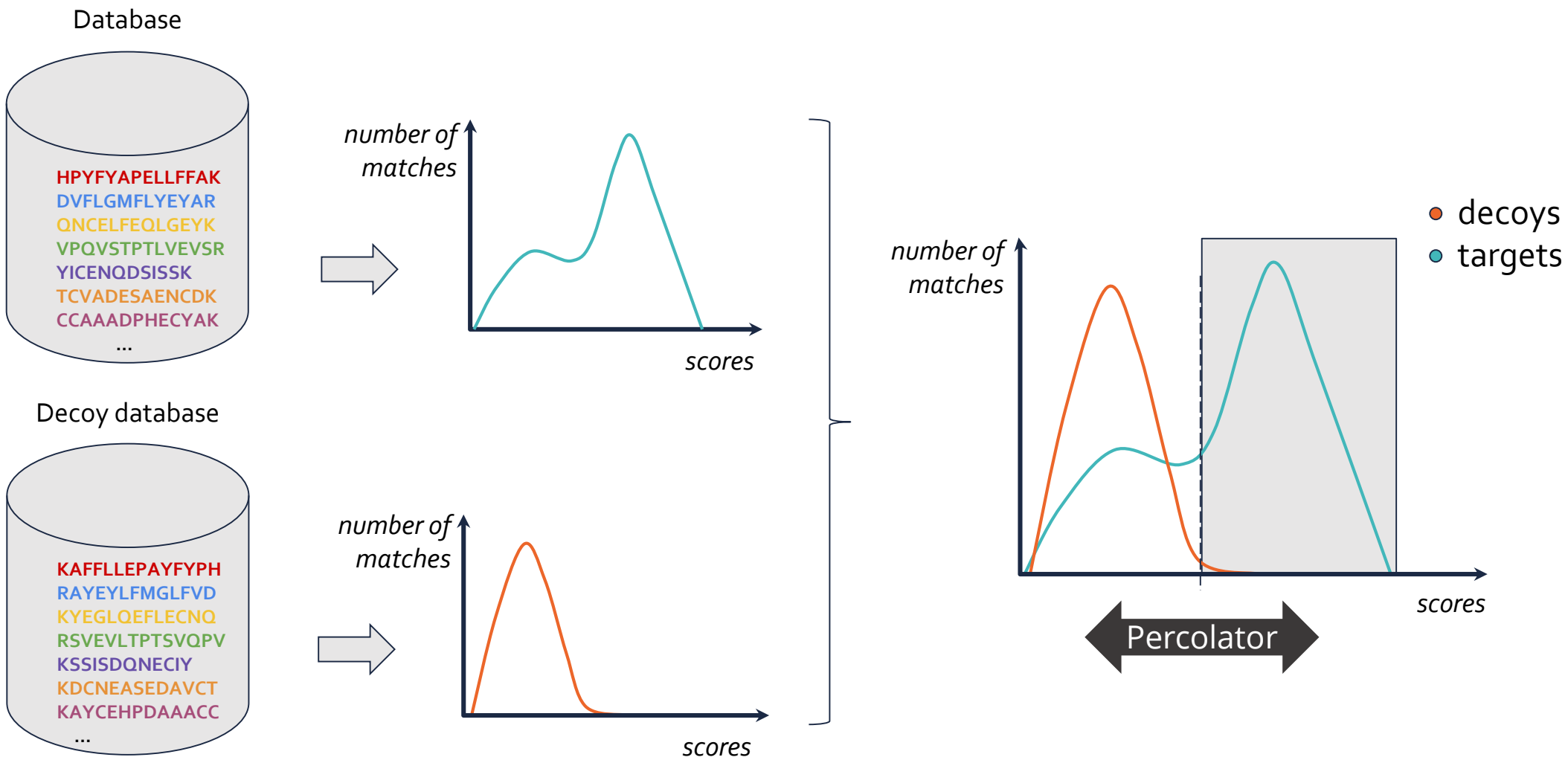
You can install MS²PIP on your machine by following our extended install instructions found on the MS²PIP GitHub repository. For a more user friendly experience, we created this web server. Below, you can easily upload a list of peptide sequences, after which the corresponding predicted MS² spectra can be downloaded in a CSV or MGF file format.

More advanced users can also access MS²PIP Server through our RESTful API. Swagger-generated documentation can be found here and an example Python script to access the API can be found here.

If you use MS²PIP for your research, please cite the following papers:

- Degroeve, S., Maddelein, D., & Martens, L. (2015). MS²PIP prediction server: compute and visualize MS2 peak intensity predictions for CID and HCD fragmentation. *Nucleic Acids Research*, *43*(W1), W326–W330.

Percolator employs a semi-supervised learning to improve target-decoy separation



Percolator's input comes from search engine derived metrics

Search engine features

- Search engine score
- Mass error of the peaks (theoretical vs measured)
- % of matched peaks
- …

We can add or replace these features with information from MS²PIP

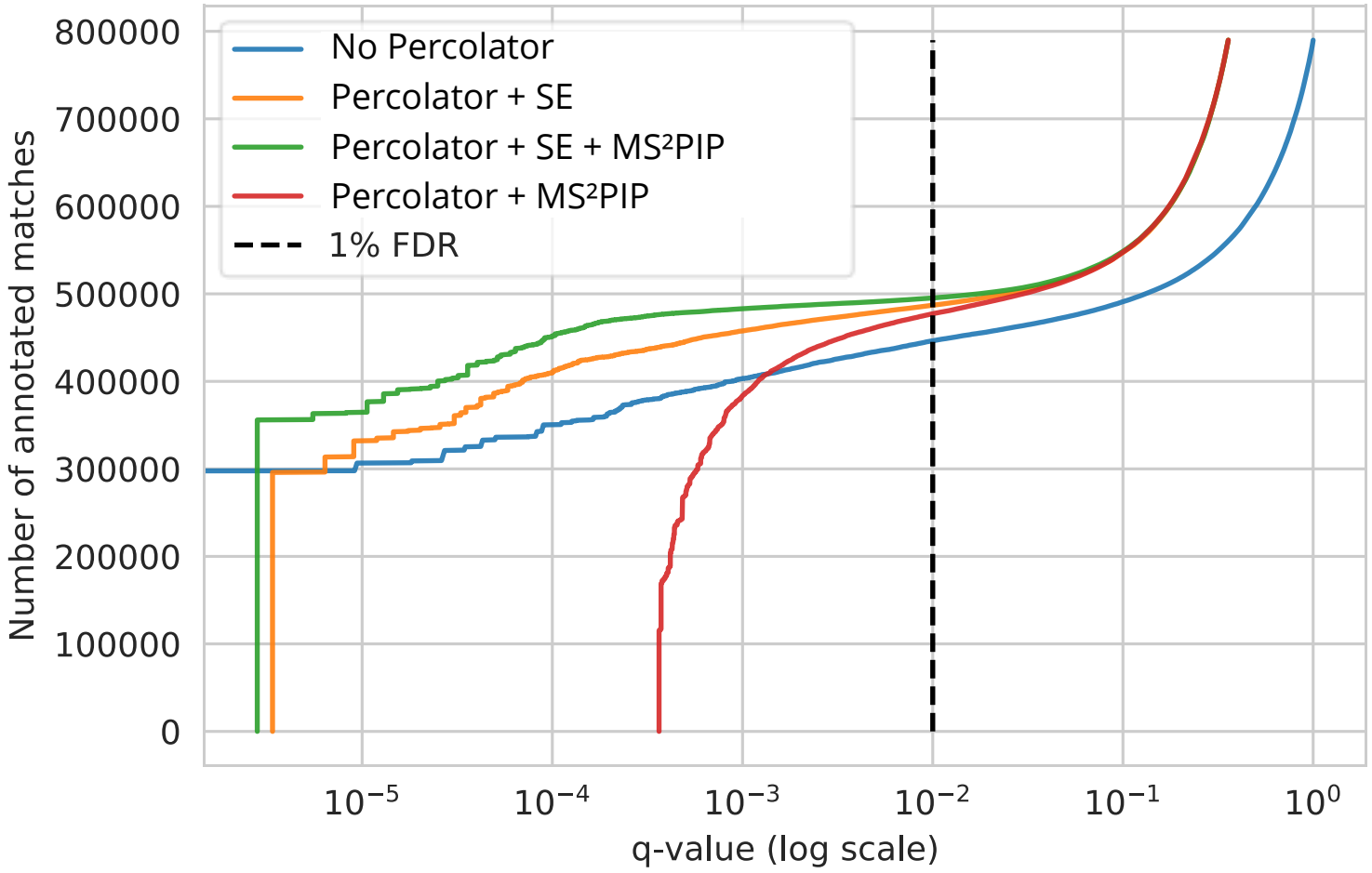
Search engine features

- Search engine score
- Mass error of the peaks (theoretical vs measured)
- % of matched peaks
- …

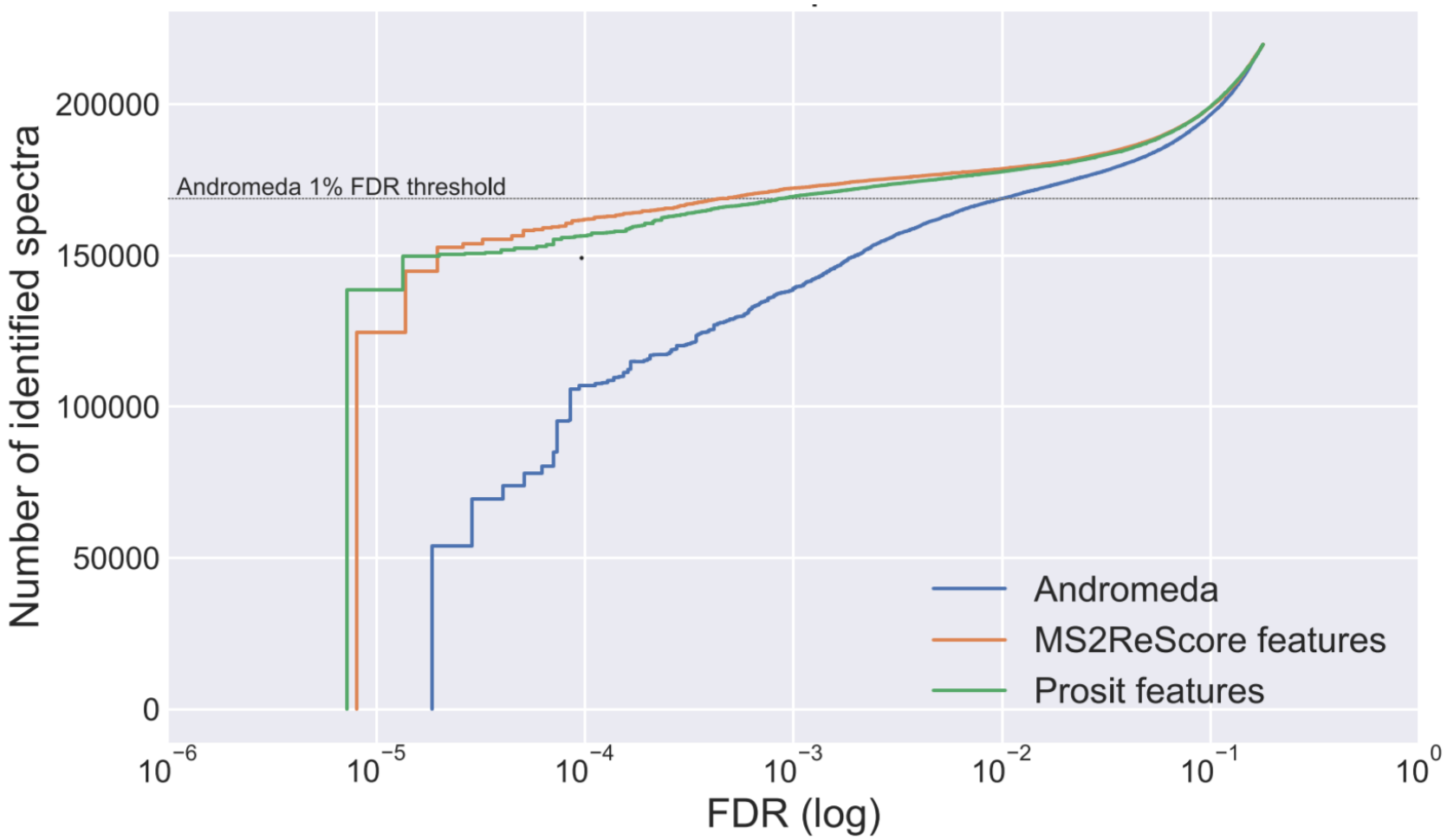
Correlation of MS²PIP prediction and measured spectrum

- Dot product
- Pearson correlation
- Spearman rank correlation
- Absolute differences
- …

MS²PIP + Percolator allows for more identifications at a more conservative FDR threshold



MS²PIP + Percolator allows for more identifications at a more conservative FDR threshold

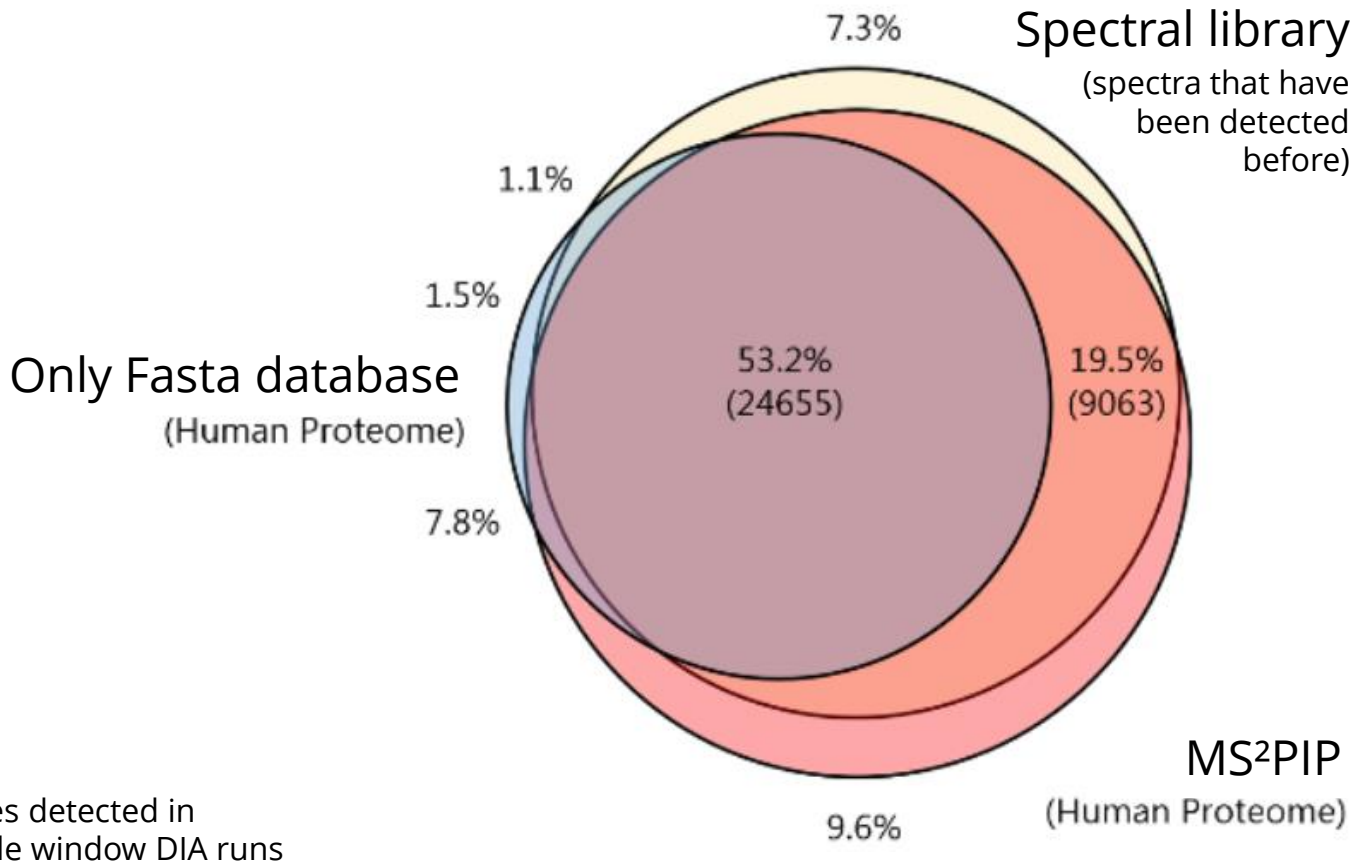


MS²PIP within a search engine enables sensitive open modification searches

ionbot

https://ionbot.cloud

MS²PIP can replace spectral libraries for Data-Independent acquisition (DIA)

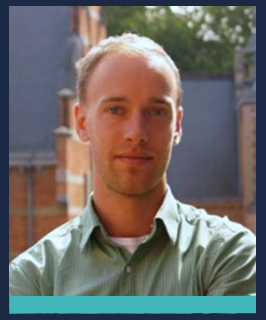


* Only peptides detected in triplicate wide window DIA runs with at least three transitions

Bart Van Puyvelde*, Sander Willems*, Ralf Gabriels* (2019) bioRxiv. doi:10.1101/681429

















VIB-UGENT CENTER FOR MEDICAL BIOTECHNOLOGY

GHENT UNIVERSITY

CompOmics

www.compomics.com
@CompOmics

Comp omics

VIB-UGENT CENTER FOR MEDICAL BIOTECHNOLOGY

UNIVERSITEIT GENT

@RalfGabriels

@CompOmics

www.compomics.com

References

- Sven Degroeve (2013). Bioinformatics. doi:10.1093/bioinformatics/btt544
- Ralf Gabriels (2019) Nucleic Acids Research doi:10.1093/nar/gkz299
- Bart Van Puyvelde*, Sander Willems*, Ralf Gabriels* (2019) bioRxiv. doi:10.1101/681429
- github.com/compomics/ms2rescore
- github.com/Biobix/proteoformer
- iomics.ugent.be/ms2pip
- ionbot.cloud