SCIENCE MEETS LIFE

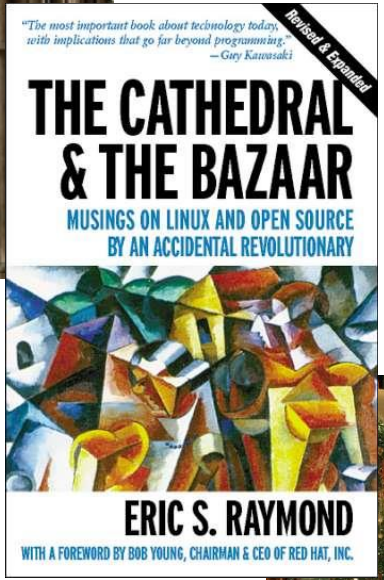
Data sharing and reuse

lennart martens @compomics

lennart.martens@ugent.be
computational omics and systems biology group
Ghent University and VIB, Ghent, Belgium

VIB GHENT UNIVERSITY Comp omics

Lincoln Cathedral, Lincolnshire UK



"The most important book about technology today, with implications that go far beyond programming."
—Guy Kawasaki

Revised & Expanded

THE CATHEDRAL & THE BAZAAR

MUSINGS ON LINUX AND OPEN SOURCE BY AN ACCIDENTAL REVOLUTIONARY

ERIC S. RAYMOND

WITH A FOREWORD BY BOB YOUNG, CHAIRMAN & CEO OF RED HAT, INC.

Grand Bazaar, Istanbul, Turkey

Why should we be re-using data?

Four types of (proteomics) data re-use

Unbiased proteome-wide (PT)M discovery as example

A subject of sociological study

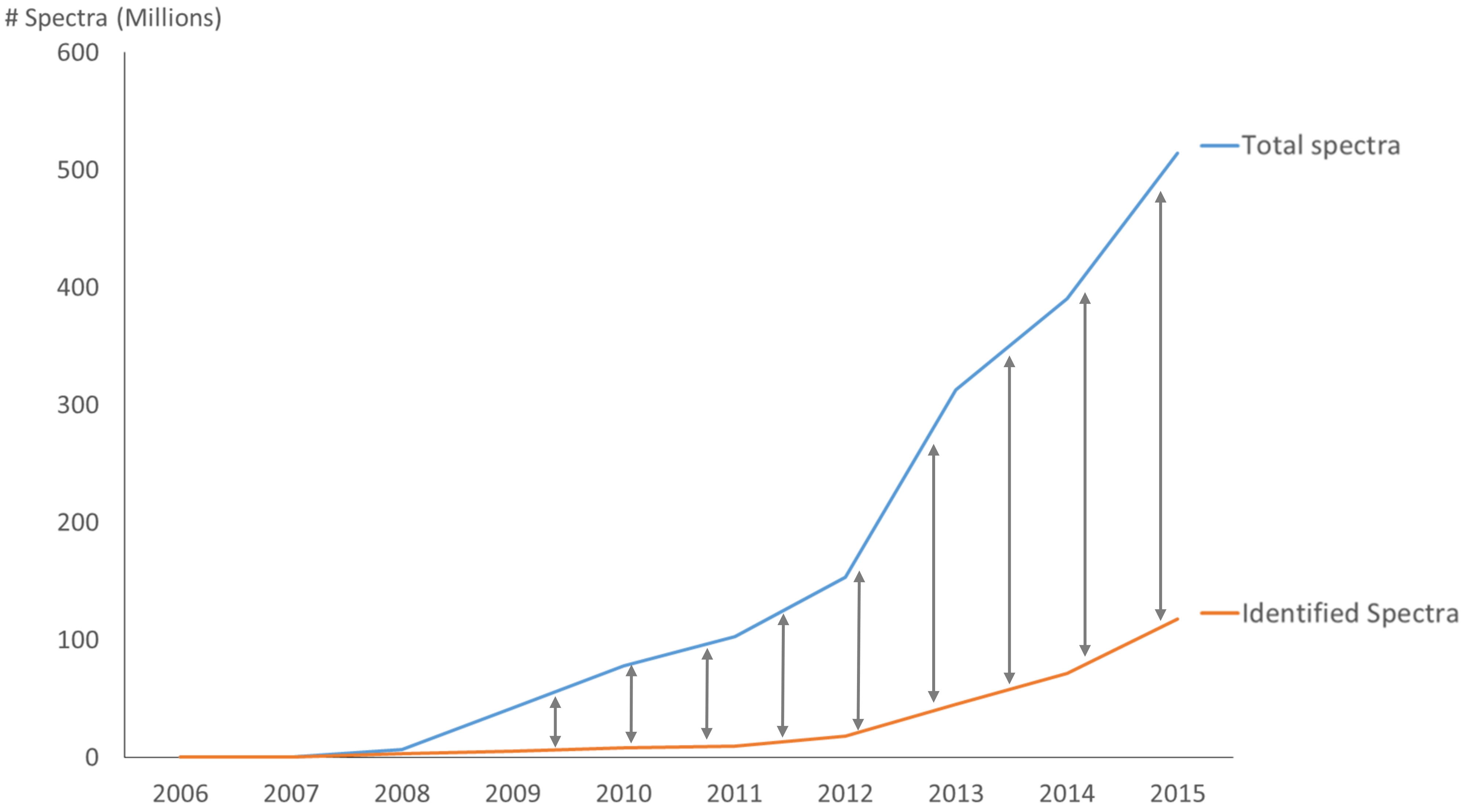
Why should we be re-using data?

Four types of (proteomics) data re-use

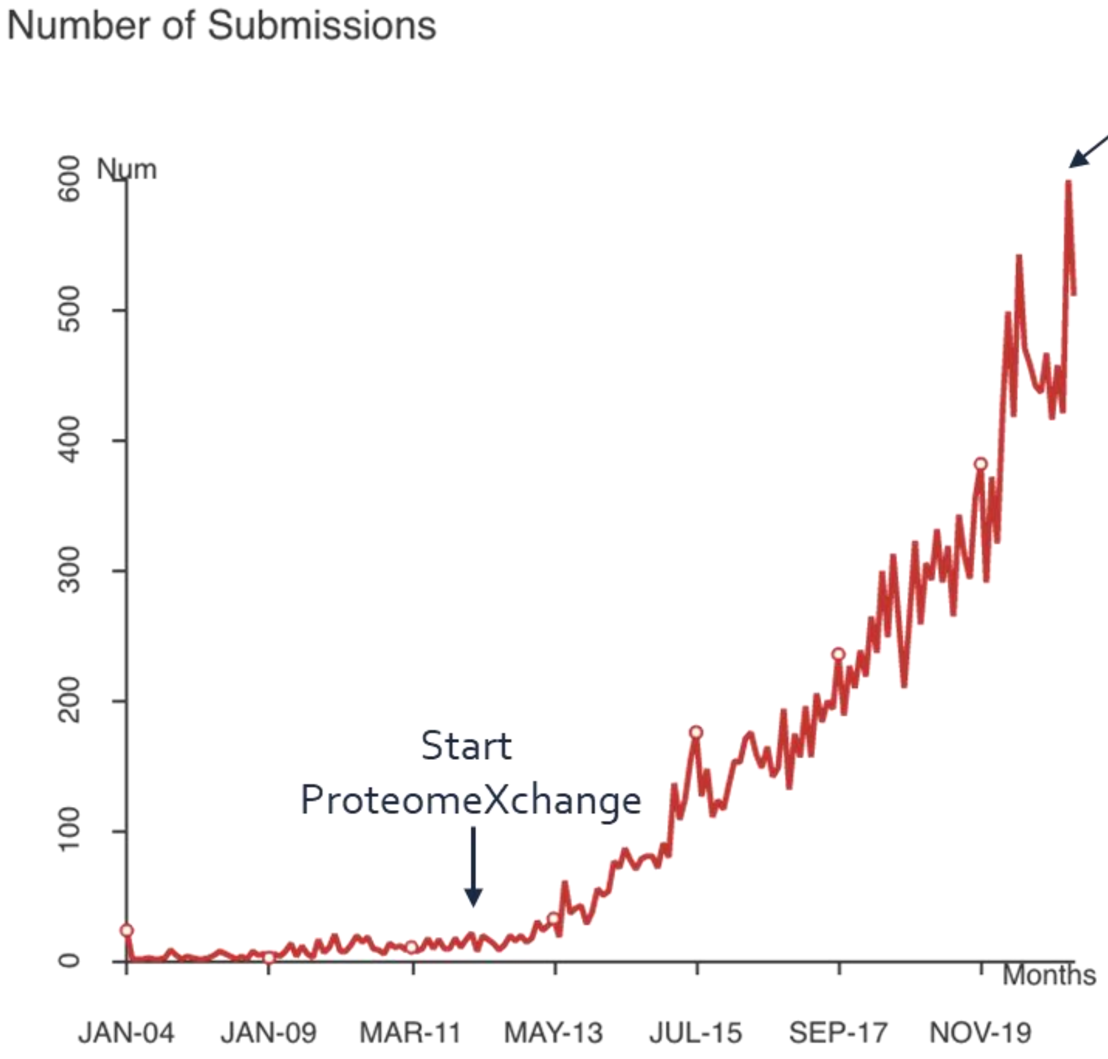
Unbiased proteome-wide (PT)M discovery as example

A subject of sociological study

Mass spectrometry data is high-content, meaning that much more data is acquired than is used in most papers



Mass spectrometry is also high throughput, meaning there is lots of data available!



March 2021 (596 datasets) – PXD025000

Dec 2019-Dec 2020

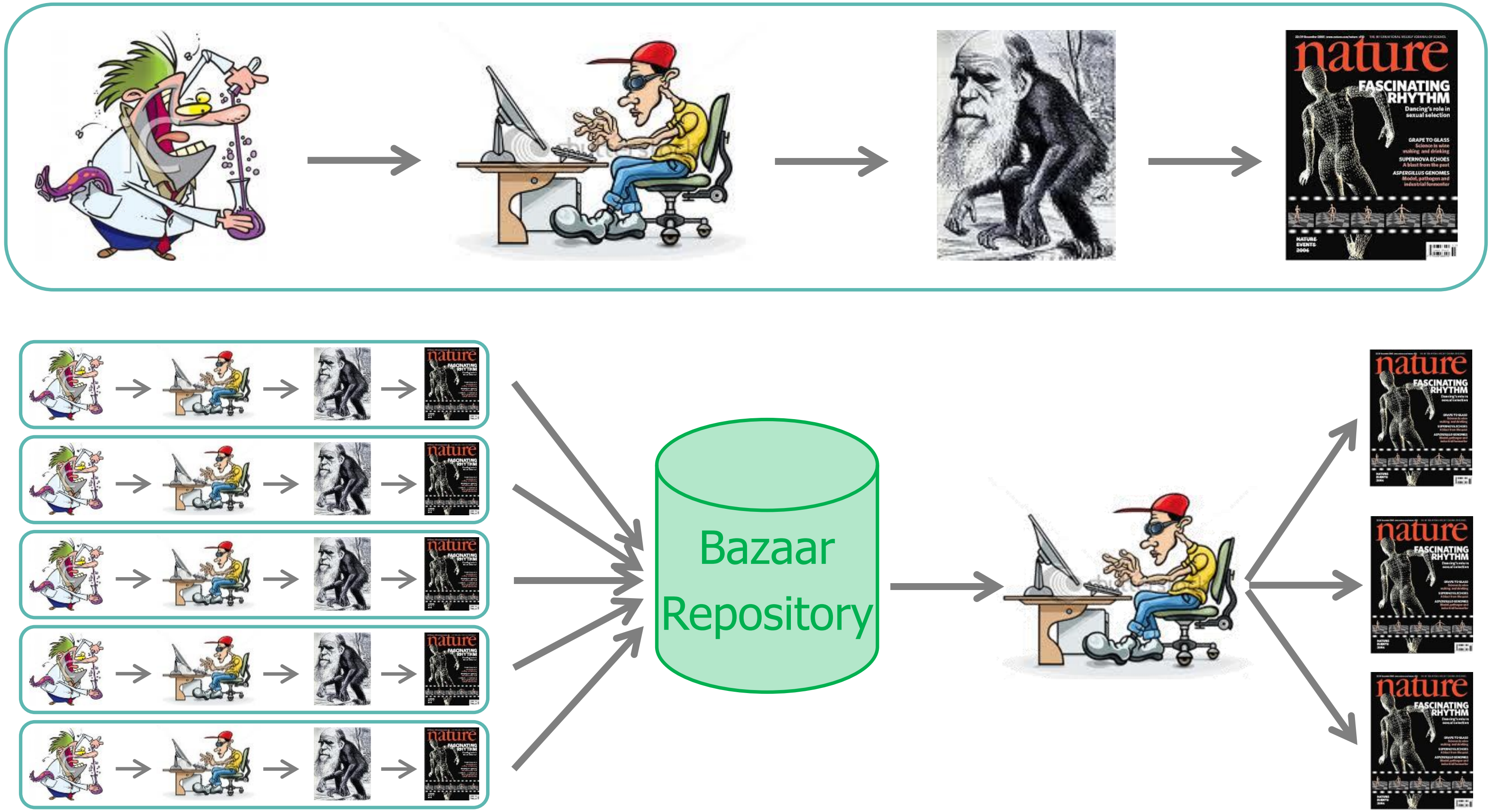
(+40% in one year)

An average of ~**450 datasets/month** during 2020

~5,300 datasets in 2020

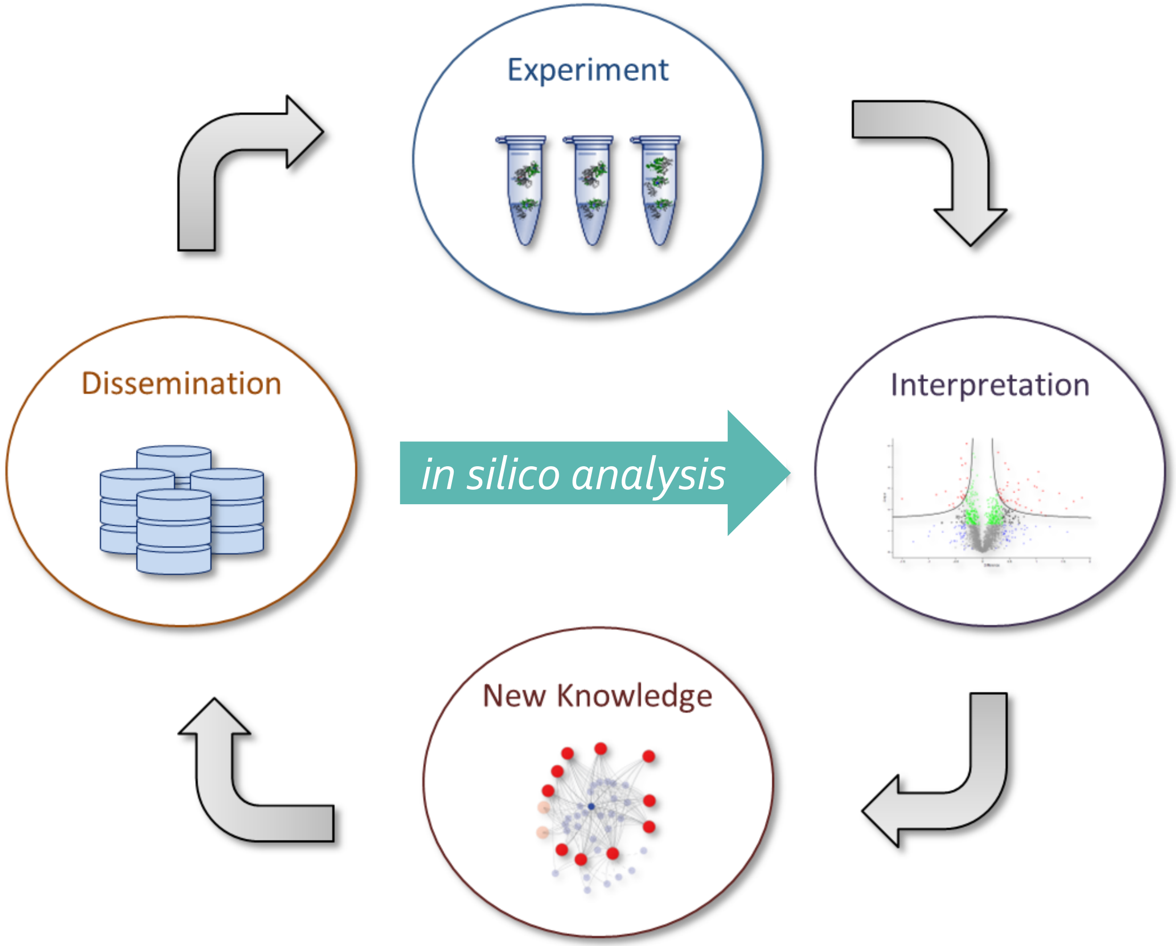
Slide courtesy of Dr. Juan Antonio Vizcaíno, Proteomics Team Leader, EMBL-EBI

As the volume and content of data increases in a field, the role of bioinformatics in that field changes as well



Bazaar Repository

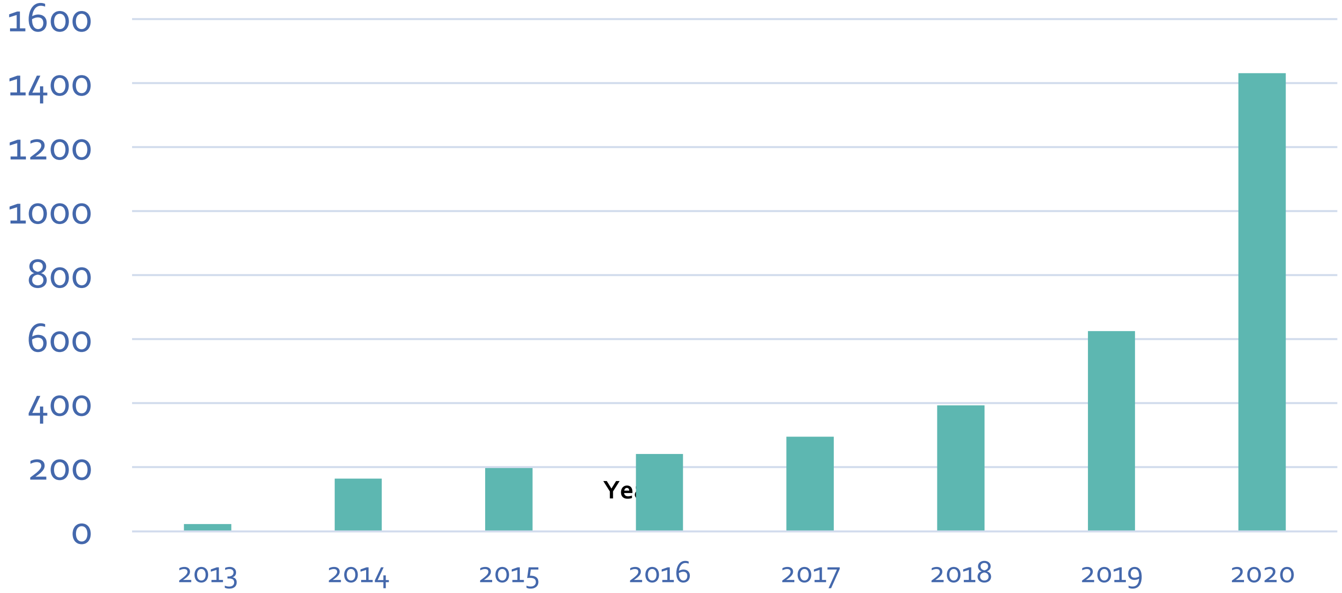
The data life cycle shows how *in silico* re-use of data fits in with the overall flow of information

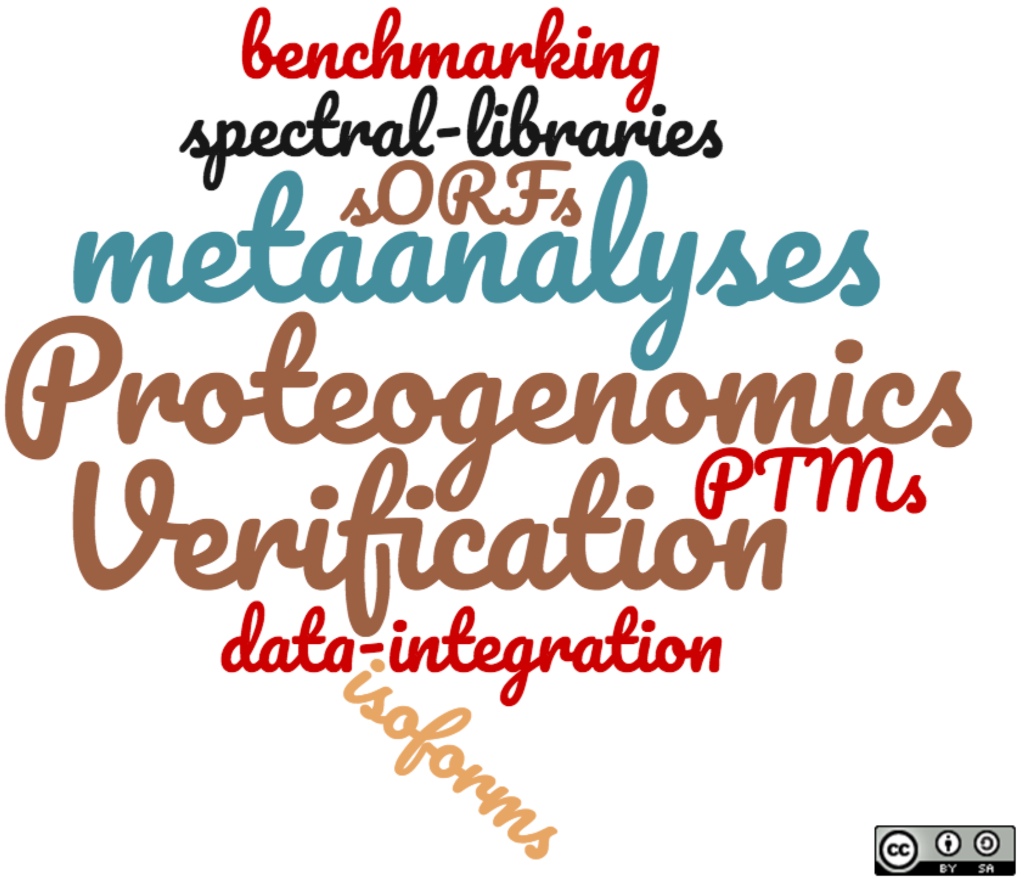


Vaudel, Proteomics, 2016

And this is seen in practice in proteomics as data is increasingly re-used

Volume of PRIDE Data Downloads (TB), 2013-2020

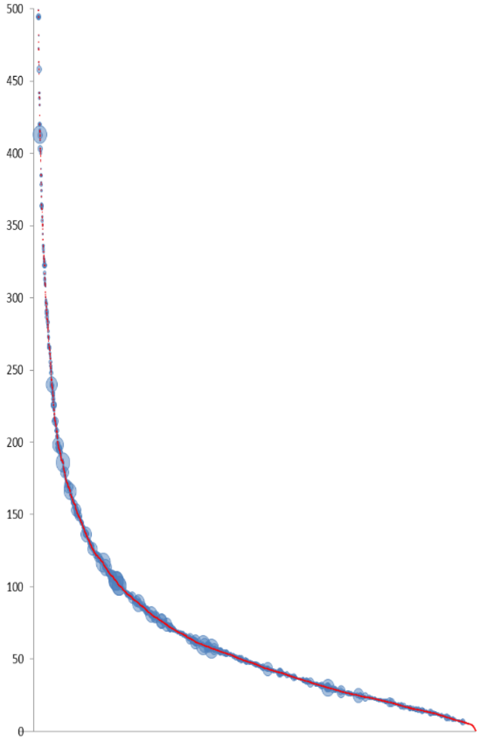




Slide courtesy of Dr. Juan Antonio Vizcaíno, Proteomics Team Leader, EMBL-EBI

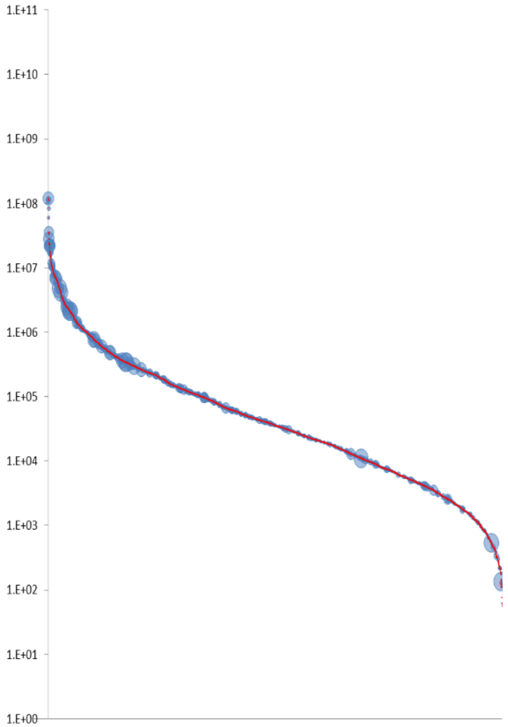
Large-scale data reprocessing can harness heterogeneity to dig very deep into the proteome

Half life (h)



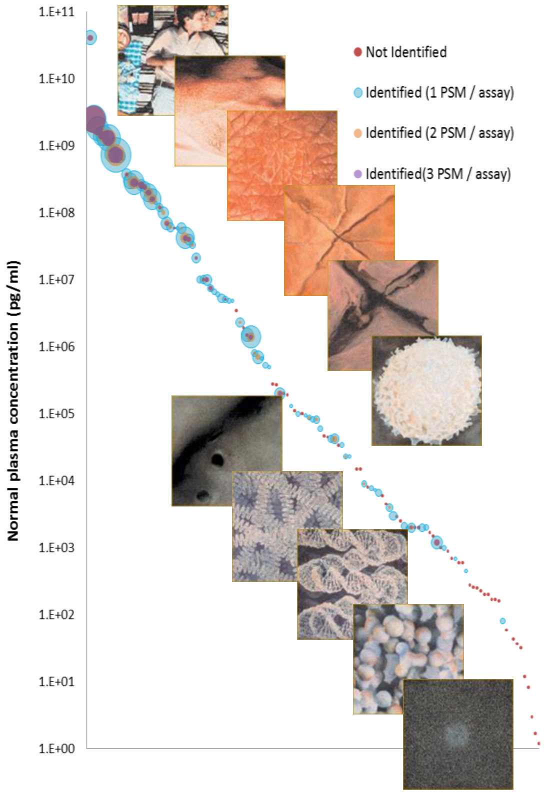
proteins

concentration (copies per cell)



proteins

concentration (pg/ml)



Verheggen, Journal of Proteome Research, 2017

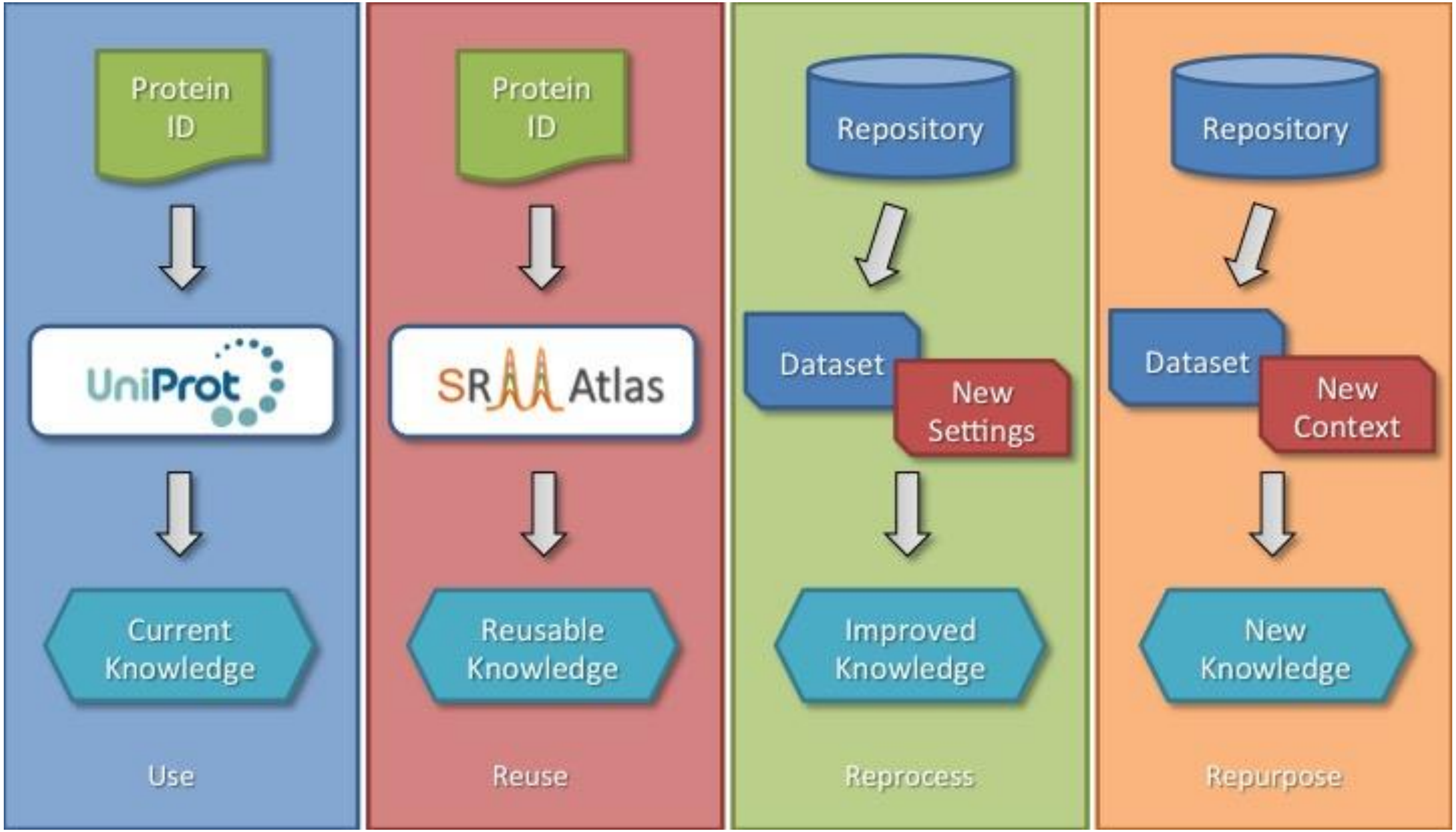
Why should we be re-using data?

Four types of (proteomics) data re-use

Unbiased proteome-wide (PT)M discovery as example

A subject of sociological study

In general, data re-use can take four distinct forms, all of which are somehow applied in our example



Vaudel, Proteomics, 2016

Why should we be re-using data?

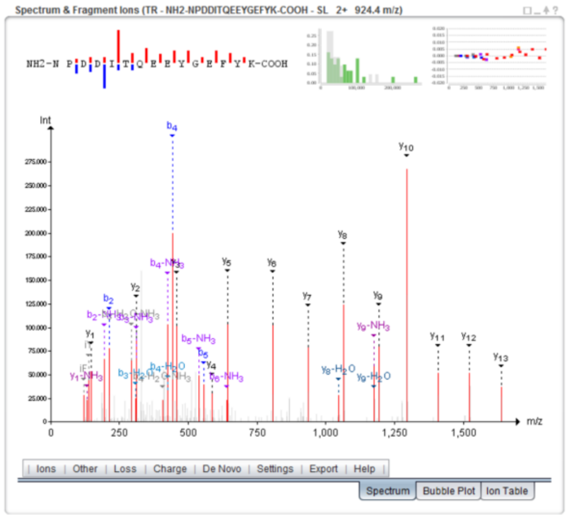
Four types of (proteomics) data re-use

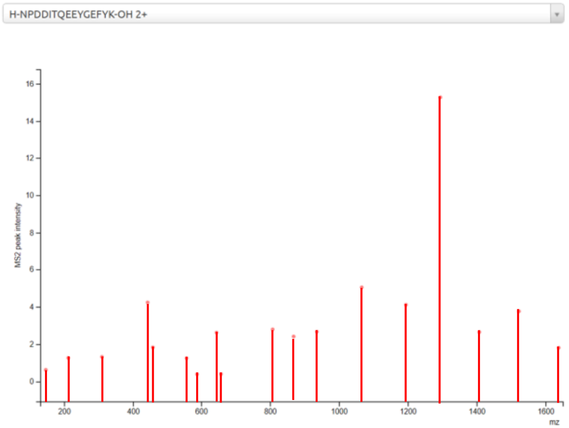
Unbiased proteome-wide (PT)M discovery as example

A subject of sociological study

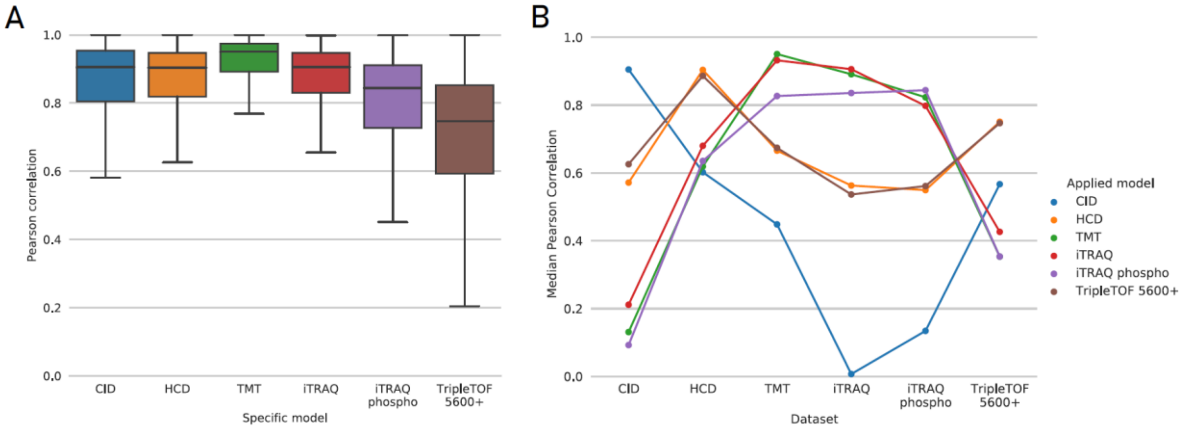
Our MS2PIP fragmentation model accurately predicts peptide behaviour in varying conditions

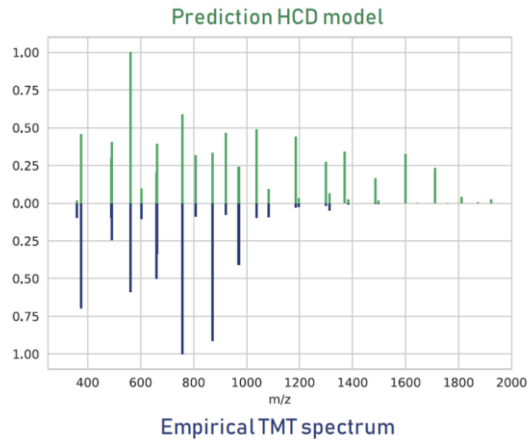
Vaudel, Nat. Biotech., 2015
PeptideShaker

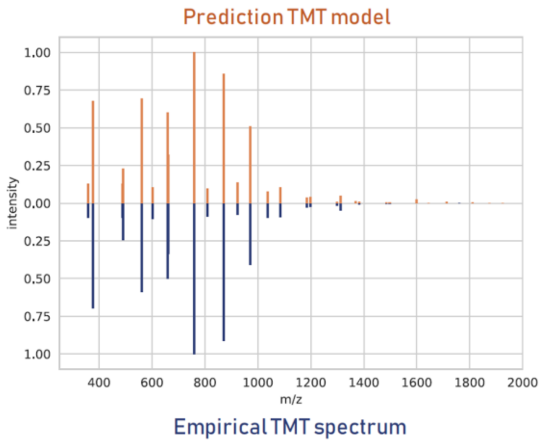

Spectrum & Fragment Ions (TR - NH2-NPDDITQEEYGEFYK-COOH - SL 2+ 924.4 m/z)
NH2-N P D D I T Q E E Y G E F Y K-COOH
Int
m/z
0
250
500
750
1,000
1,250
1,500
b4
y10
y8
y5
y6
y7
y9
y11
y12
y13
Ions | Other | Loss | Charge | De Novo | Settings | Export | Help
Spectrum
Bubble Plot
Ion Table



https://iomics.ugent.be/ms2pip
Degroeve, Bioinformatics, 2013
Degroeve, Nucleic Acids Research, 2015

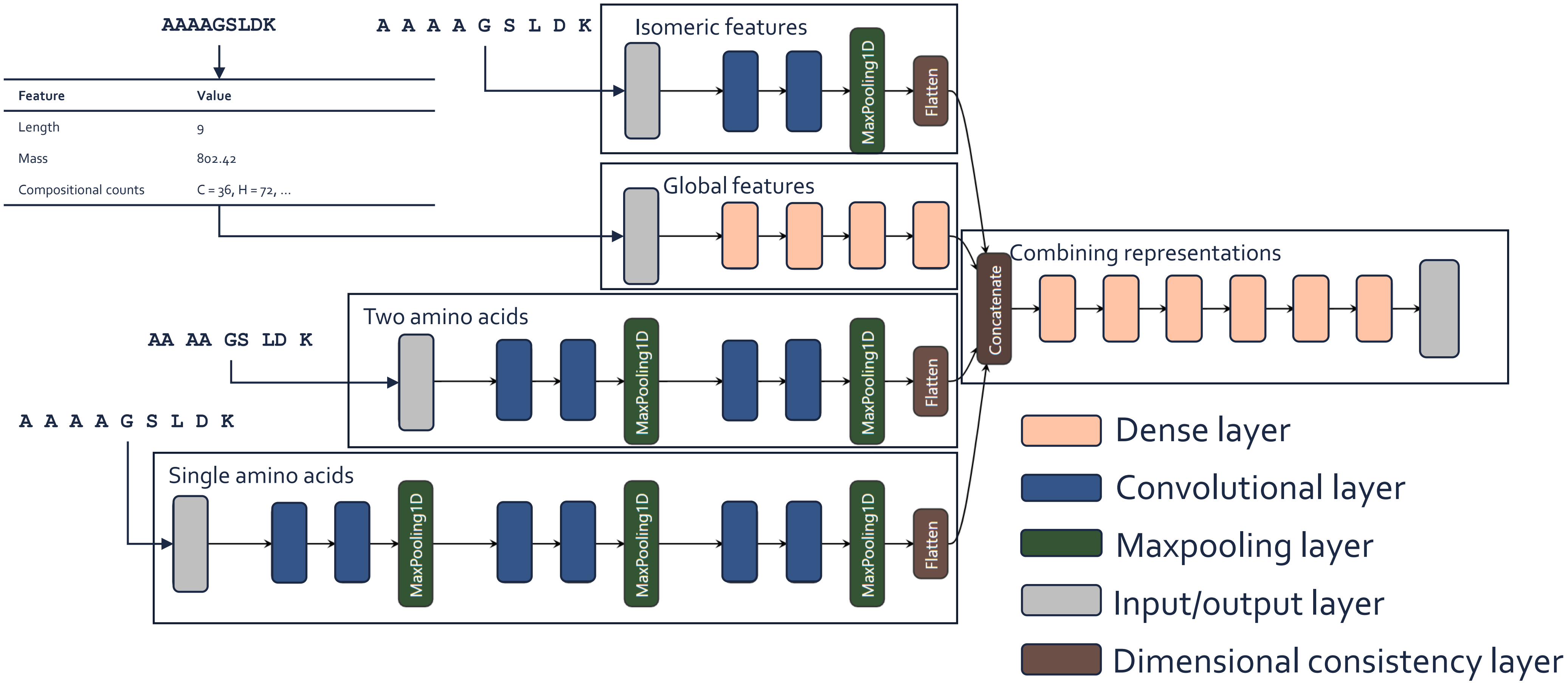






Gabriels, Nucleic Acids Research, 2019

DeepLC is a retention time predictor that can accurately predict retention times of as-yet unseen modifications

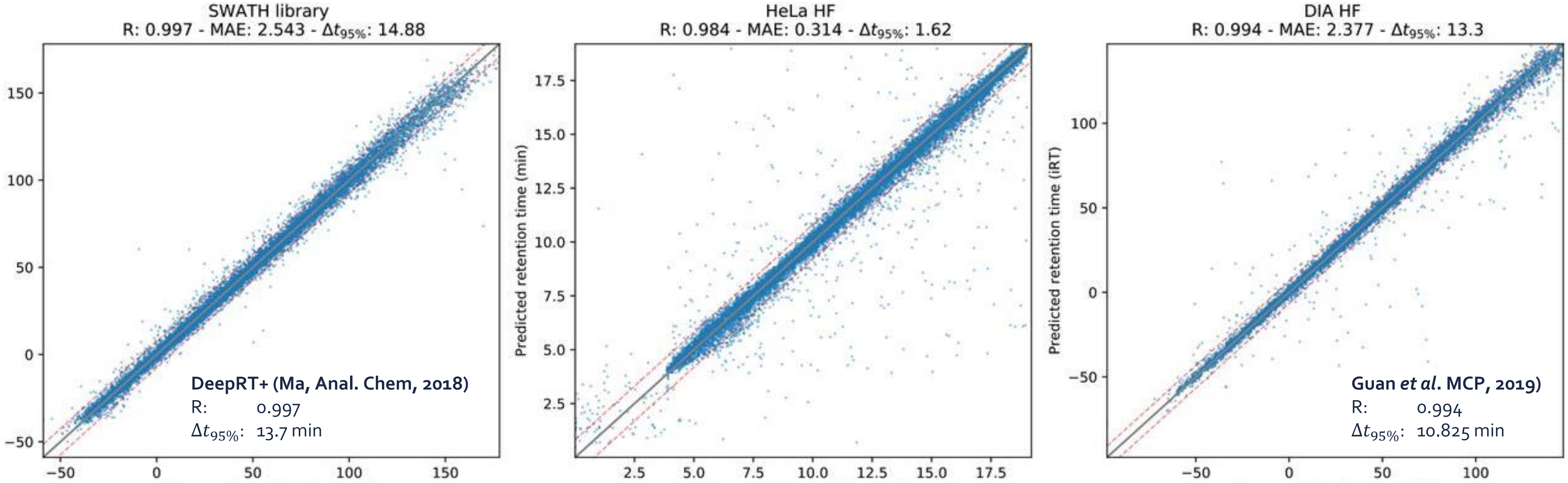


Feature	Value
Length	9
Mass	802.42
Compositional counts	C = 36, H = 72, ...

Bouwmeester, Nature Methods, *accepted* - https://www.biorxiv.org/content/10.1101/2020.03.28.013003v2

The performance of DeepLC is on par with the state-of-the-art

Predicted retention time (min)



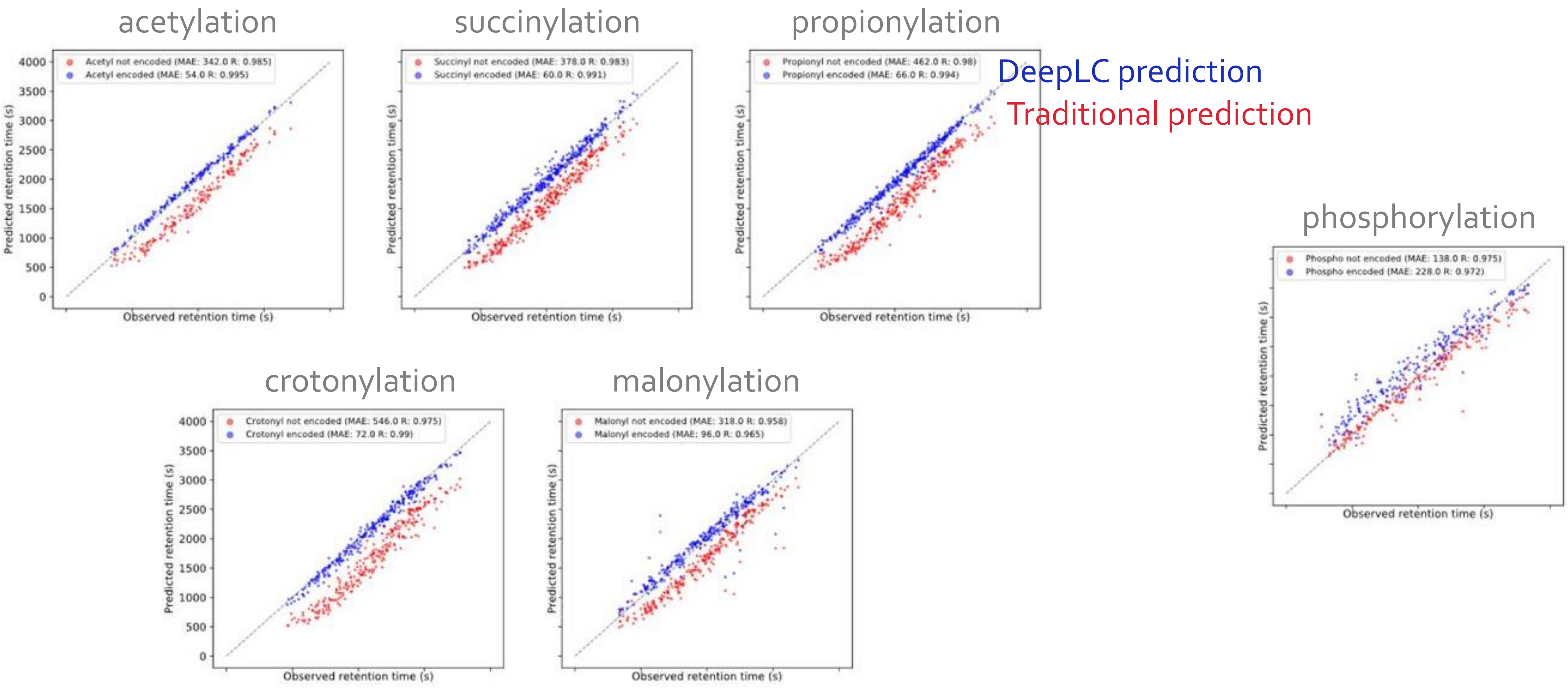
DeepRT+ (Ma, Anal. Chem, 2018)
R: 0.997
$\Delta t_{95\%}$: 13.7 min

Guan *et al*. MCP, 2019)
R: 0.994
$\Delta t_{95\%}$: 10.825 min

Observed retention time (min)

Dataset from Zolg, MCP, 2018
Bouwmeester, Nature Methods, *accepted* - https://www.biorxiv.org/content/10.1101/2020.03.28.013003v2

DeepLC can accurately predict t_R for many modifications, despite never having seen these before



acetylation

succinylation

propionylation

phosphorylation

crotonylation

malonylation

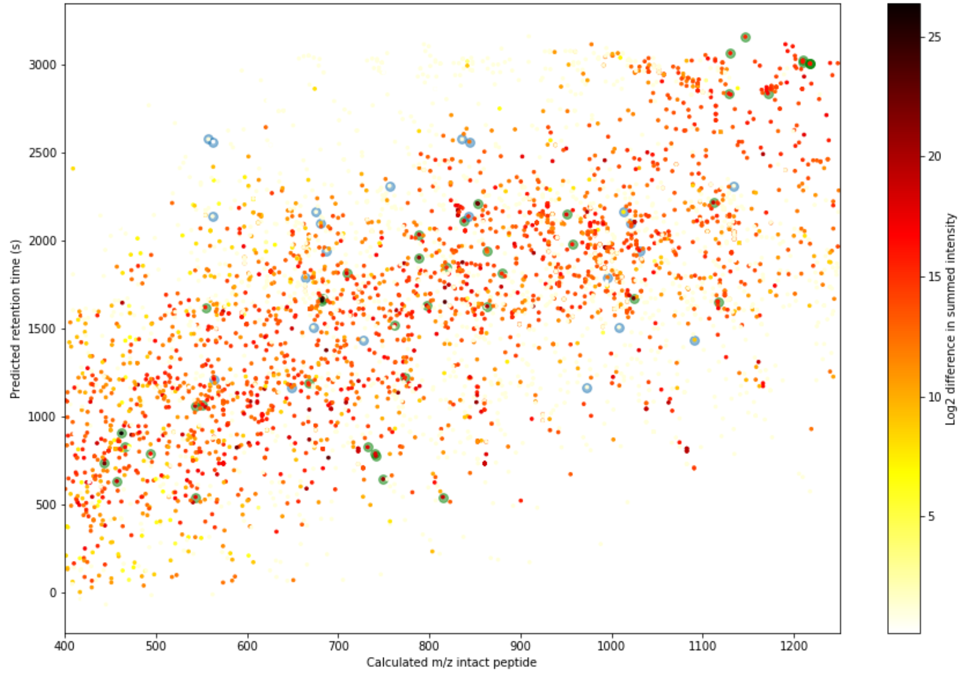
DeepLC prediction

Traditional prediction

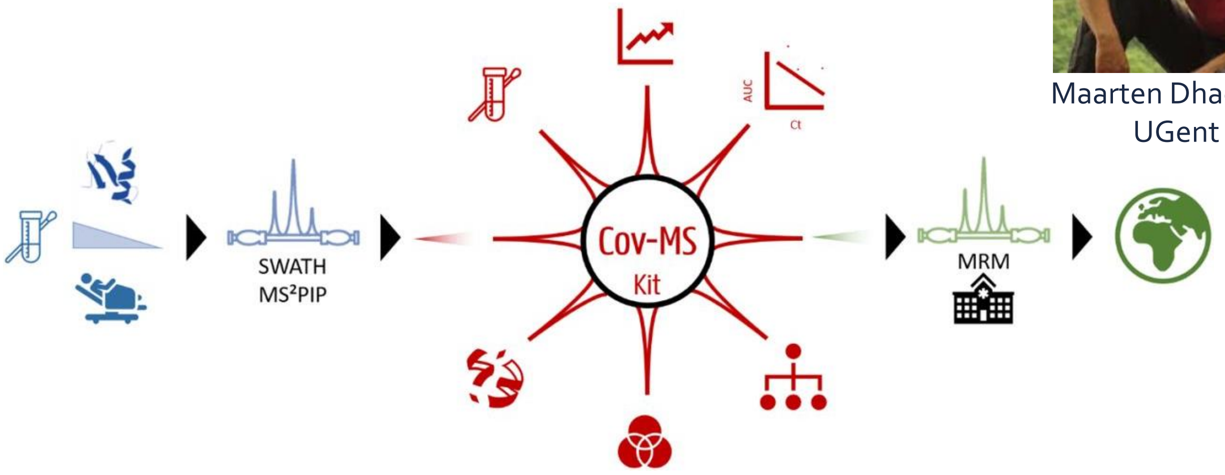
Dataset from Zolg, MCP, 2018

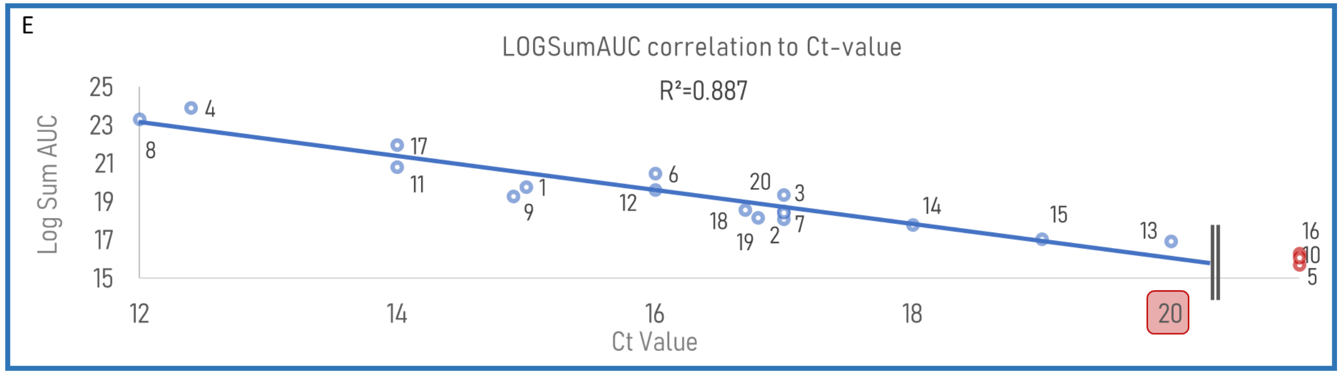
Bouwmeester, Nature Methods, *accepted* - https://www.biorxiv.org/content/10.1101/2020.03.28.013003v2

MS2PIP and DeepLC were crucial in the development of a targeted MS-based COVID-19 test that runs in 38 mins



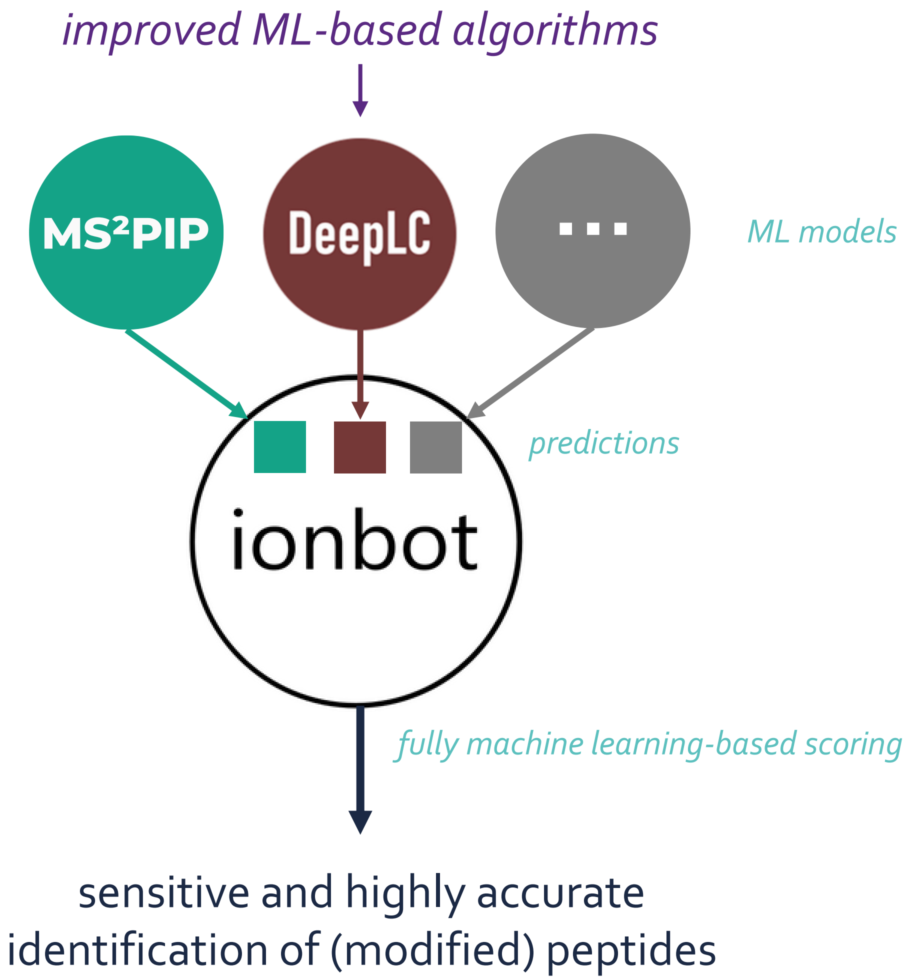
Maarten Dhaenens
UGent





Van Puyvelde, JACS Au, 2021

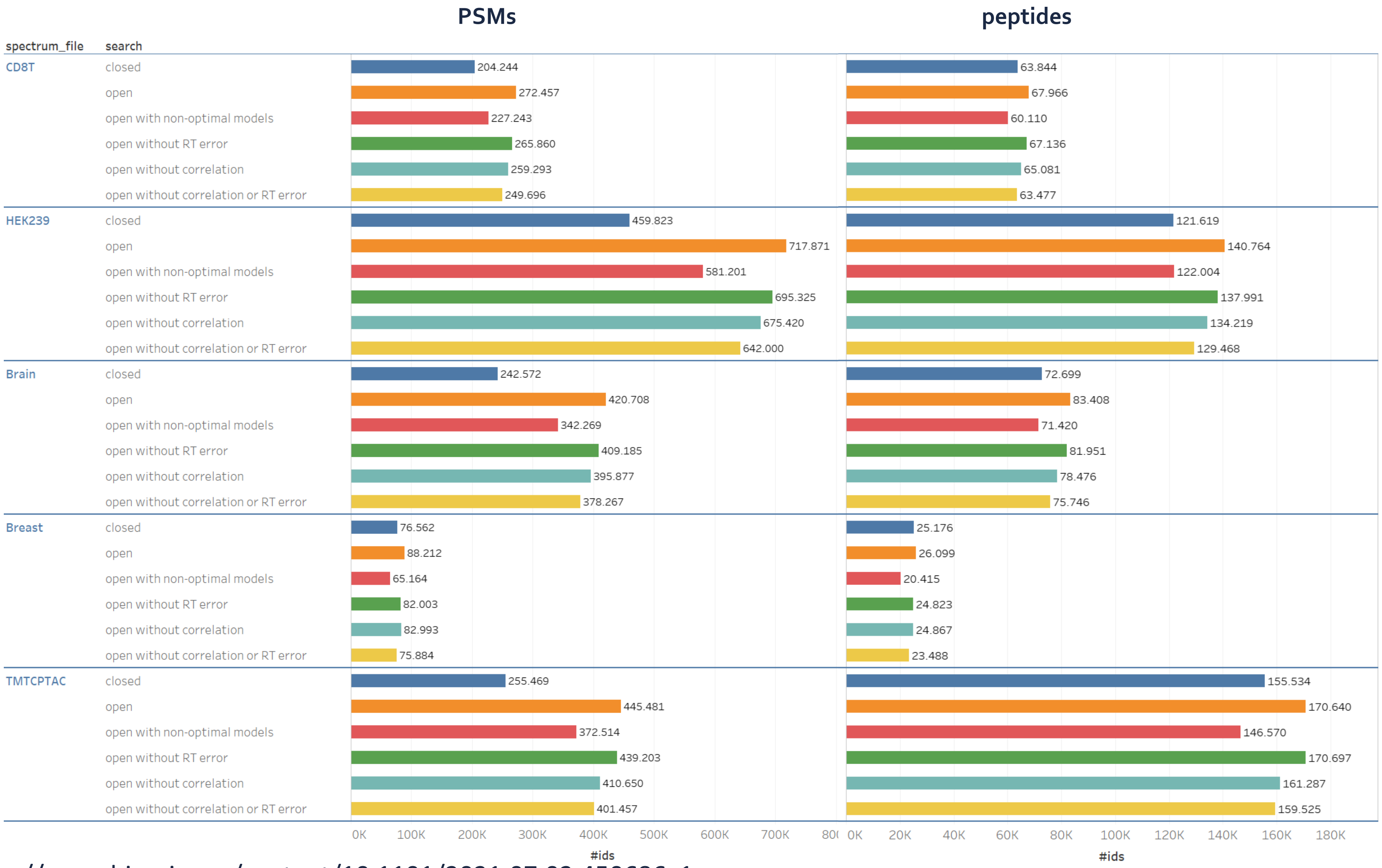
MS2PIP and DeepLC power ionbot, a novel and extensible open modification search engine with high reliability


improved ML-based algorithms
MS²PIP
DeepLC
...
ML models
predictions
ionbot
fully machine learning-based scoring

sensitive and highly accurate identification of (modified) peptides

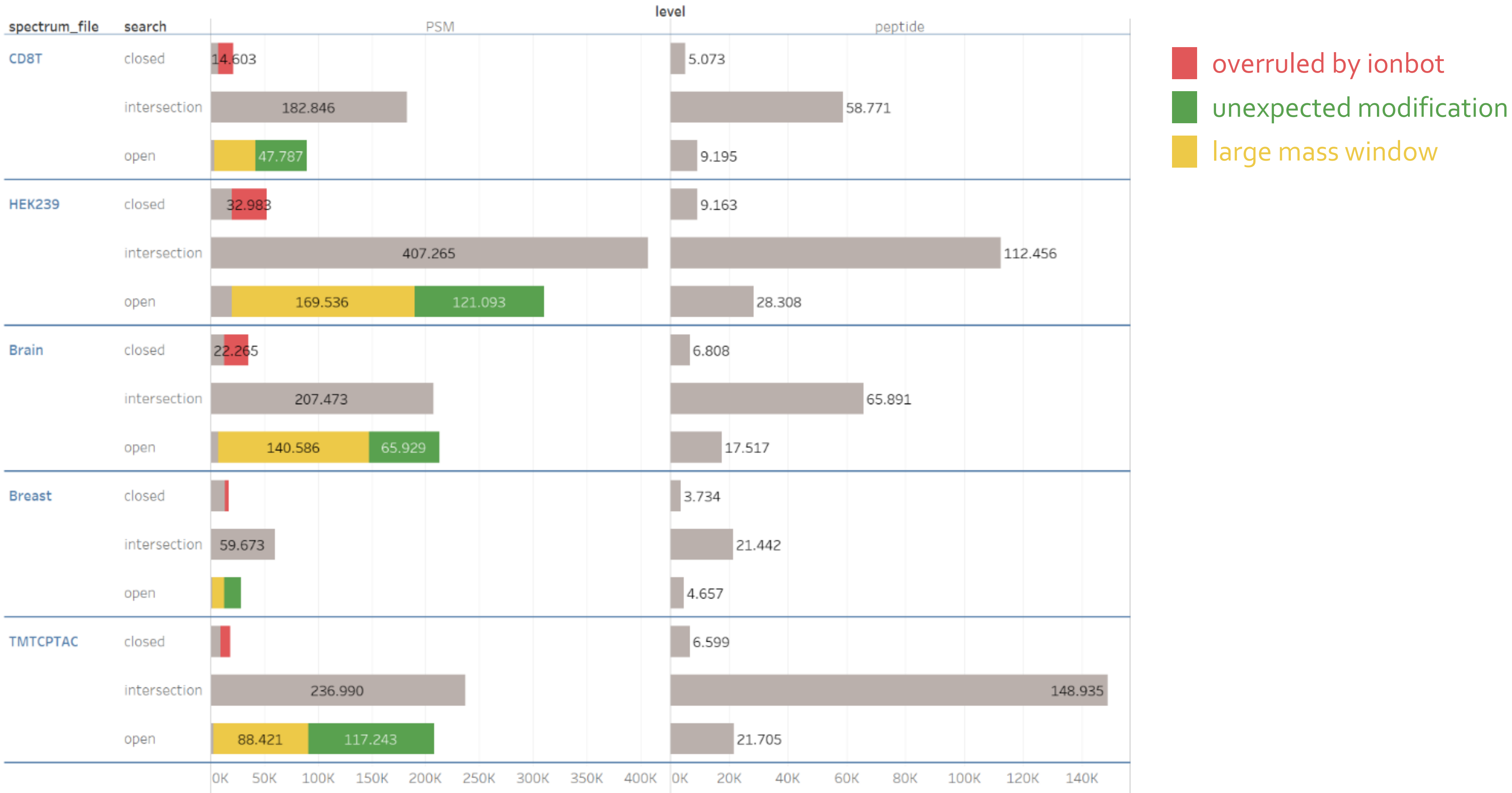
Degroeve, https://www.biorxiv.org/content/10.1101/2021.07.02.450686v1

Ionbot shows the value of open modification searches, as well as the value of accurate prediction models



Degroeve, https://www.biorxiv.org/content/10.1101/2021.07.02.450686v1

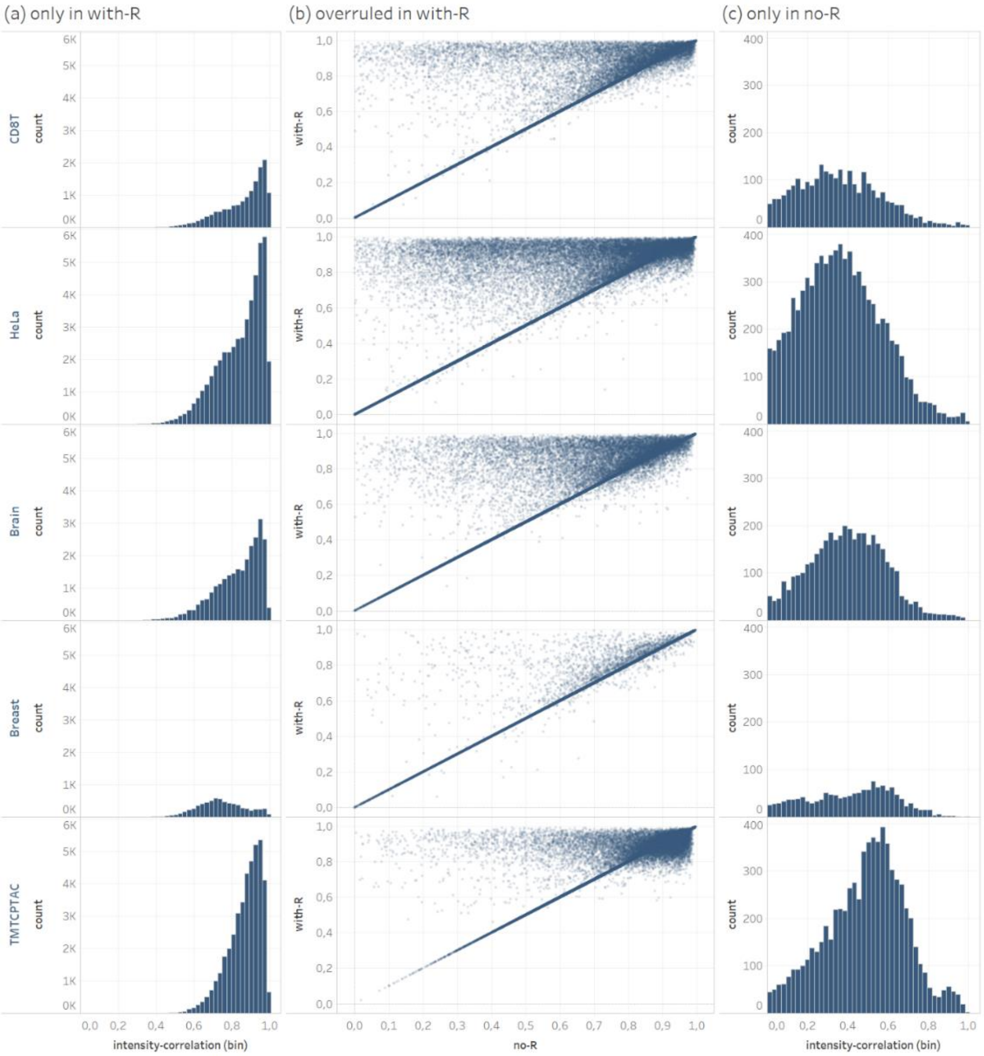
Interestingly, many identifications from the closed search are overruled by ionbot in the open search



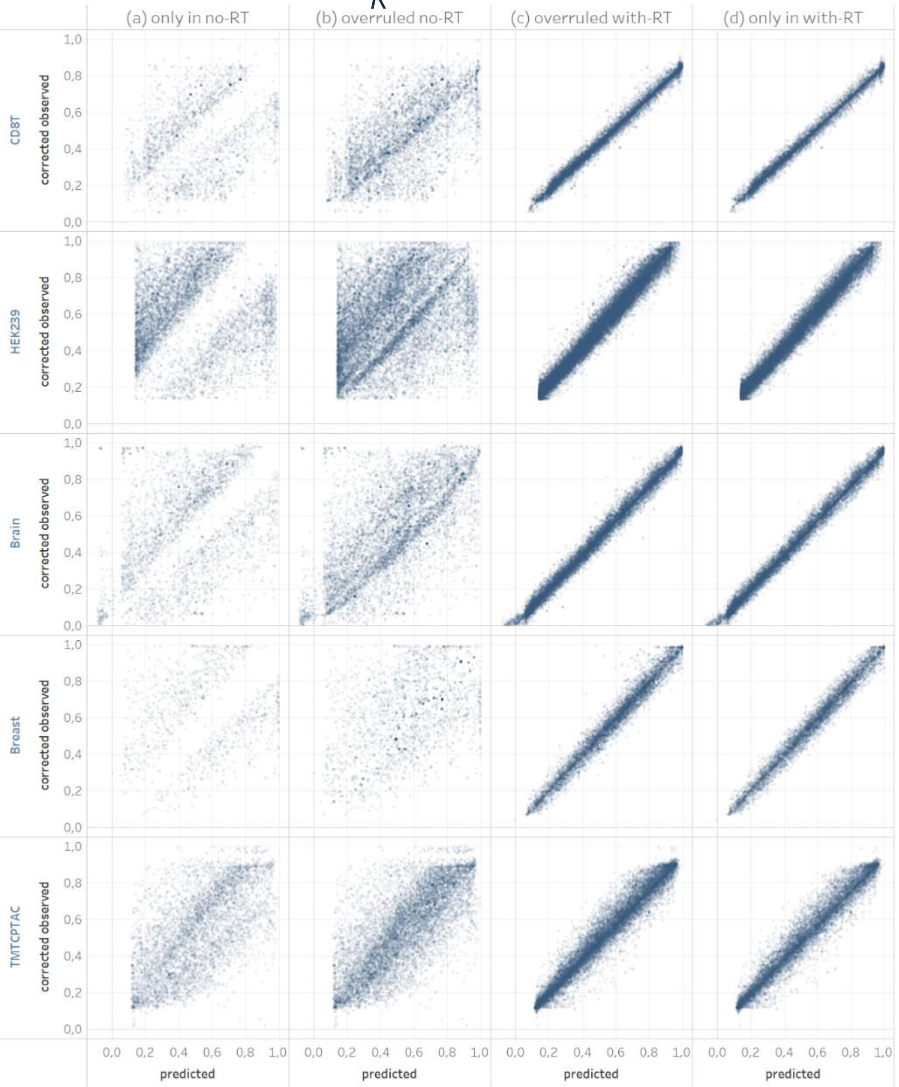
Degroeve, https://www.biorxiv.org/content/10.1101/2021.07.02.450686v1

Overruled identifications are better, as shown for results obtained with, and without predictions provided to ionbot

Intensity correlation

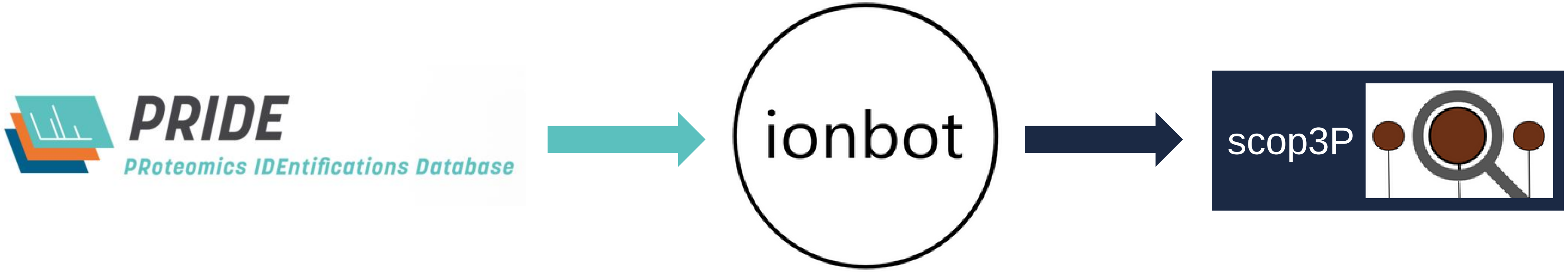


t_R correlation



Degroeve, https://www.biorxiv.org/content/10.1101/2021.07.02.450686v1

We reprocessed a large amount of phosphoproteomics data using ionbot, and made it available through Scop3P



www.ebi.ac.uk/pride

2016 RAW files
60.2 million spectra

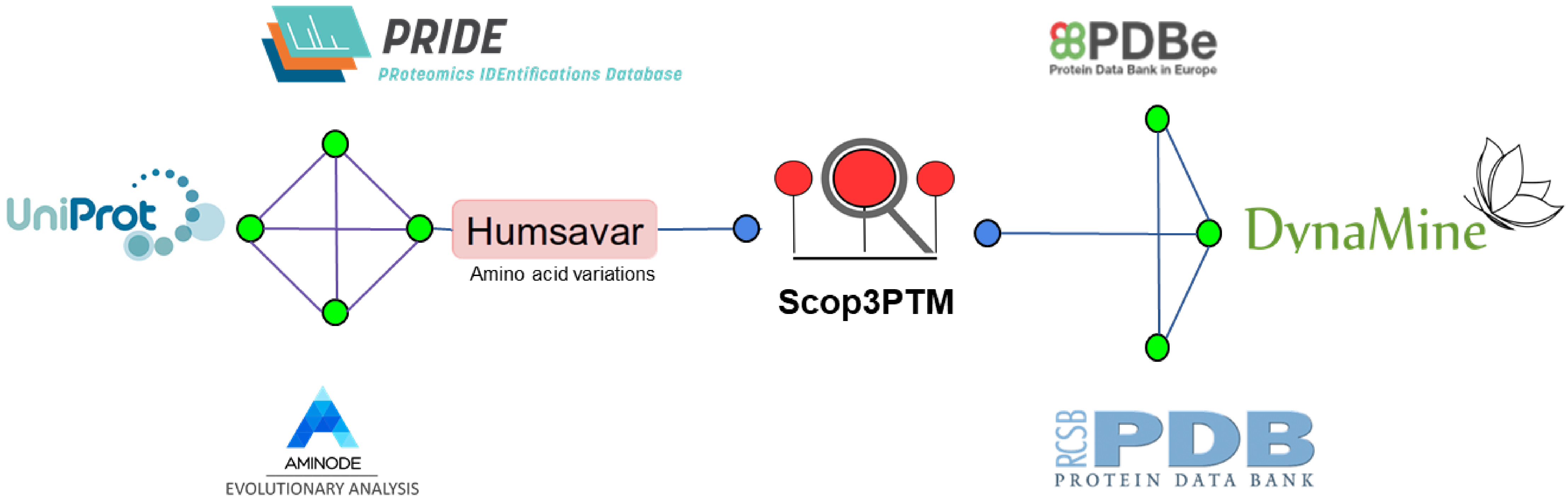
1490 UniMod modifications
all possible AA mutations

iomics.ugent.be/scop3p

19.2 million PSMs
139 048 P-sites
94 111 sites, PhosphoRS > 0.5
14 261 phosphoproteins

https://iomics.ugent.be/scop3p
Ramasamy, Journal of Proteome Research, 2020

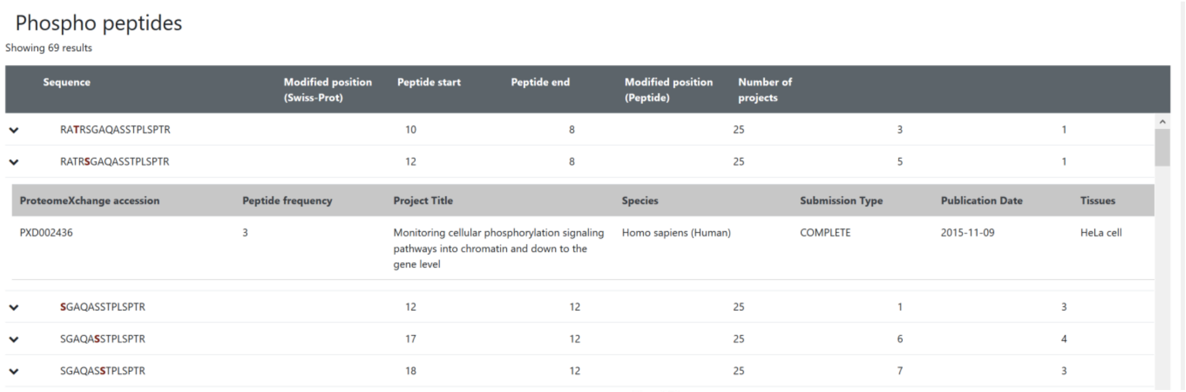
Scop3PTM integrates protein information at the residue level from a variety of resources



sequence information

structural information

Scop3P shows these results interactively on the web, and presents REST APIs for 3rd party re-use



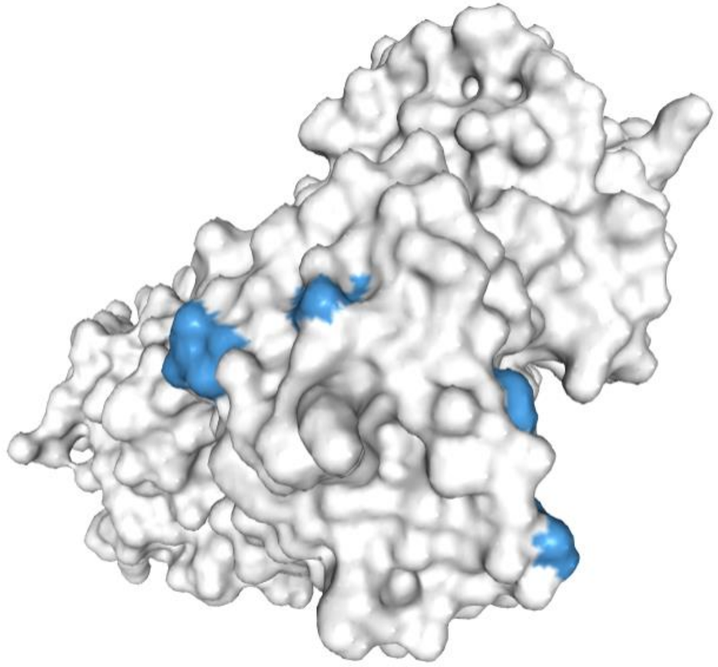
Phospho peptides

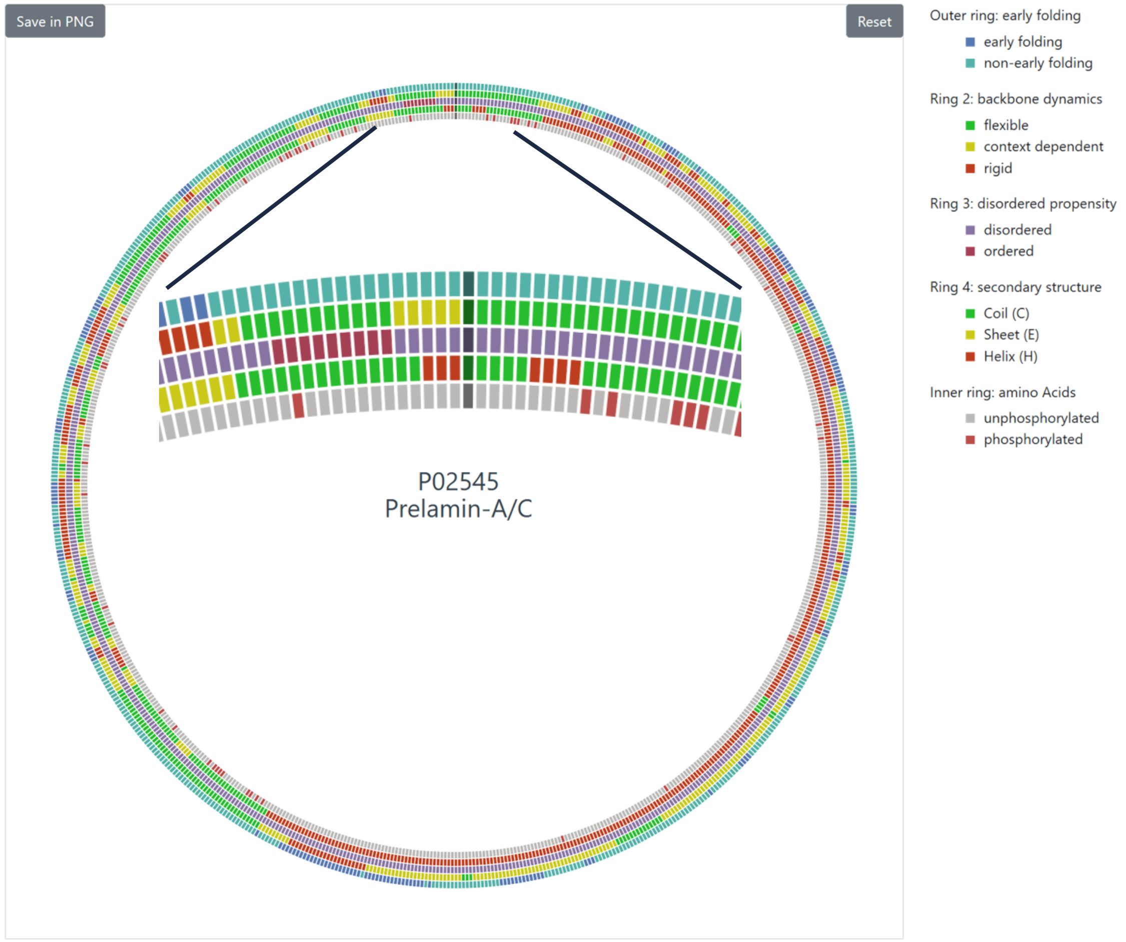
Showing 69 results

	Sequence	Modified position (Swiss-Prot)	Peptide start	Peptide end	Modified position (Peptide)	Number of projects
⌄	RA**T**RSGAQASSTPLSPTR	10	8	25	3	1
⌄	RATR**S**GAQASSTPLSPTR	12	8	25	5	1

ProteomeXchange accession	Peptide frequency	Project Title	Species	Submission Type	Publication Date	Tissues
PXD002436	3	Monitoring cellular phosphorylation signaling pathways into chromatin and down to the gene level	Homo sapiens (Human)	COMPLETE	2015-11-09	HeLa cell

	Sequence	Modified position (Swiss-Prot)	Peptide start	Peptide end	Modified position (Peptide)	Number of projects
⌄	**S**GAQASSTPLSPTR	12	12	25	1	3
⌄	SGAQA**S**STPLSPTR	17	12	25	6	4
⌄	SGAQAS**S**TPLSPTR	18	12	25	7	3

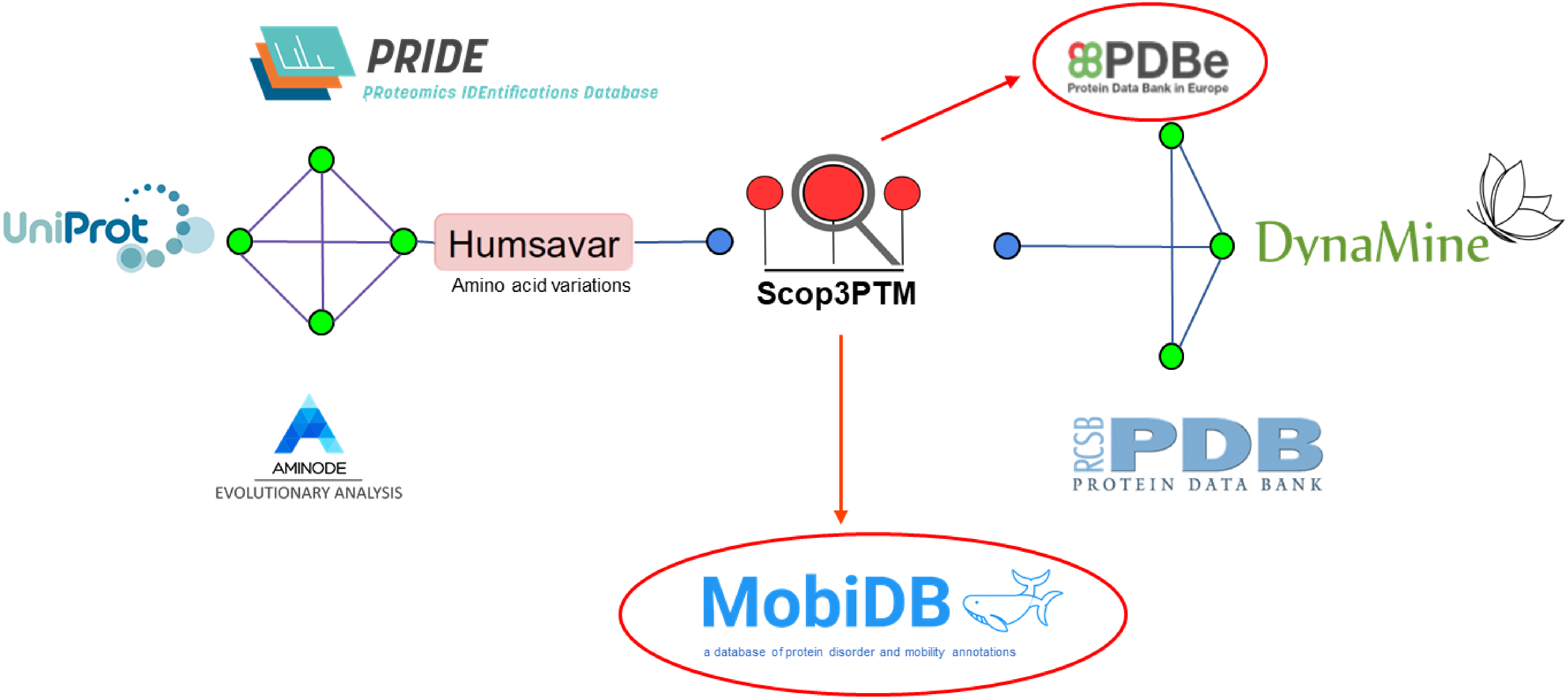




https://iomics.ugent.be/scop3p

Ramasamy, Journal of Proteome Research, 2020

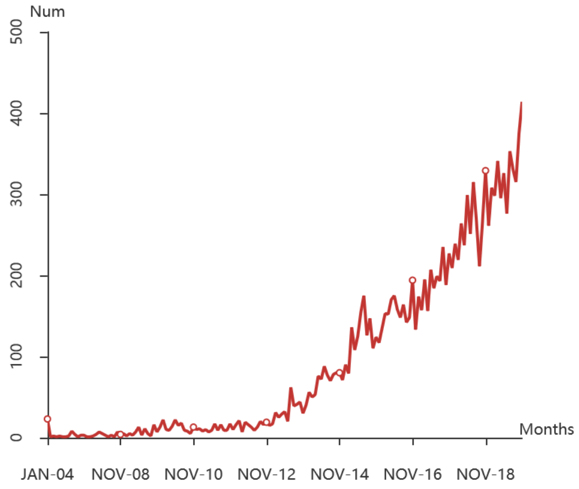
Closing the circle, Scop3P itself becomes a resource for use and re-use by others!



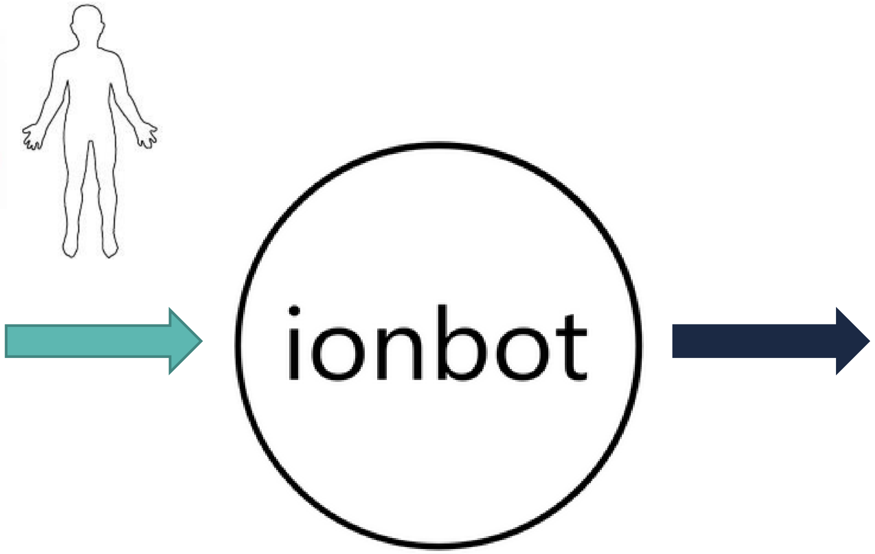
And we are now running ionbot on all human spectra contained in the PRIDE database

PRIDE
PRoteomics IDEntifications Database

Number of Submissions



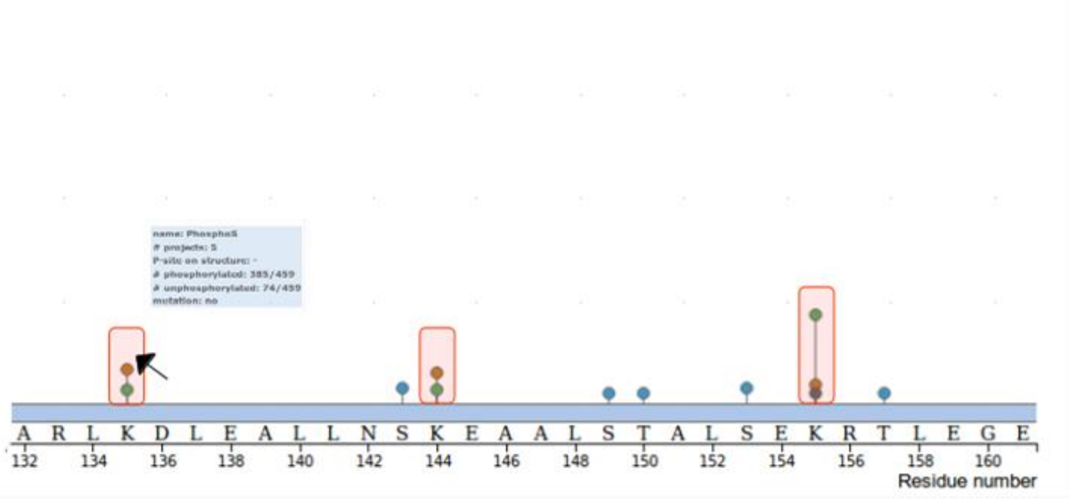
25 314 RAW files
1 *billion* spectra

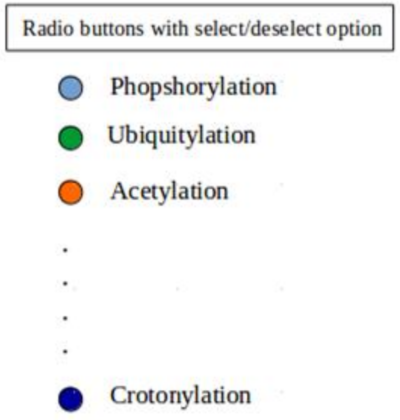


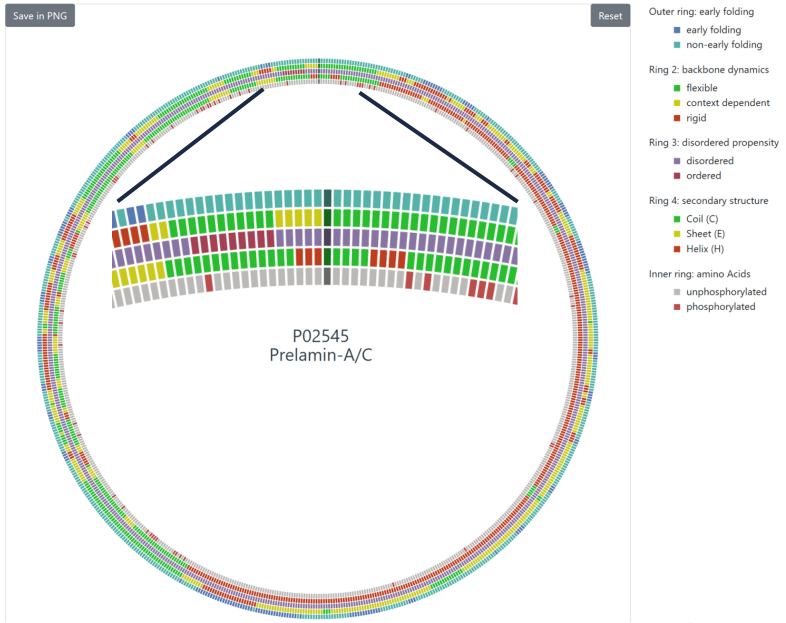
1490 UniMod modifications
all possible AA mutations

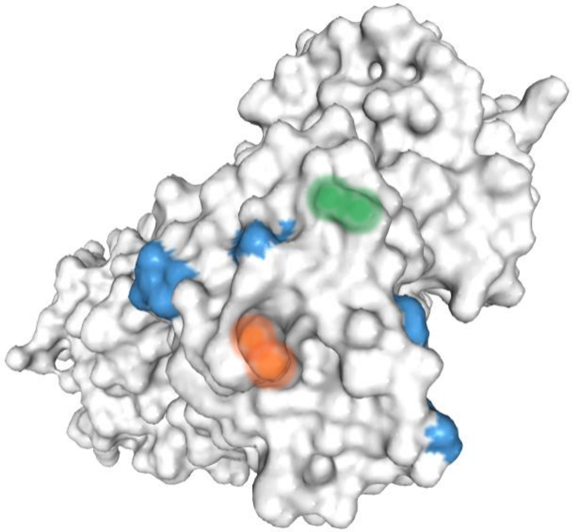
215,046,444	PSMs
742,422	unique peptides
20,246	proteins
444	modifications
4,869,660	modified residues

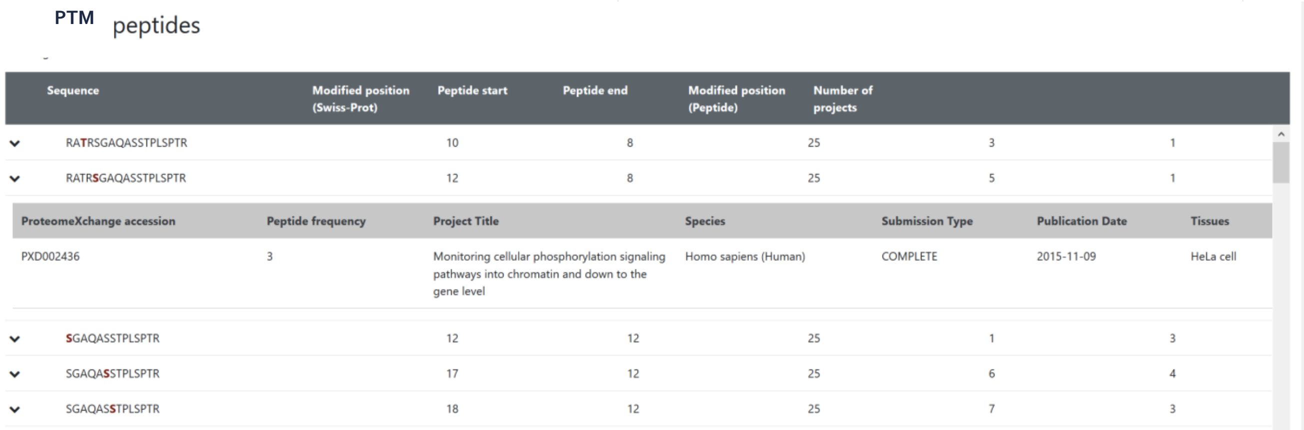
Scop3PTM will become a proteome-wide PTM detection knowledgebase











PTM peptides

Sequence	Modified position (Swiss-Prot)	Peptide start	Peptide end	Modified position (Peptide)	Number of projects
RATRSGAQASSTPLSPTR	10	8	25	3	1
RATRSGAQASSTPLSPTR	12	8	25	5	1

ProteomeXchange accession	Peptide frequency	Project Title	Species	Submission Type	Publication Date	Tissues
PXD002436	3	Monitoring cellular phosphorylation signaling pathways into chromatin and down to the gene level	Homo sapiens (Human)	COMPLETE	2015-11-09	HeLa cell

Sequence	Modified position (Swiss-Prot)	Peptide start	Peptide end	Modified position (Peptide)	Number of projects
SGAQASSTPLSPTR	12	12	25	1	3
SGAQASSTPLSPTR	17	12	25	6	4
SGAQASSTPLSPTR	18	12	25	7	3

Why should we be re-using data?

Four types of (proteomics) data re-use

Unbiased proteome-wide (PT)M discovery as example

A subject of sociological study

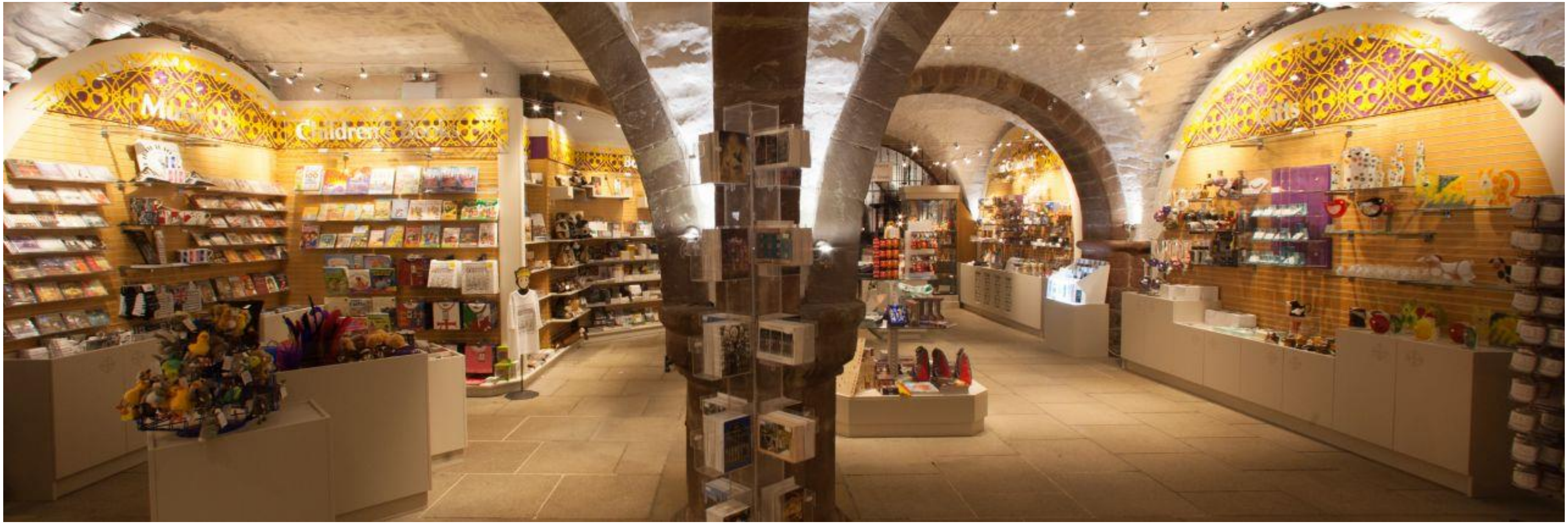
A sociological take on work of my group highlights the key benefits of open science

"This desire to reactivate data is widespread, and Klie et al. are not alone in wanting to show that 'far from being places where data goes to die' (Klie et al., 2007: 190), **such data collections can be mined for valuable information that could not be obtained in any other way**."

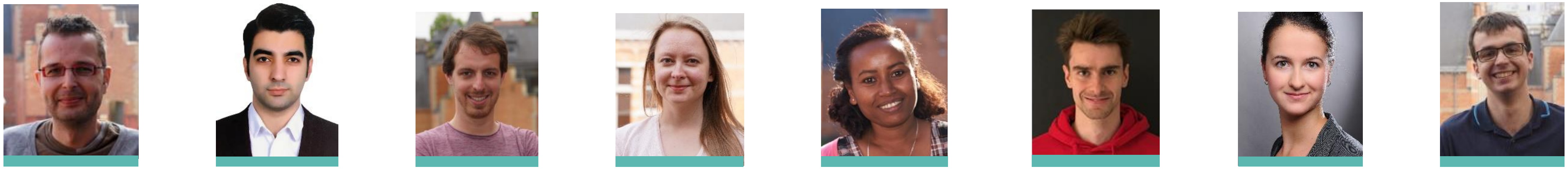
"In attempting to **reactivate sedimented data** in order to enable its re-use, their first step was …"

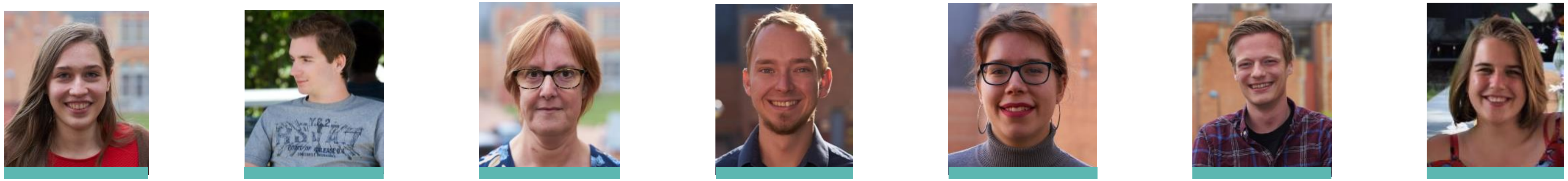
"… they are experiments in seeing, in furnishing ways of seeing how data on proteins could become re-usable, could be reactivated as **collective property rather than the by-product of publication**."

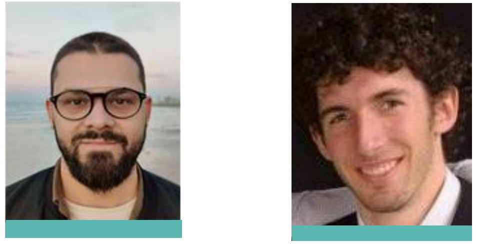
Mackenzie and McNally, *Theory, Culture and Society*, 2013



Gift shop at Chester Cathedral, UK







Comp
omics

VIB
science meets life

GHENT
UNIVERSITY

www.compomics.com

@compomics