

SCIENCE MEETS LIFE

open science

opportunity knocks at your door

lennart martens

lennart.martens@ugent.be

computational omics and systems biology group

Ghent University and VIB, Ghent, Belgium



CC BY-SA 4.0



Vincent Van Gogh – Starry Night

Why should we be re-using data?

The weird and wonderful world of proteomics

Four types of data re-use

Re-using available data to build machine learning models

Reprocessing data with new models for new insights

Repurposing large-scale data for new knowledge

Why should we be re-using data?

The weird and wonderful world of proteomics

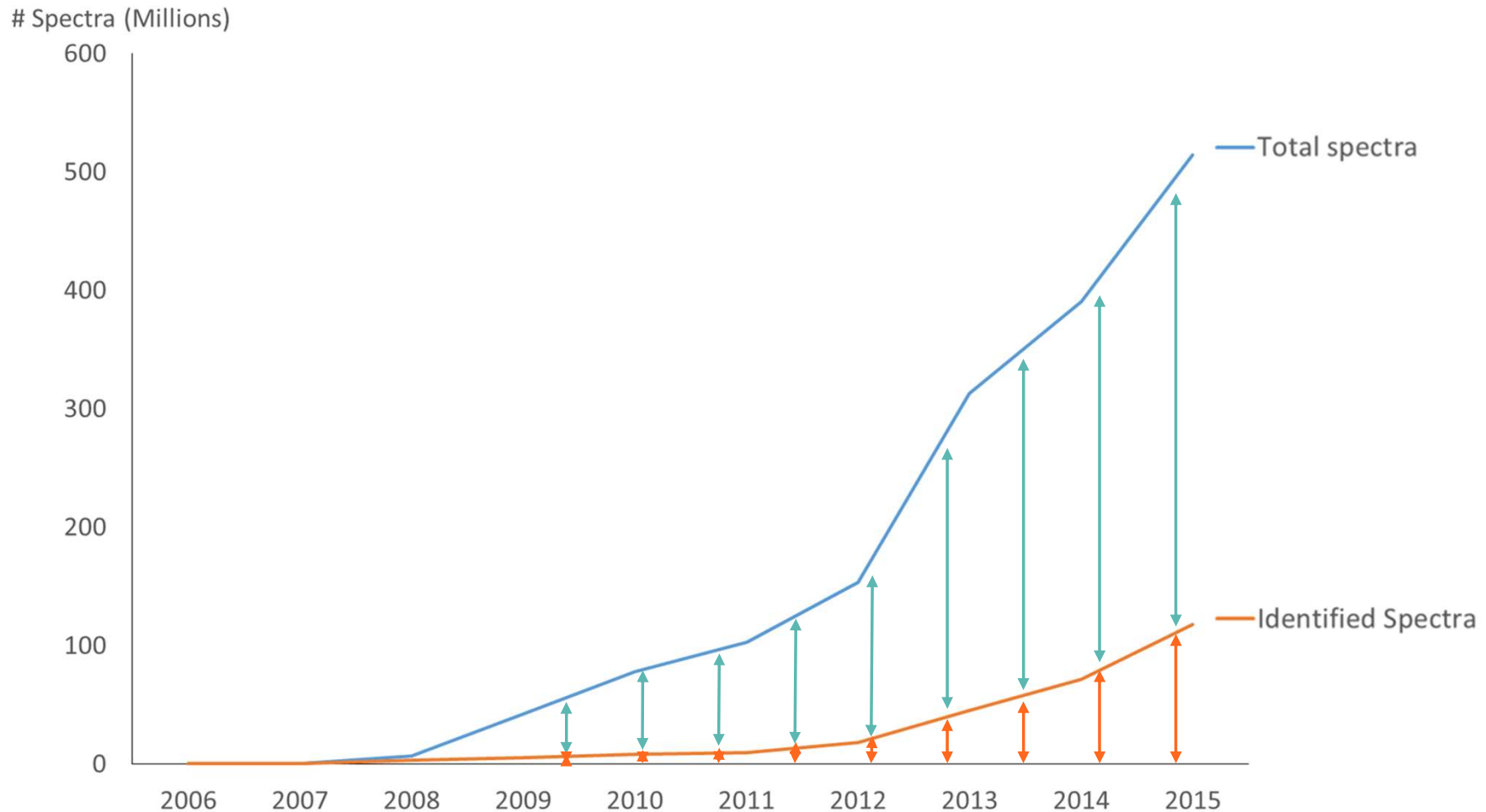
Four types of data re-use

Re-using available data to build machine learning models

Reprocessing data with new models for new insights

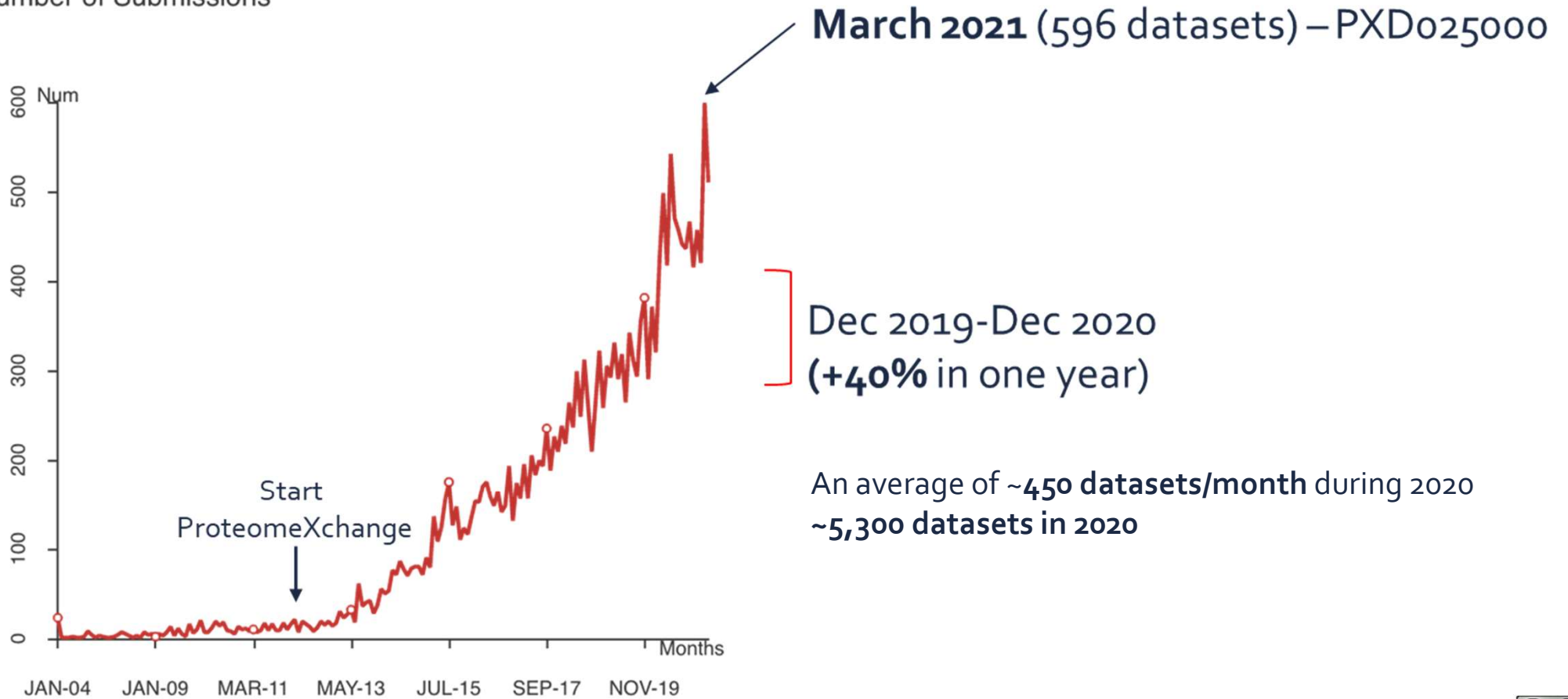
Repurposing large-scale data for new knowledge

A lot of data these days is high-content, meaning that much more data is acquired than is used in most papers



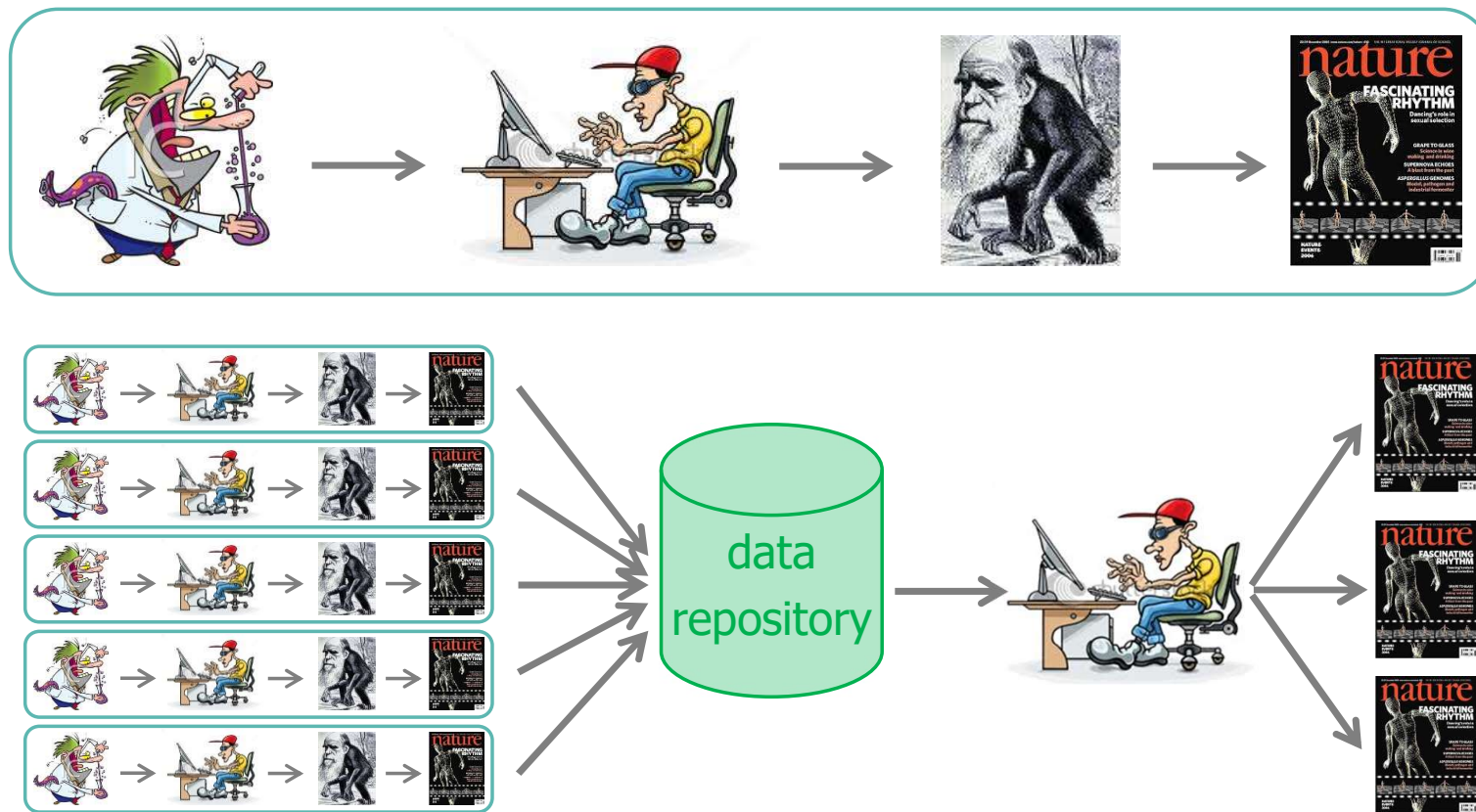
Most of our data is also high throughput, meaning there is lots of data available!

Number of Submissions



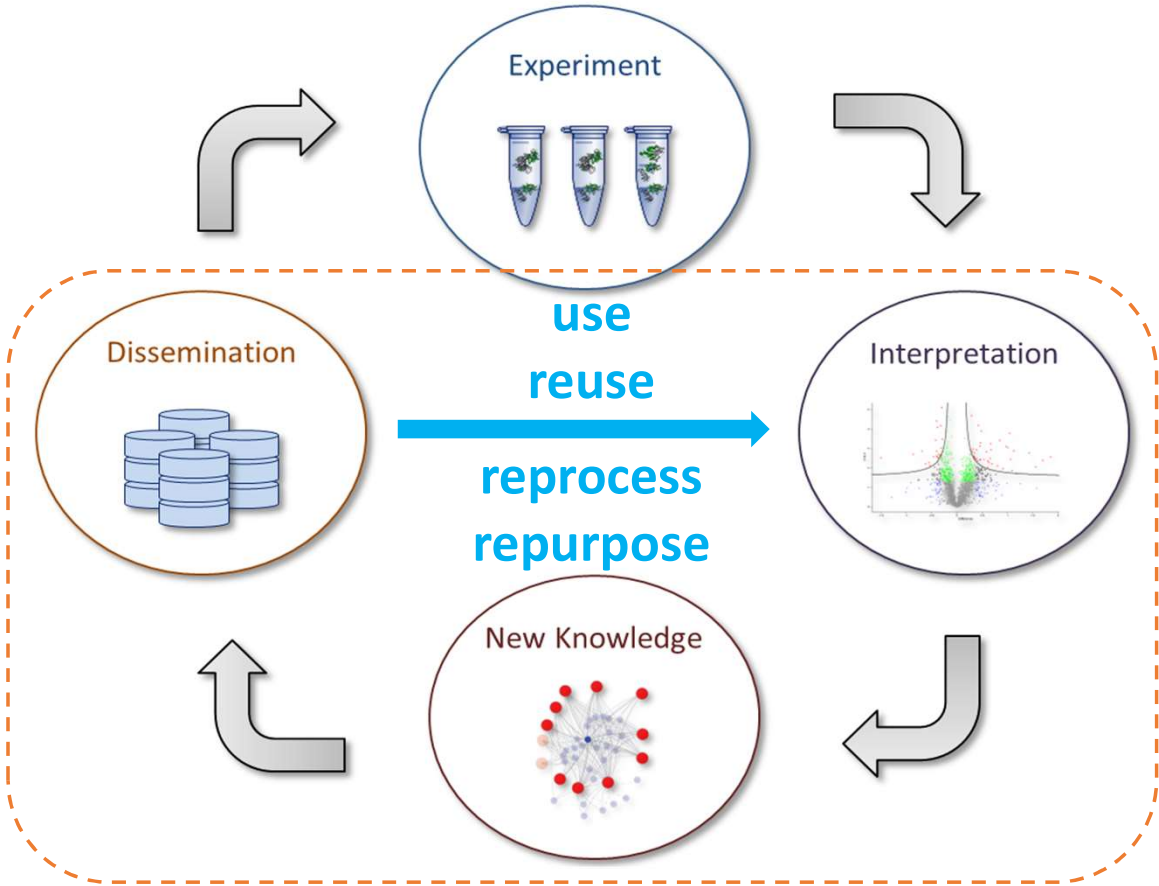
Slide courtesy of Dr. Juan Antonio Vizcaíno, Proteomics Team Leader, EMBL-EBI

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



An open data exchange ecosystem allows for productive (and completely novel!) data uses

Open Data Exchange Ecosystem



Adapted from: Vaudel, Proteomics, 2016

Why should we be re-using data?

The weird and wonderful world of proteomics

Four types of data re-use

Re-using available data to build machine learning models

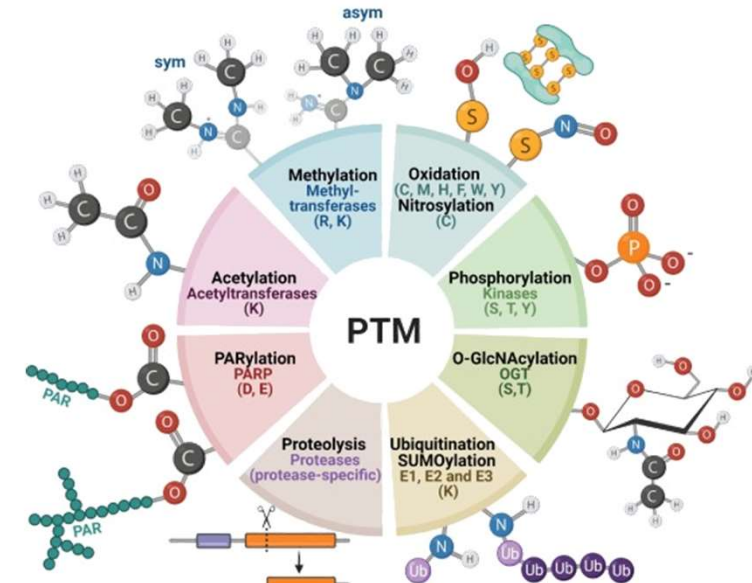
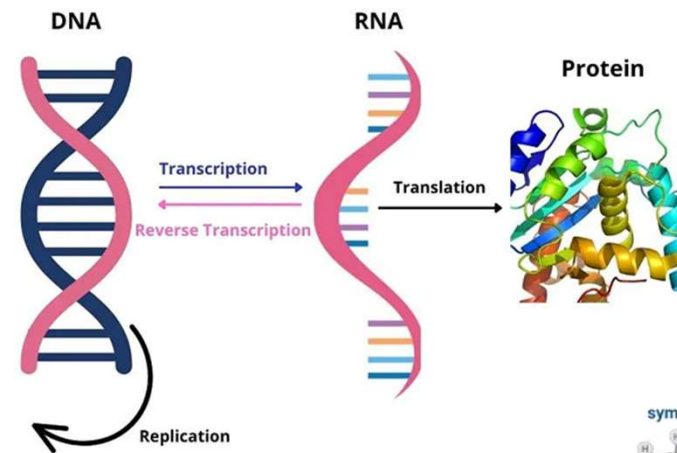
Reprocessing data with new models for new insights

Repurposing large-scale data for new knowledge

Proteomics studies proteins, the executive molecules in a cell, which are regulated by, and exposed to, chemical modifications

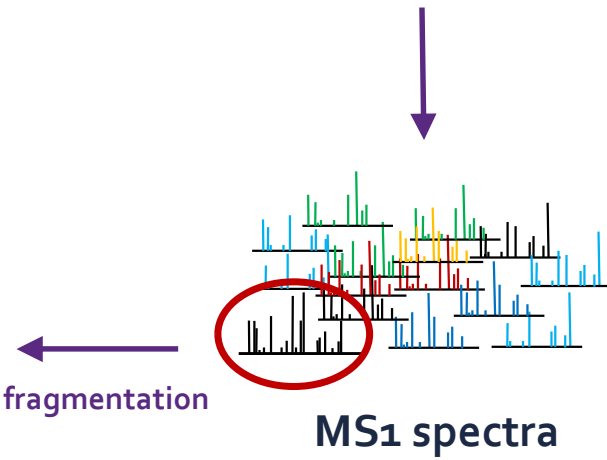
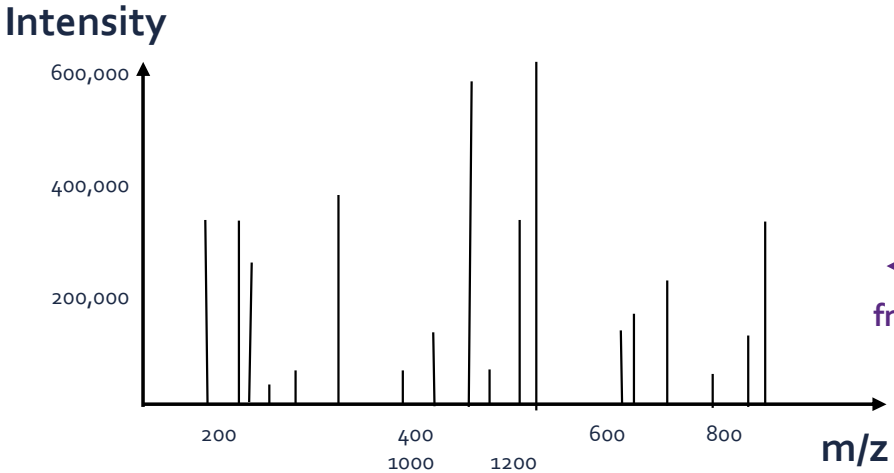
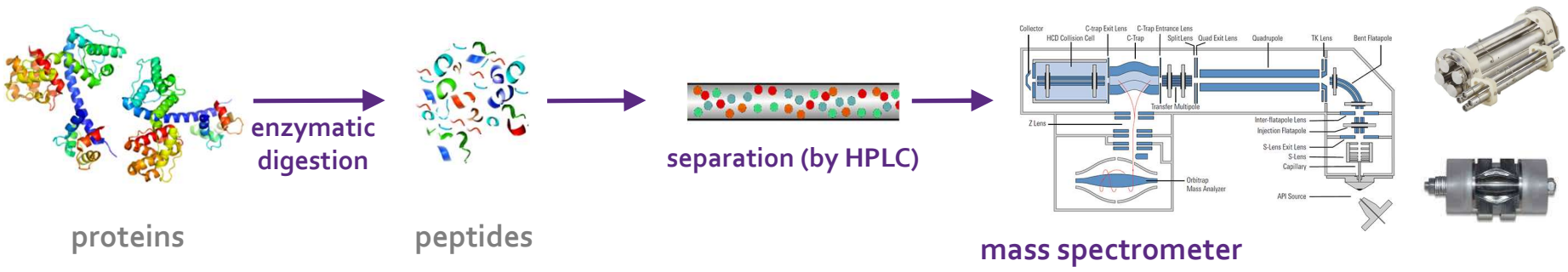


Gael McGill
<https://gaelmcgill.artstation.com/projects/PmoJL1>



Sternburg, Trends Biochem. Sci., 2022

A typical proteomics workflow from sample to data involves sample preparation, and mass spectrometry



We resolve this ambiguity by using a database as filter, and we make important assumptions in the process

database

```
>sw|Q9NZ18|IF2B1_HUMAN Insulin-like grow
protein 1 OS=Homo sapiens GN=IGF2BP1 PE=
MNKLYIGNLNESTVPALLEKVFAEHKISYSGQLVRSQYA
GKVELQGRLEIEHSVFKRQRSRKIQIRNIFQLRWEVLD
SETAVVNVVYISNREYRQAIMKLNHGHLENHALKVSYIFD
GQFRQQSPVAAAGAPAKQQVDIFLRLLVPTQVYGAIGRE
RRENAGAAEKASVHSTPEGCSSACKMILEIMHEAKDTK
LIGREGNLLKVEQDTETKRTISSLQULTLYNPERTITVK
EAYENDVAAMSLQSHLIPGLNLAAVGLFPASSAVPPFPB
MVQVFI PAQAVGAIIGKKGHIKQLSRPASASIKIAPPET
KAQRITYGKLEENFFGKREEVKLETHIRYPAABAAGRVIG
VPRDQTPDENQIVIKIIGHFYASQMAQRKIRDILAQVKQ
>sw|Q8TF68|ZN384_HUMAN Zinc finger prote
GN=ZNF384 PE=1 SV=2
VREGLSSTGKPKDQATYDPTGQVCFSTLREYKPKATYRSP
```

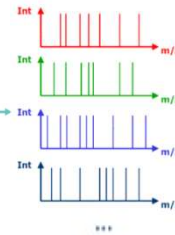
in silico
digest

peptide seq.

YSVATAER
HETSINGK
MILQEESTVYRR
SEFASTPINK
...

in silico
MS/MS

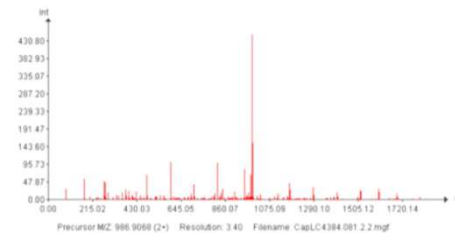
theoretical spectra



in silico
matching

peptide scores

- 1) YSFVATAER 34
- 2) YSFVSAIR 12
- 3) FFLIGGGGK 12
- ...



experimental spectra

Why should we be re-using data?

The weird and wonderful world of proteomics

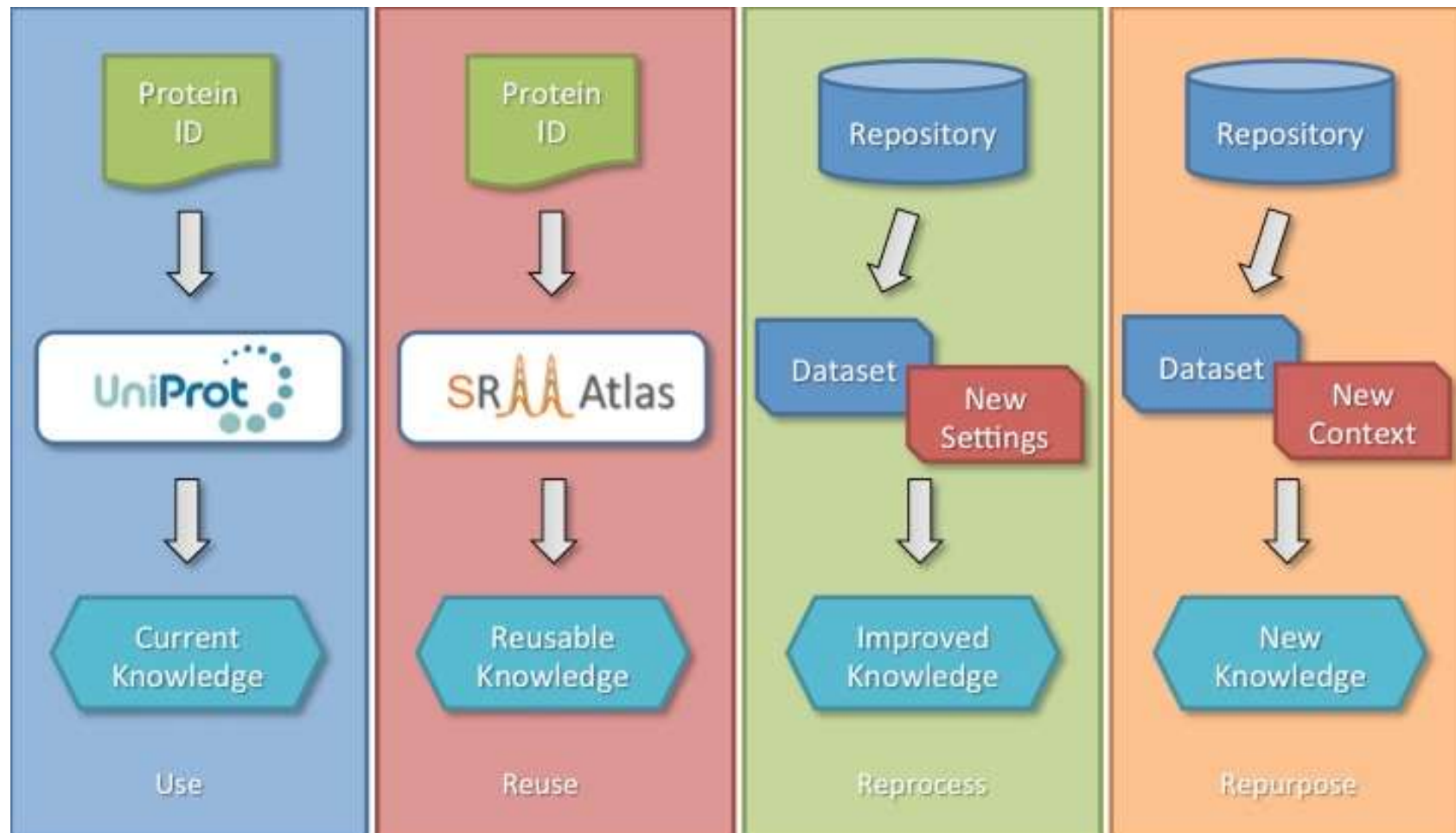
Four types of data re-use

Re-using available data to build machine learning models

Reprocessing data with new models for new insights

Repurposing large-scale data for new knowledge

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



We may want to make a choice of how we frame open data

Show me your data, now!

I don't trust you!

I'll find all your mistakes!

This will not end well!



Could I look at your data?

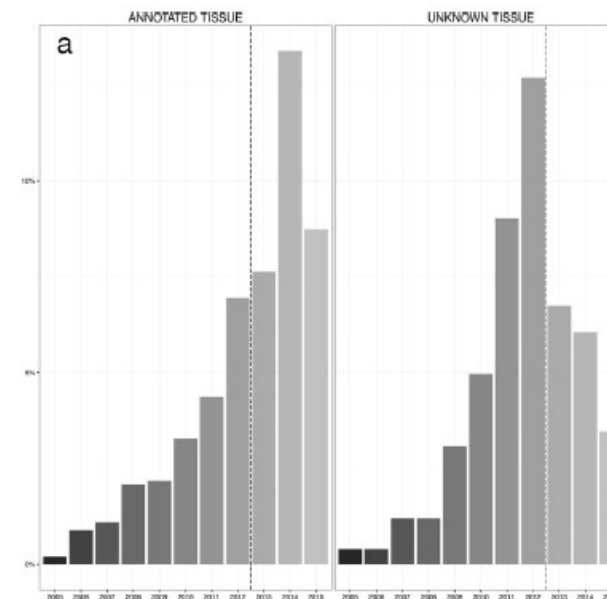
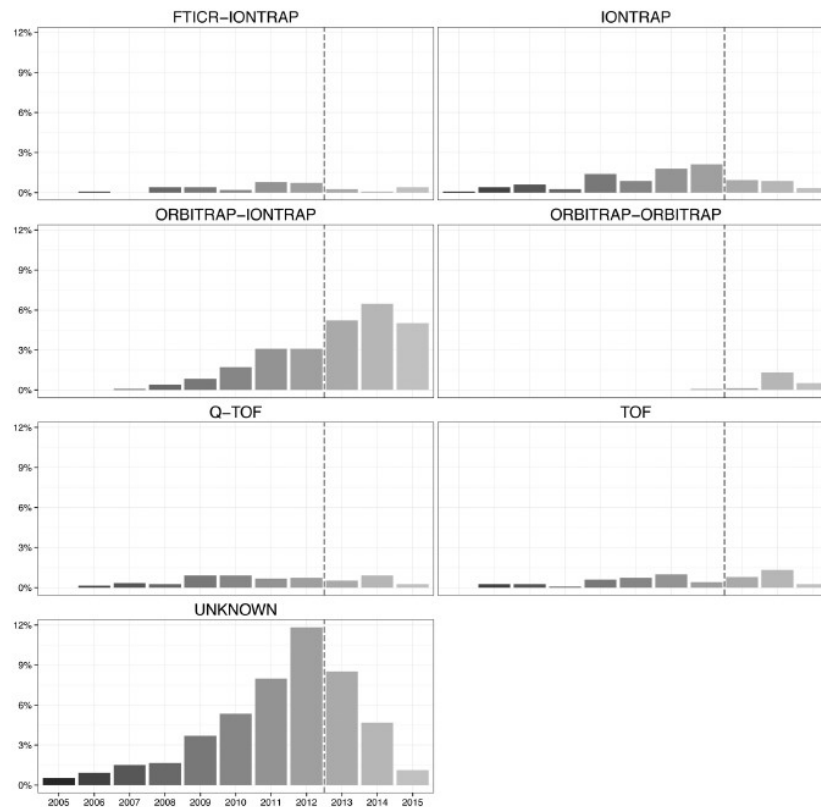
OK, this is pretty cool!

Look what I found in here!

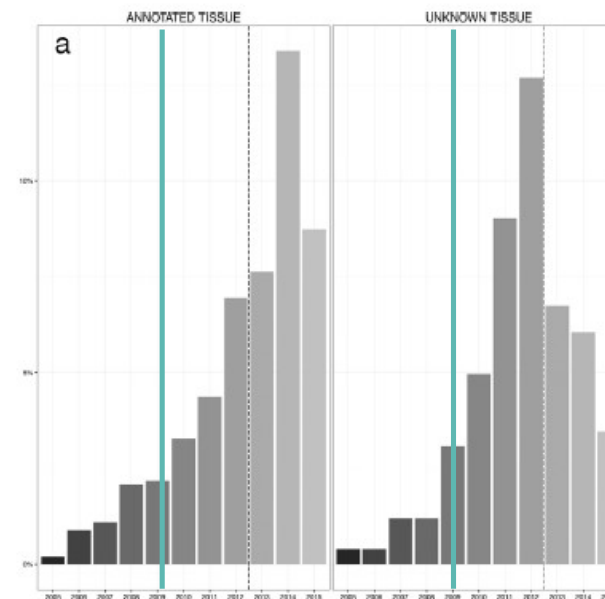
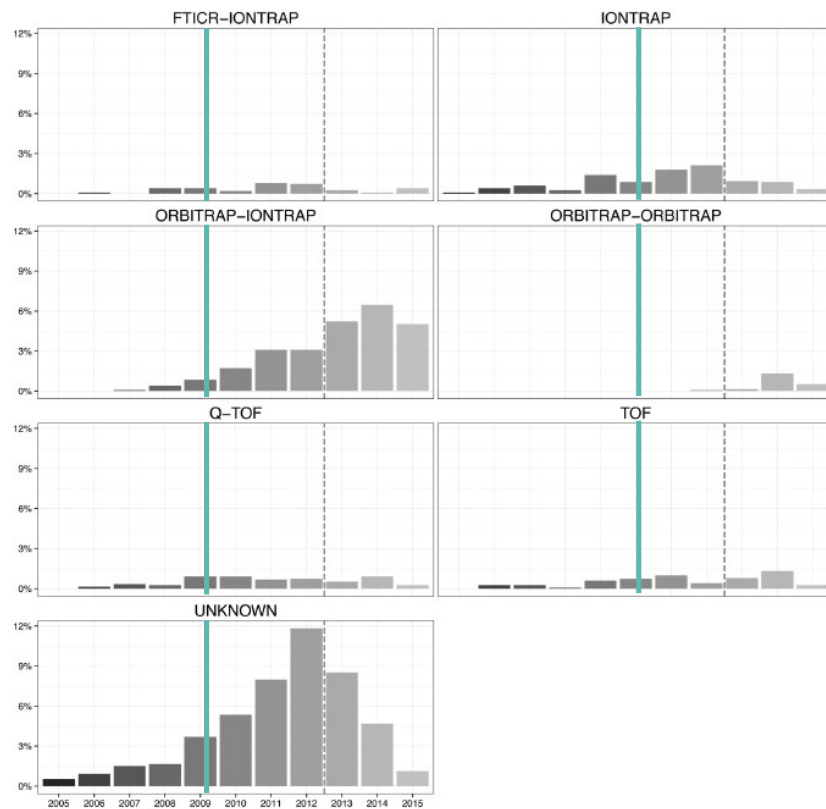
Your data is so useful!

And let us not forget that your data will most likely live a much longer and more useful life than your publication!

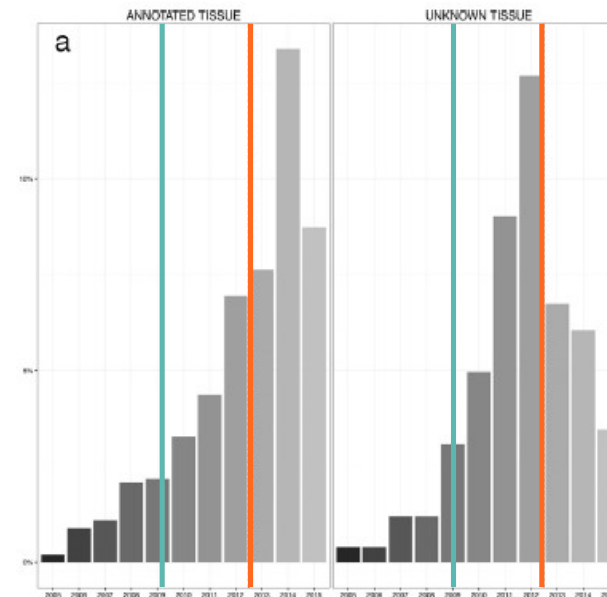
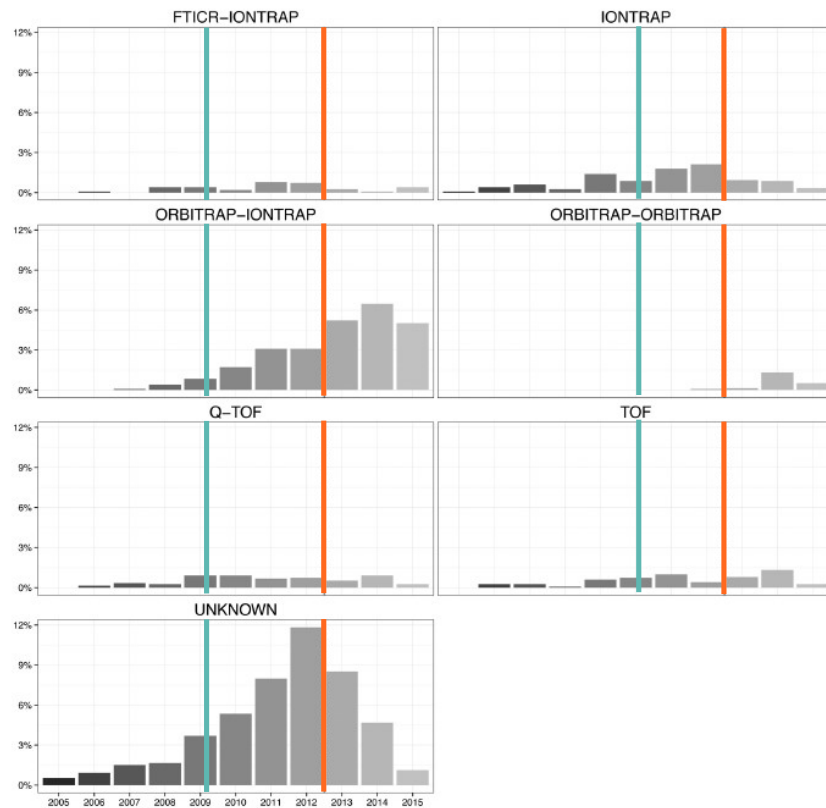
Metadata is often the key issue, as it requires manual work



Even user-friendly submission tools cannot correct for a lack of elementary motivation

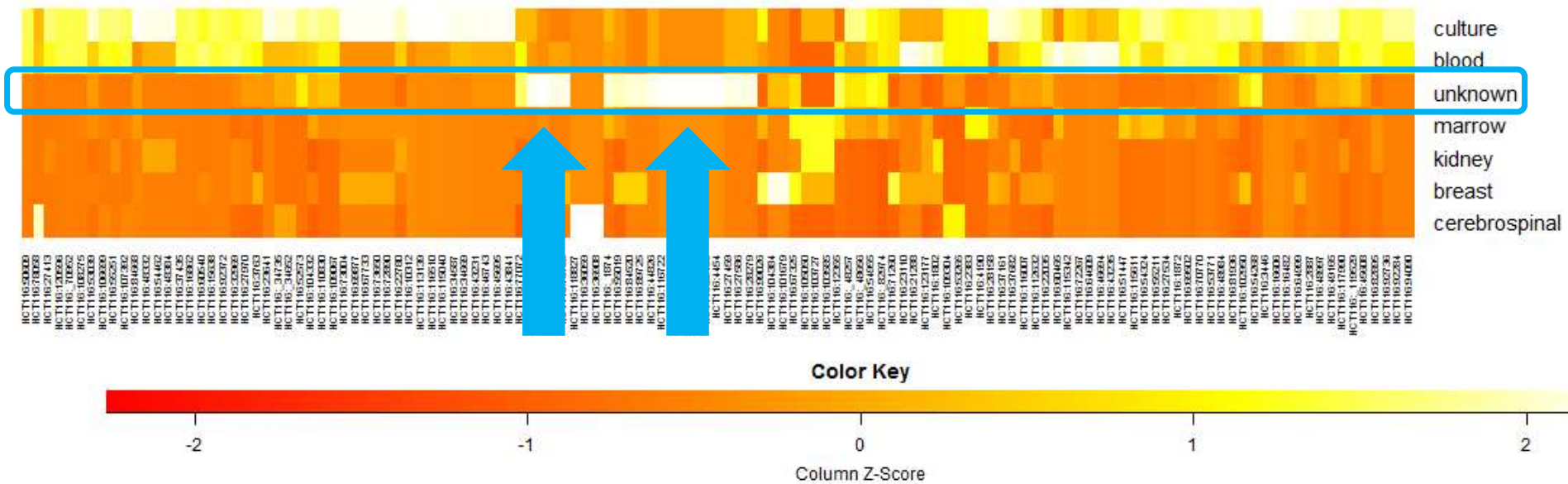


Manual curation of submissions, equivalent to restrictive policing, does help

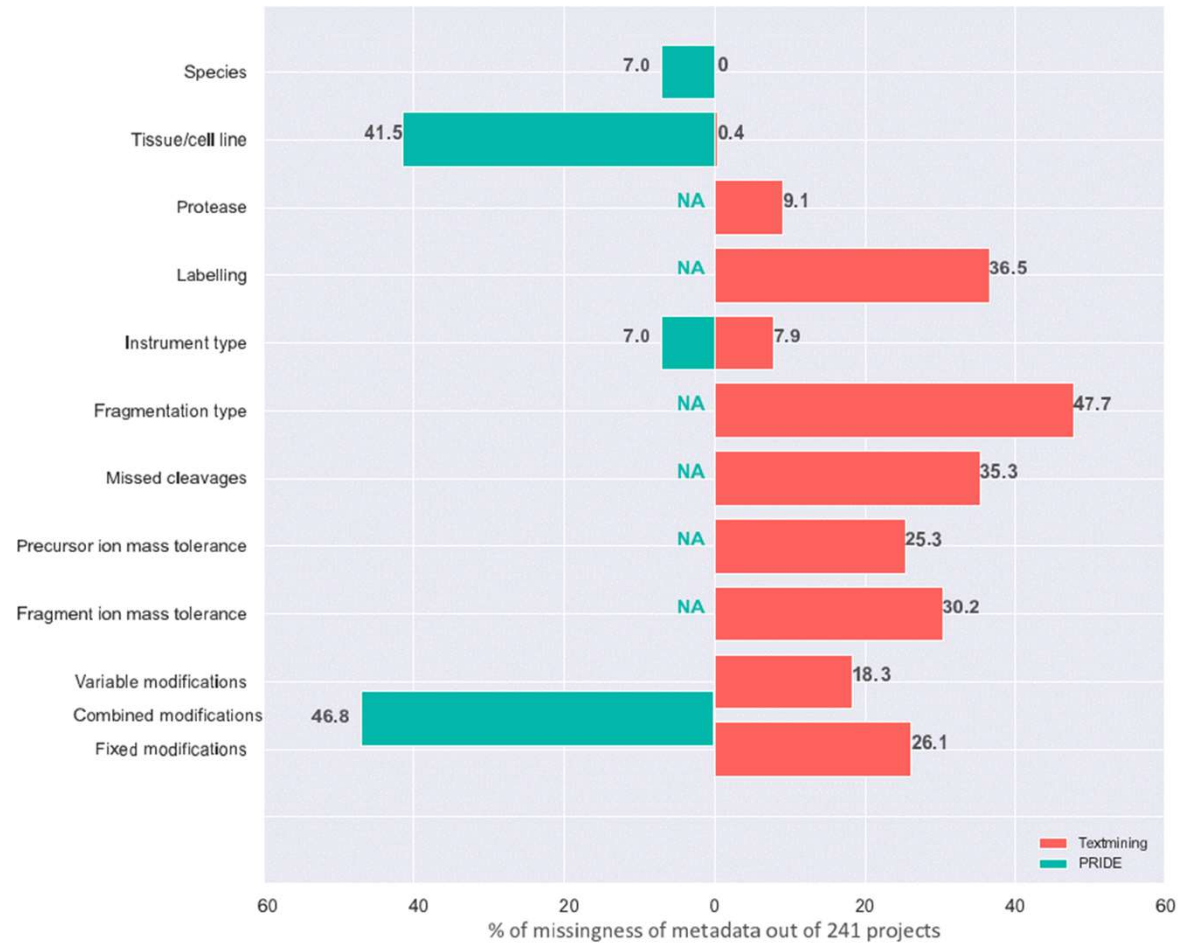
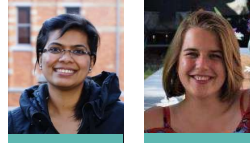


Missing metadata becomes pretty annoying when people successfully re-use your data

#PSMs per tissue per sORFs with more than 5 occurrences



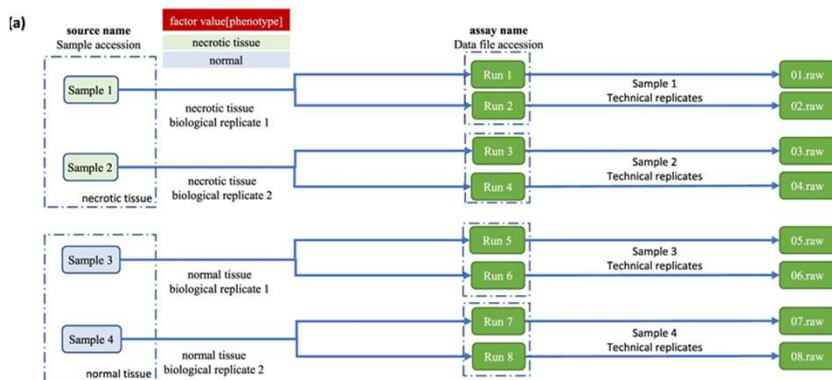
Metadata annotation in both PRIDE and articles(!) remains a major problem in proteomics



SDRF is a metadata annotation format meant to tackle this, but it is far from easy to work with, hampering adoption



Sample to Data Relationship Format (SDRF)



(b) SDRF table structure showing columns for Sample properties, Data file properties, and Study variables.

Sample properties					Data file properties				Study variables		
source name	characteristics [organism]	characteristics [disease]	characteristics [phenotype]	characteristics [biological replicate]	assay name	comment [fraction identifier]	comment [label]	comment [technical replicate]	comment [data file]	factor value [phenotype]	
Sample 1	homo sapiens	...	liver cancer	necrotic tissue	1	Run 1	1	label free sample	1	01.raw	necrotic tissue
Sample 1	homo sapiens	...	liver cancer	necrotic tissue	1	Run 2	1	label free sample	2	02.raw	necrotic tissue
Sample 2	homo sapiens	...	liver cancer	necrotic tissue	2	Run 3	1	label free sample	1	03.raw	necrotic tissue
Sample 2	homo sapiens	...	liver cancer	necrotic tissue	2	Run 4	1	label free sample	2	04.raw	necrotic tissue
Sample 3	homo sapiens	...	liver cancer	normal	1	Run 5	1	label free sample	1	05.raw	normal
Sample 3	homo sapiens	...	liver cancer	normal	1	Run 6	1	label free sample	2	06.raw	normal
Sample 4	homo sapiens	...	liver cancer	normal	2	Run 7	1	label free sample	1	07.Raw	normal
Sample 4	homo sapiens	...	liver cancer	normal	2	Run 8	1	label free sample	2	08.raw	normal

So we built an online, free, open, and easy-to-use tool to provide proteomics metadata: lesSDRF



Ontologies

SDRF guidelines
Local metadata

lesSDRF
freely available
open source
no install required

Welcome to lesSDRF

The SDRF annotation tool, because spending less time on SDRF creates more time for amazing research

SDRF structure questions

Map local metadata to ontology terms

1. Map local metadata to SDRF

If you have a local metadata file available, you can use this file to map the data to the required SDRF information.

Important: you can upload the file in csv, tsv or xls format.
The order of your raw file names should match the order in which you inputted them in the previous step

This is your current SDRF file:

	source name	characteristics[organism]	characteristics[organism part]	characteristics[cell type]	characteristics[ancestry category]	characteristics[age]	characteristics[sex]	characteristics[disease]	characteristics[individual]	characteristics[biological replicate]	technology type	assay name	comment[technical replicate]	comment[fraction identifier]	comment[label]	comment[instrument]	comment[cleavage agent]
0	None	None	liver	None	None	None	None	None	None	None	None	None	None	None	None	None	None
1	None	None	liver	None	None	None	None	None	None	None	None	None	None	None	None	None	None
2	None	None	liver	None	None	None	None	None	None	None	None	None	None	None	None	None	None
3	None	None	liver	None	None	None	None	None	None	None	None	None	None	None	None	None	None

Upload your local metadata file (.csv, .tsv or .xls)

Drag and drop file here
Link: 1028 per file - CSV, TSV, XLSX

Browse files

demo_metadata.tsv 383.00B

Your metadata file:

	source name	organism	part	cell	assay name	cleavage	instrument
0	Mix1	Homo sapiens	liver	Hepatocyte_mixture	hep_mix1.raw	Trypsin	Orbitrap-velos
1	Mix2	Homo sapiens	liver	Hepatocyte_mixture	hep_mix2.raw	Trypsin	Orbitrap-velos
2	Mix3	Homo sapiens	liver	Hepatocyte_mixture	hep_mix3.raw	Trypsin	Orbitrap-velos
3	Mix4	Homo sapiens	liver	Hepatocyte_mixture	hep_mix4.raw	Trypsin	Orbitrap-velos

Select columns containing data that will be used in the SDRF data

Ready to match?

Select column 1 to match in your metadata file: part

Select the corresponding column from the SDRF file: characteristics[organism p...

Match and check ontology

Select column 2 to match in your metadata file: cell

Select the corresponding column from the SDRF file: characteristics[cell line]

Match and check ontology

Ontology check

Great! The local metadata values are valid terms and are mapped to the SDRF file.

⚠️ (Hepatocyte_mixture) are not ontology terms. Select the correct terms in the next steps directly from the ontology

To-do list of columns

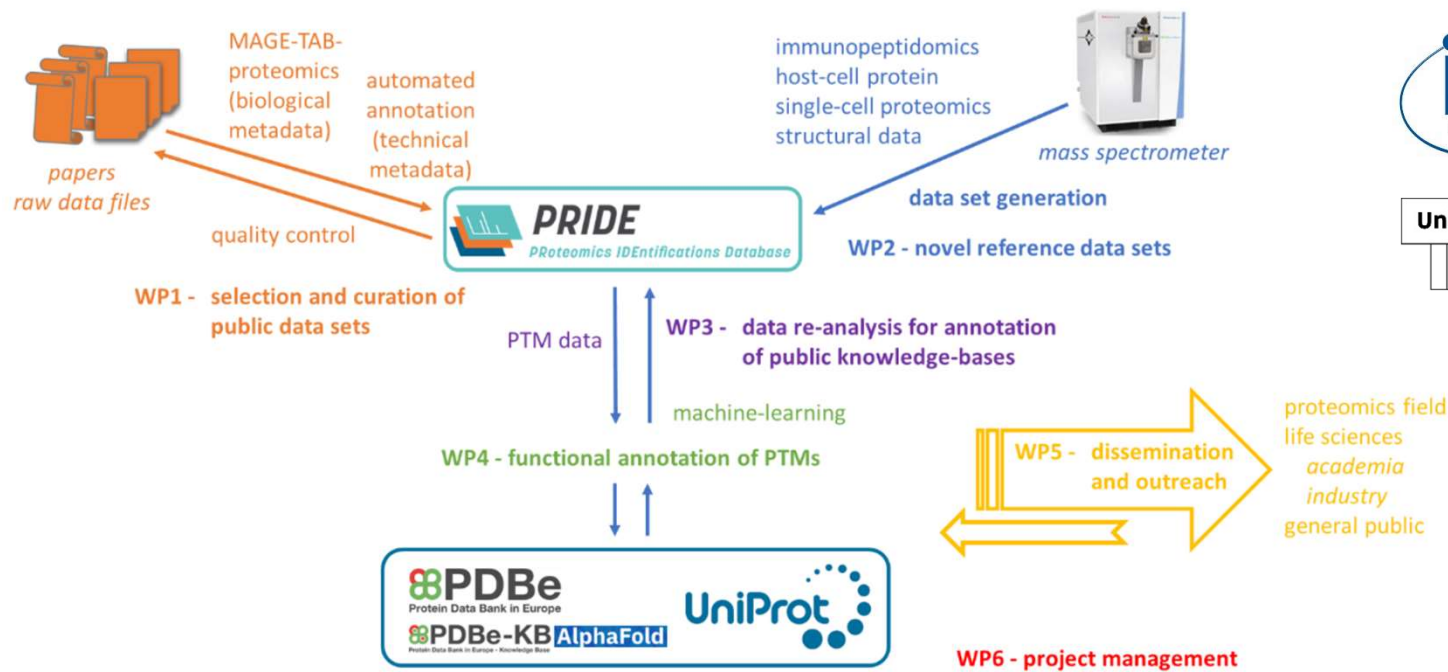
Autocomplete

<https://github.com/compomics/lesSDRF>
Clayes, Nature Communications, 2023

Our joint CHIST-ERA “ODEEP-EU” project has been selected as a generic showcase for what open science can deliver today

Open & Re-usable Research Data & Software (CHIST-ERA ORD call)

This call tackles the challenge of **open research data and software** from the perspective of their **possible reuse**. The objective is to create the conditions for **research in any domain** based on open or shared data and software.



CC BY-SA 4.0

Why should we be re-using data?

The weird and wonderful world of proteomics

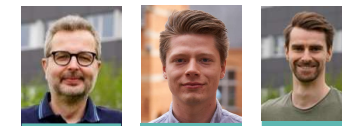
Four types of data re-use

Re-using available data to build machine learning models

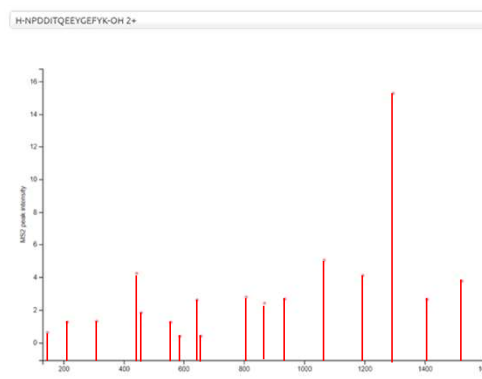
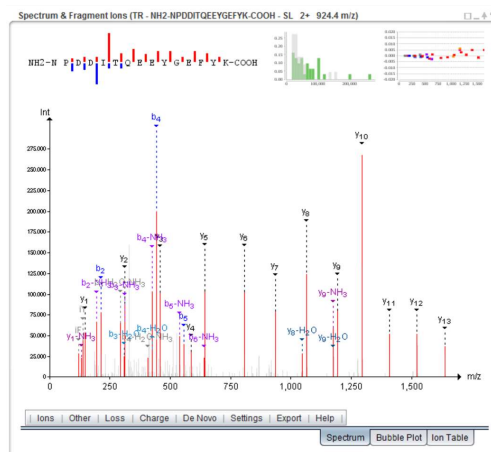
Reprocessing data with new models for new insights

Repurposing large-scale data for new knowledge

Our MS²PIP fragmentation model accurately predicts peptide fragmentation behaviour in varying conditions

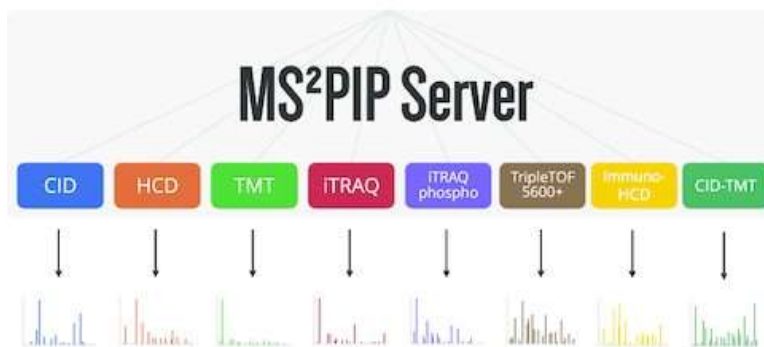


Vaudel, Nat. Biotech., 2015

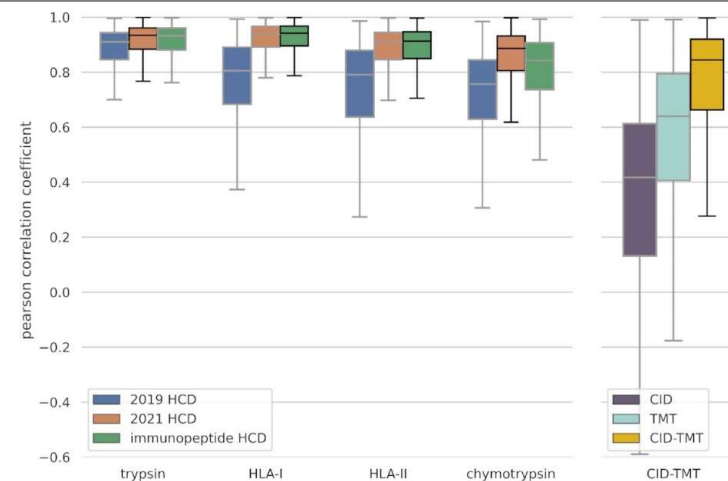


<https://iomics.ugent.be/ms2pip>
 Degroeve, Bioinformatics, 2013
 Degroeve, Nucleic Acids Research, 2015
 Gabriels, Nucleic Acids Research, 2019

CCDLHTREEARK/2

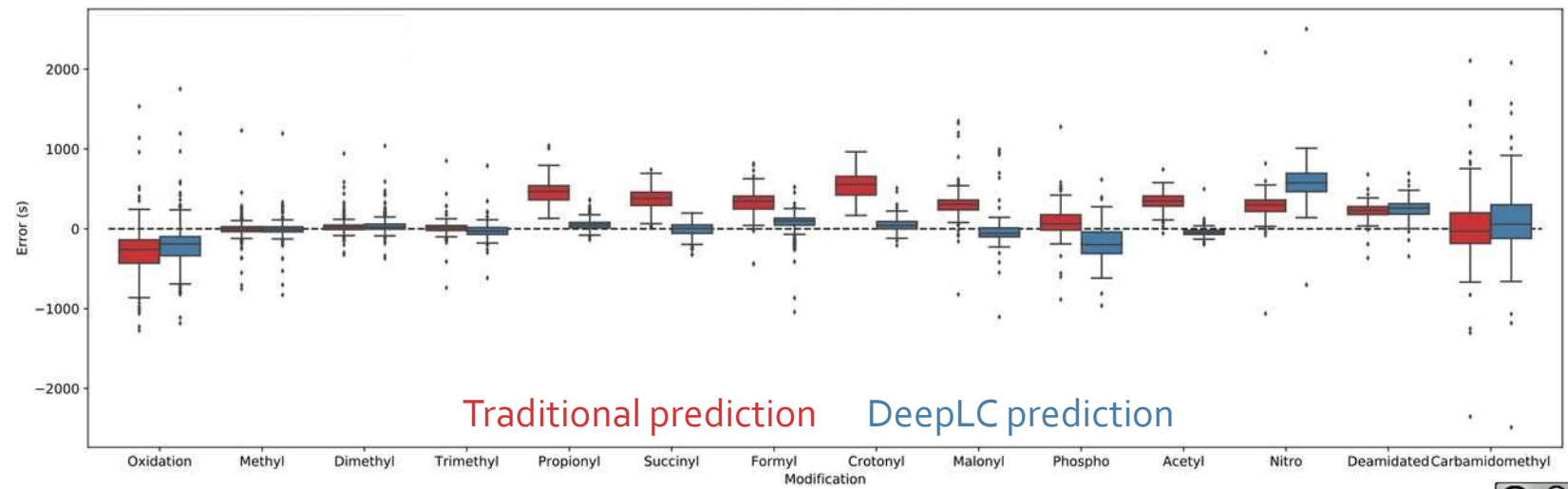
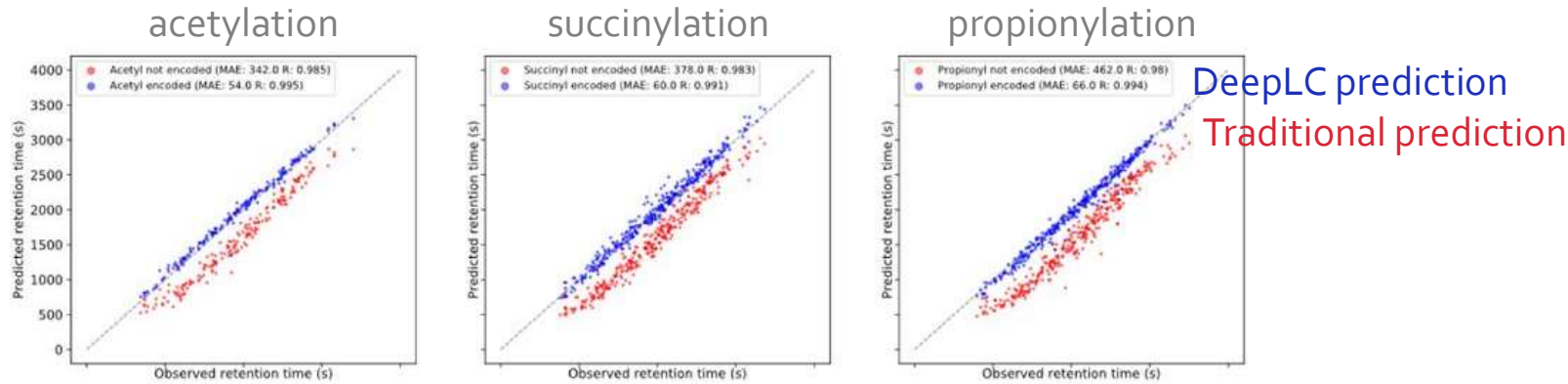


<https://iomics.ugent.be/ms2pip>
 Declercq, Nucleic Acids Research, 2023



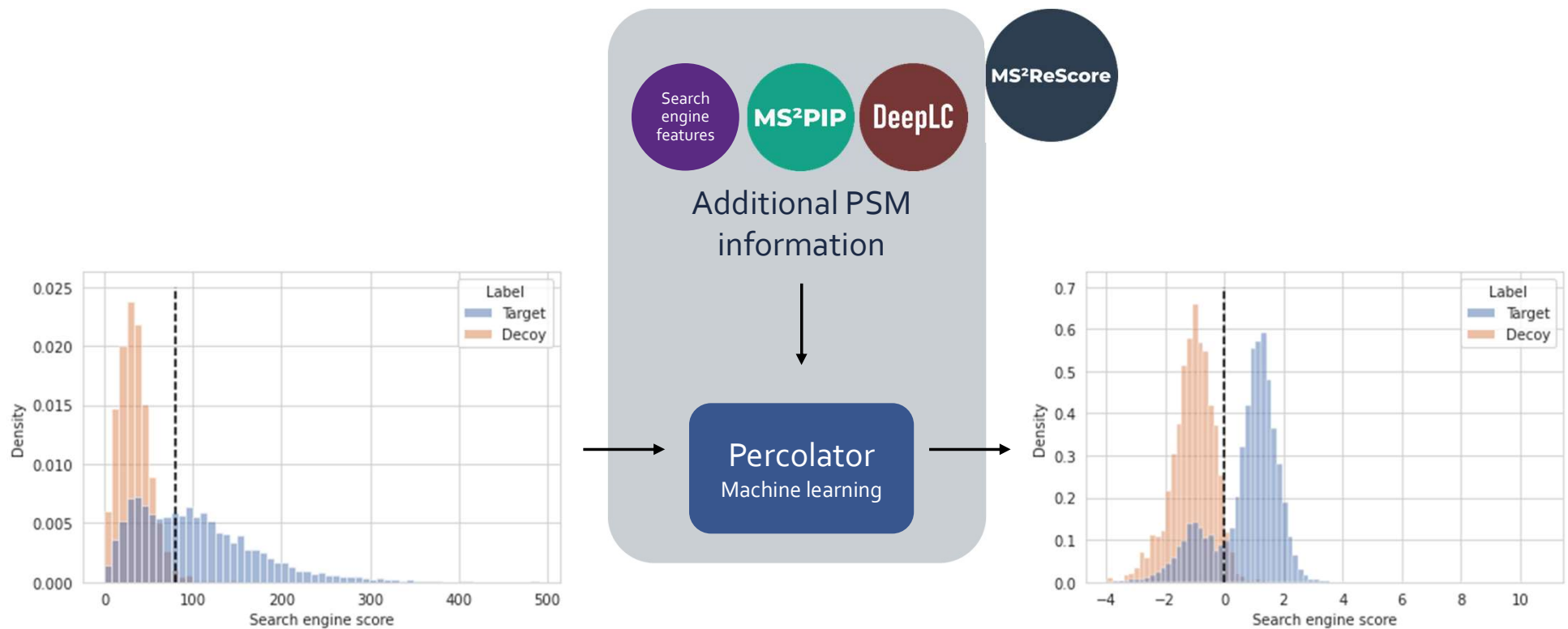
CC BY-SA 4.0

Our DeepLC retention time model accurately predicts retention times of peptides with as-yet unseen modifications



MS²Rescore makes use of these machine learning predictions to rescore identifications for improved sensitivity

<https://github.com/compomics/ms2rescore>

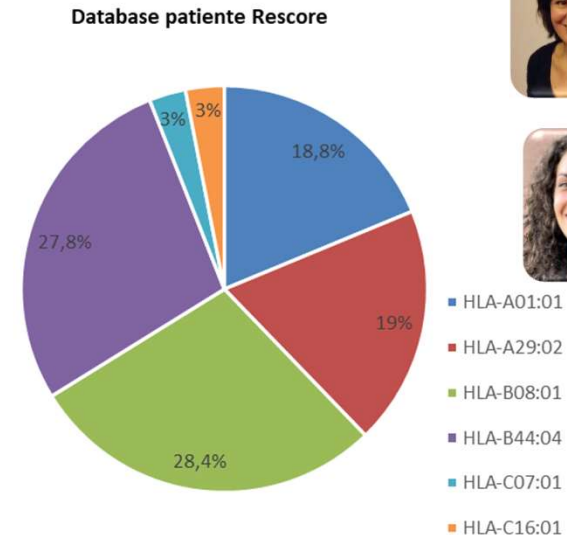
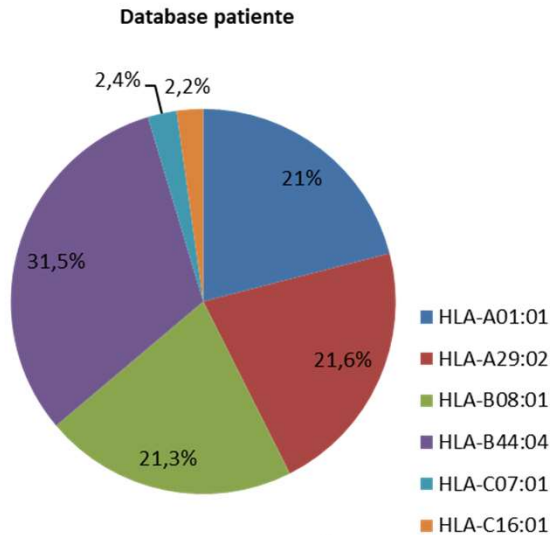


MS²Rescore: C. Silva, Bioinformatics (2019) & Declercq, MCP, 2022
Percolator: L. Käll, Journal of Proteome Research (2009)



CC BY-SA 4.0

On patient-derived tumour samples from Uni Strasbourg, MS2Rescore is proven to be reliable and sensitive



Analyse affinités		
	Nb	%
total source 8-14 AA (soumis à NetMHC pan)	2090	100
total affinités	1978	94.641148
total sans affinités	112	5.3588517
	Nb	%
total affinités	1978	100
total fortes affinités	1896	95.854398
total moyennes affinités	82	4.1456016

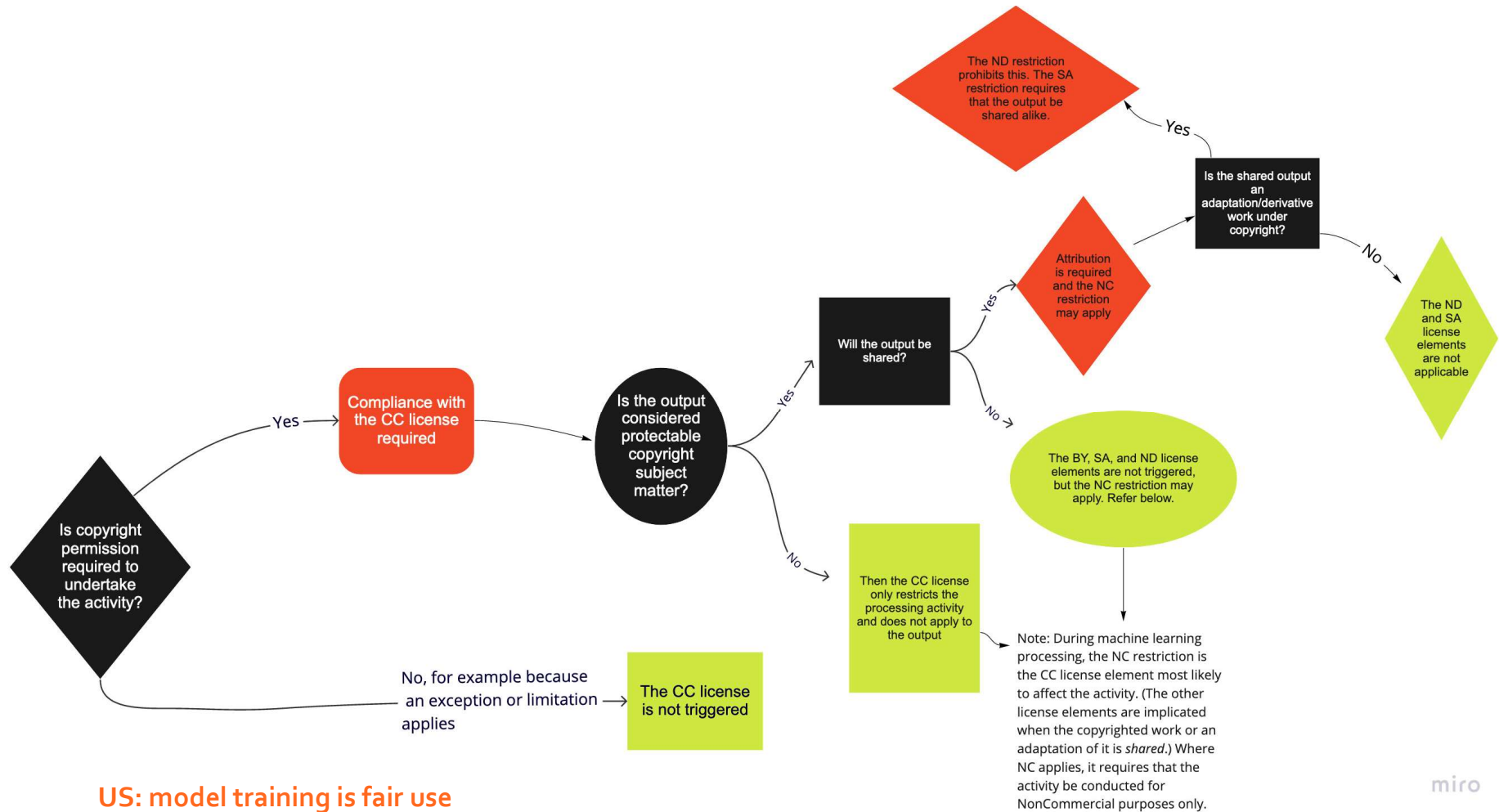
Analyse affinités		
	Nb	%
total source 8-14 AA (soumis à NetMHC pan)	3663	100
total affinités	3433	93.7209937
total sans affinités	230	6.27900628
	Nb	%
total affinités	3433	100
total fortes affinités	3242	94.436353
total moyennes affinités	191	5.56364696

Slide credit: Perrine Spinnhirny



CC BY-SA 4.0

A word on licensing of data, and data re-usability in AI models



US: model training is fair use

EU: Articles 3 and 4 of the Directive on Copyright in the Digital Single Market (DSM) 8

<https://creativecommons.org/faq/#artificial-intelligence-and-cc-licenses>

<https://creativecommons.org/2021/03/04/should-cc-licensed-content-be-used-to-train-ai-it-depends/>

Why should we be re-using data?

The weird and wonderful world of proteomics

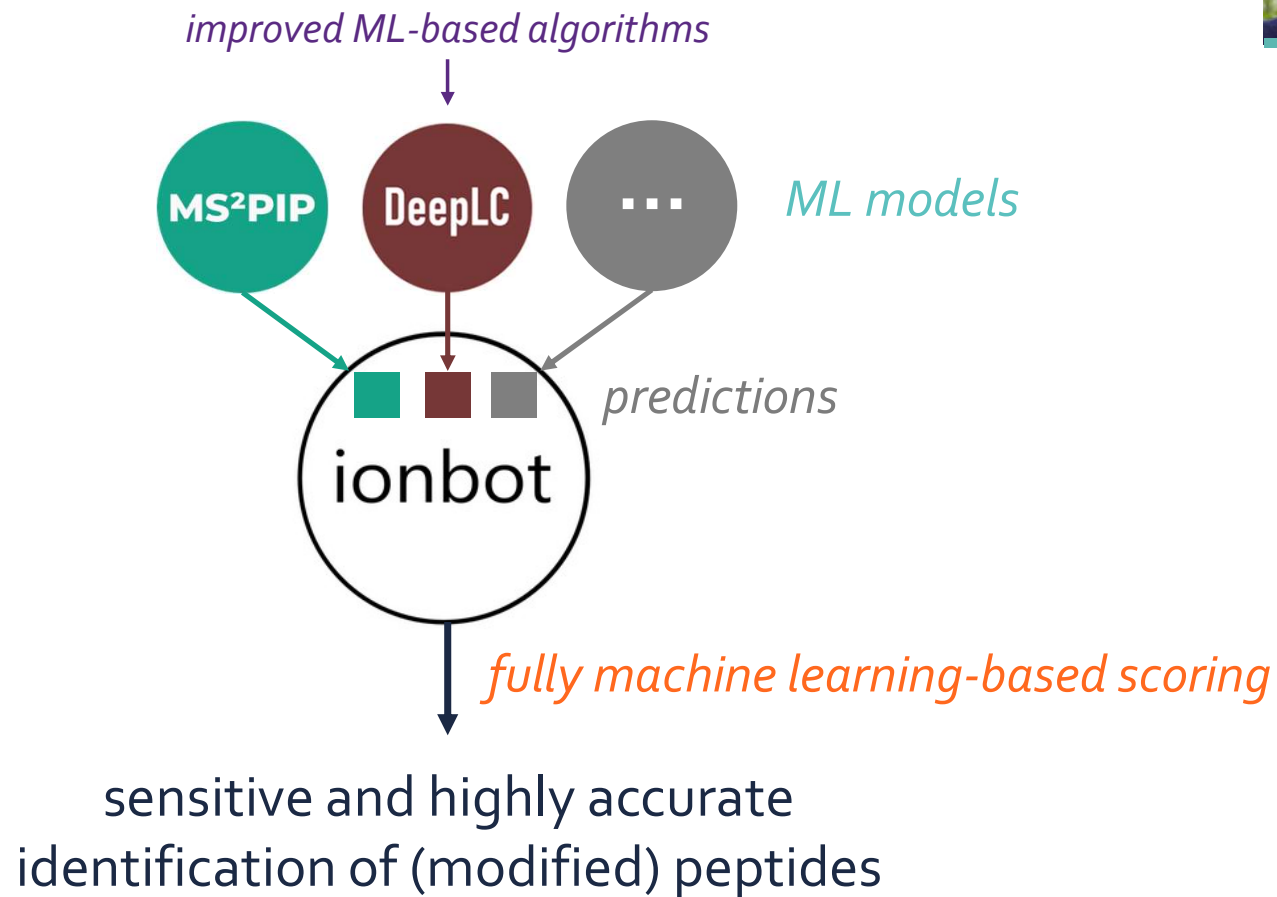
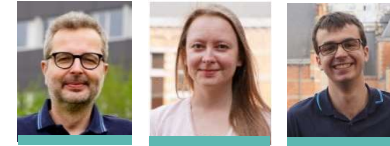
Four types of data re-use

Re-using available data to build machine learning models

Reprocessing data with new models for new insights

Repurposing large-scale data for new knowledge

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



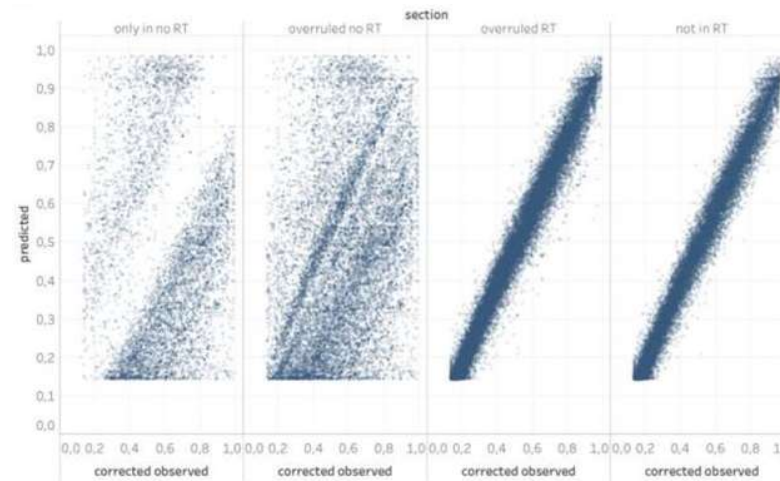
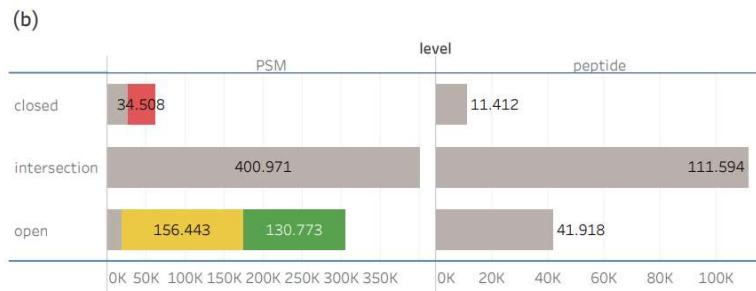
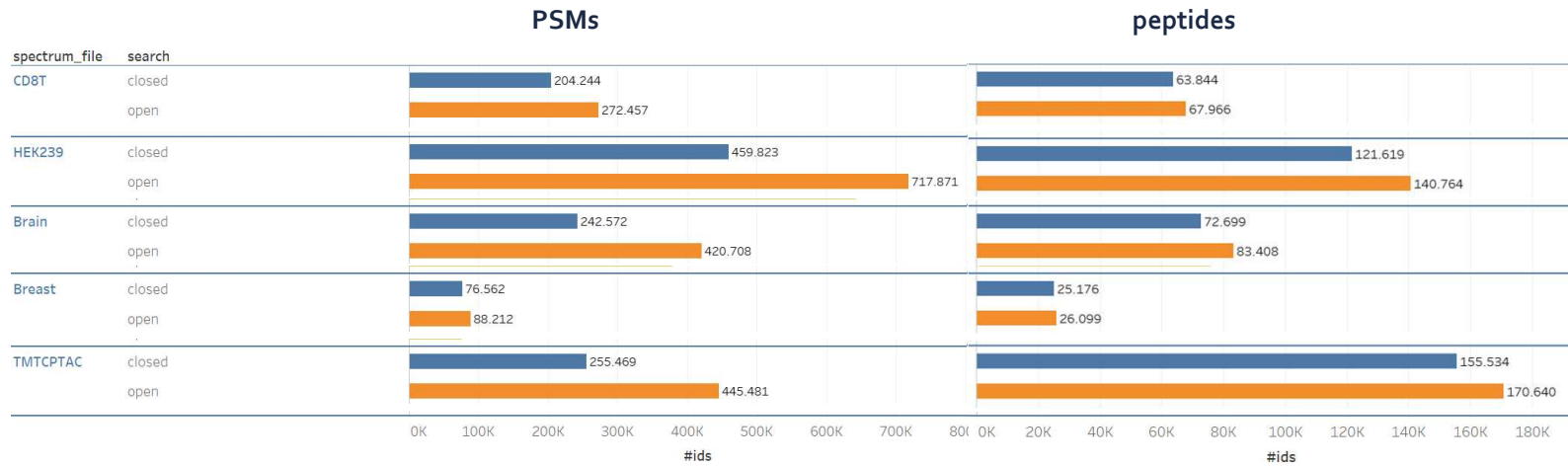
<https://ionbot.cloud>

Degroeve, <https://www.biorxiv.org/content/10.1101/2021.07.02.450686v2>



CC BY-SA 4.0

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



<https://ionbot.cloud>

Degroeve, <https://www.biorxiv.org/content/10.1101/2021.07.02.450686v2>

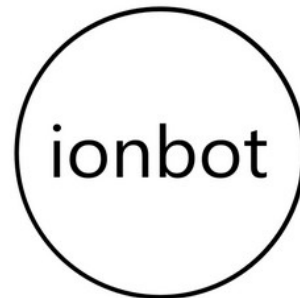


CC BY-SA 4.0

We are now running ionbot on all human and mouse spectra contained in the PRIDE database

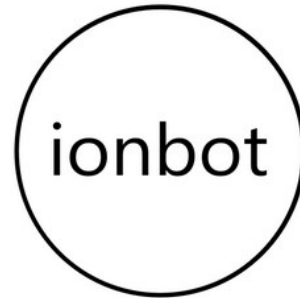
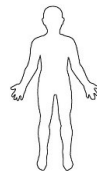


277 mouse data sets
appr. 600 million spectra
14.5 k raw files
from PRIDE



632 k peptides
16 808 proteins
(99% of SwissProt -canonical)

539 human data sets
appr. 924 million spectra
25 k raw files
from PRIDE



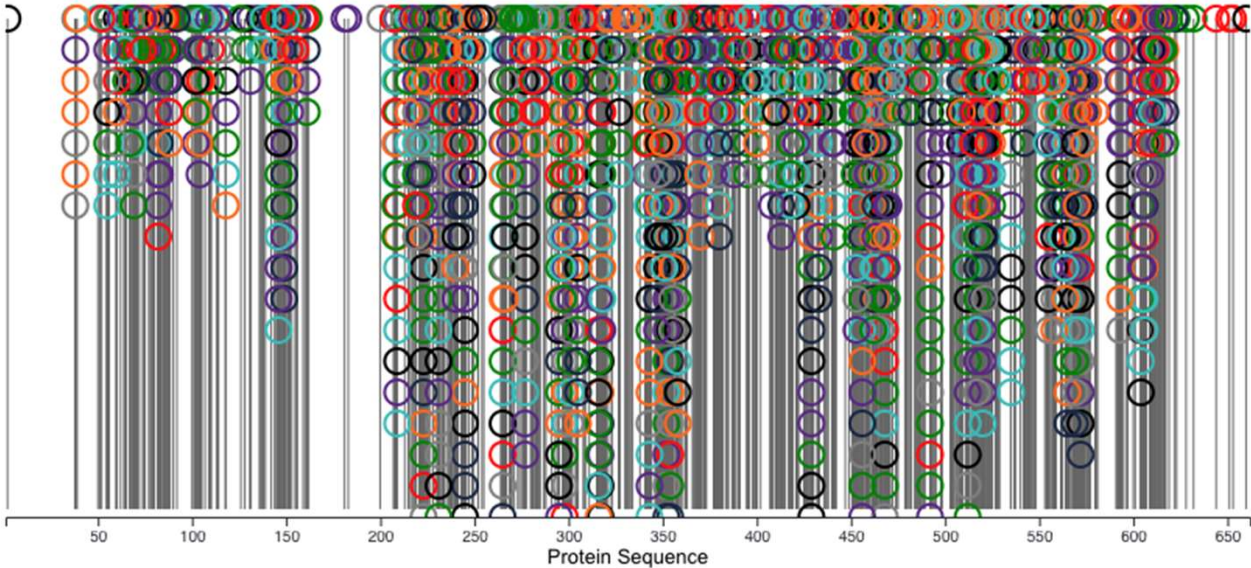
742 k peptides
20 246 proteins
(99% of SwissProt -canonical)

Proteome-wide open modification searches make it clear that proteins are really very complex molecules

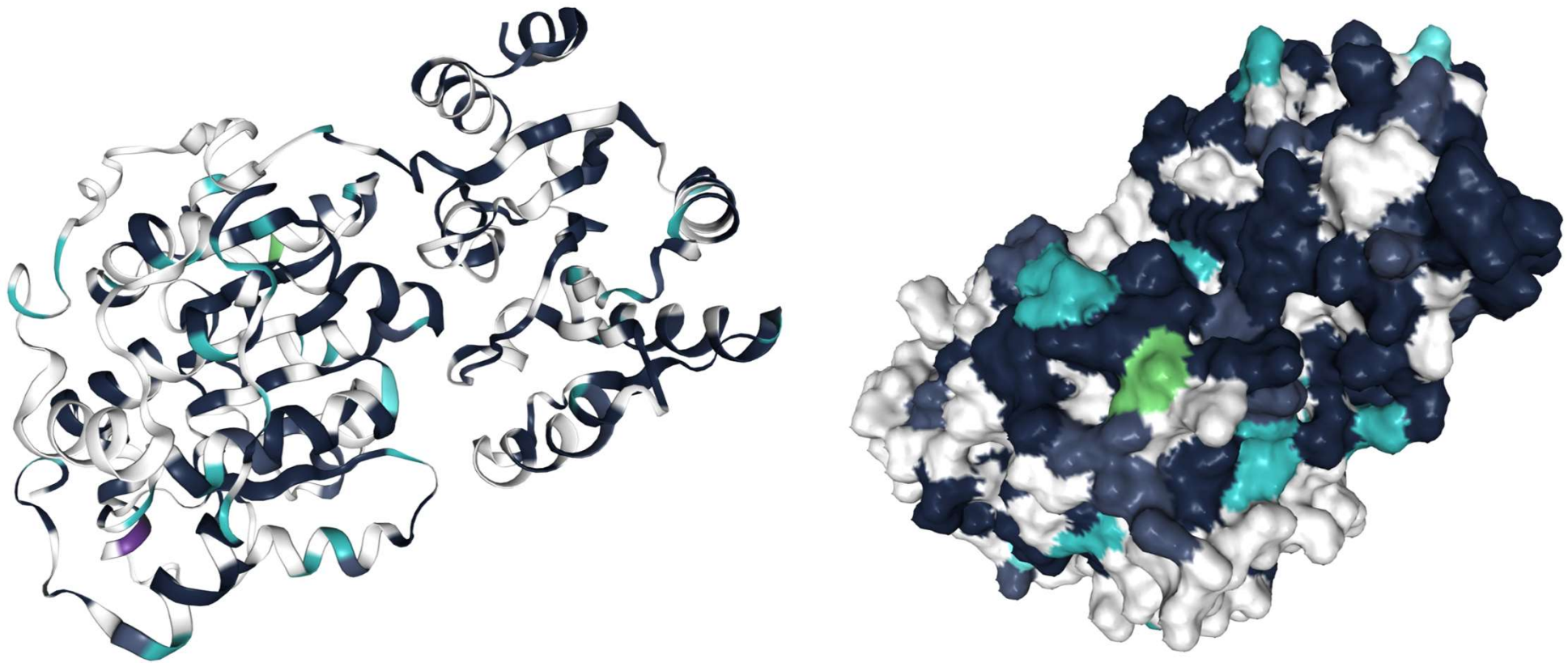


00571 Summary Peptides Structures Mutations

ATP-dependent RNA helicase DDX3X – 00571



The 3D structure view drives home the point that proteins have tremendous modification potential

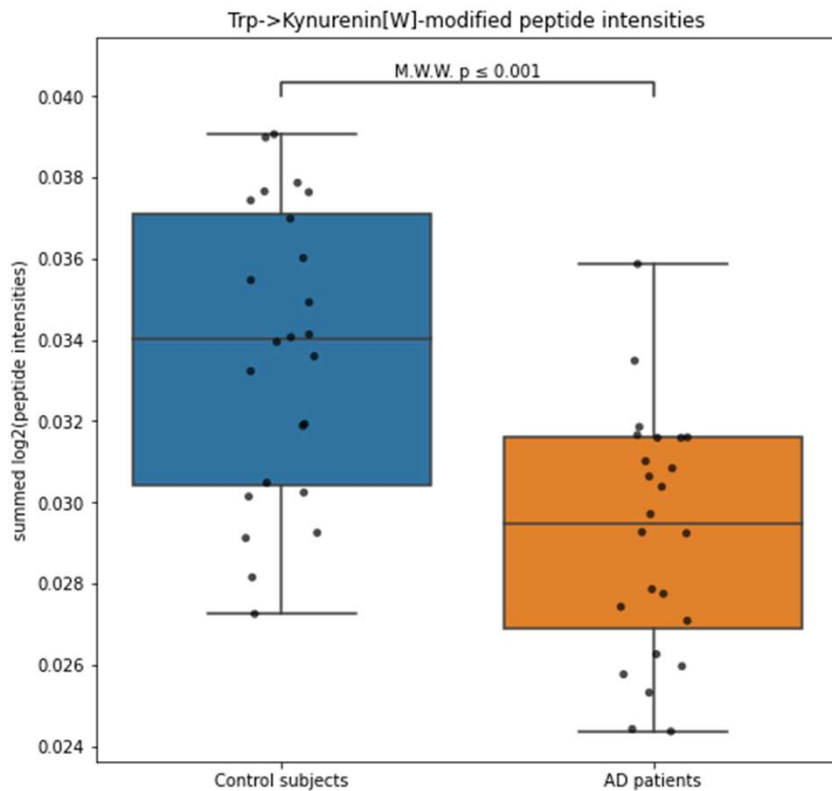


ATP-dependent RNA helicase DDX₃X – O00571 – PDB 2l4l

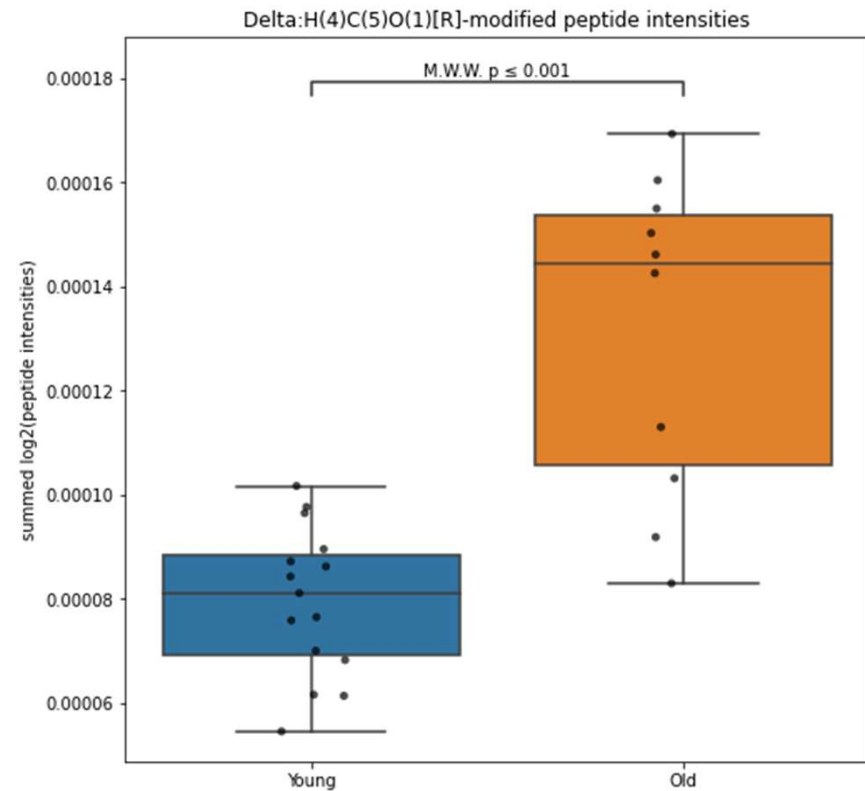


CC BY-SA 4.0

We started analysing specific data sets for differential PTMs, and possible sentinel modifications can readily be found



bio-availability of serotonin and kynurenin is reduced in urine and serum of AD



methylglyoxal is related to blood glucose and oxidation, and both are dysregulated in ageing

Why should we be re-using data?

The weird and wonderful world of proteomics

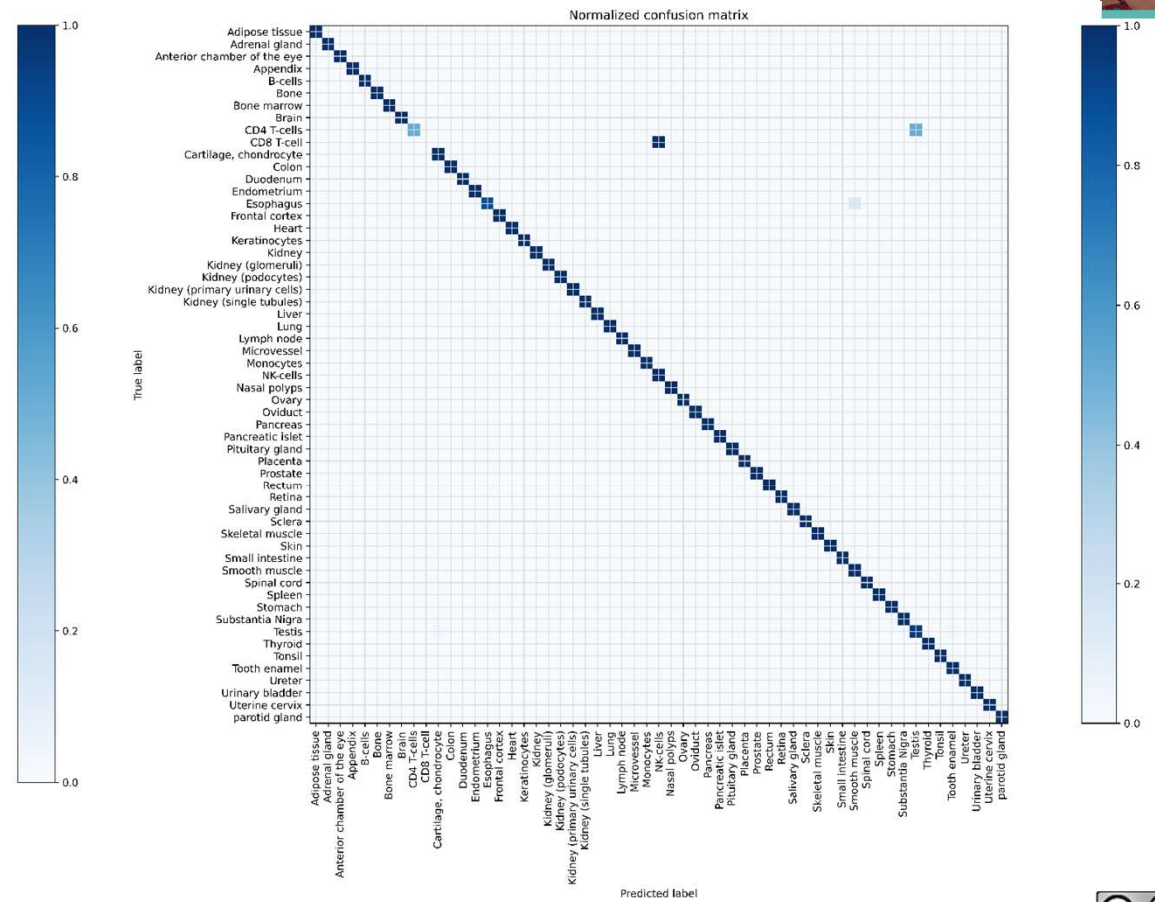
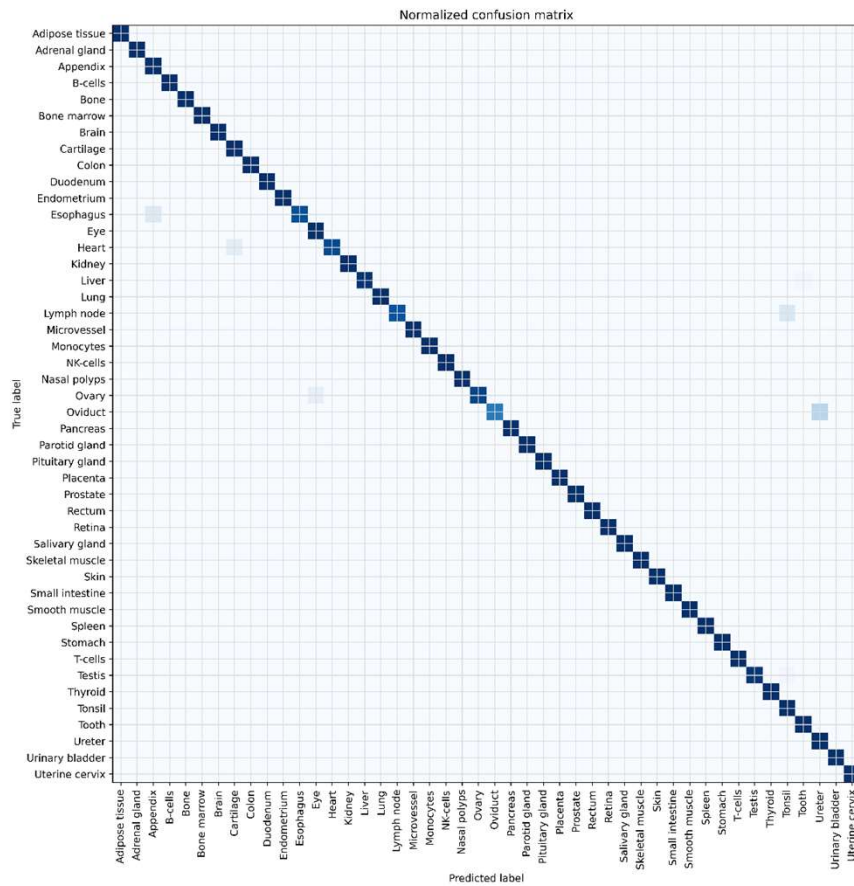
Four types of data re-use

Re-using available data to build machine learning models

Reprocessing data with new models for new insights

Repurposing large-scale data for new knowledge

We built an AI model that predicts tissue or cell type purely based on the proteome, at 98% accuracy



Although the model is only trained on healthy tissue data, it can provide insight into changes in cancerous samples

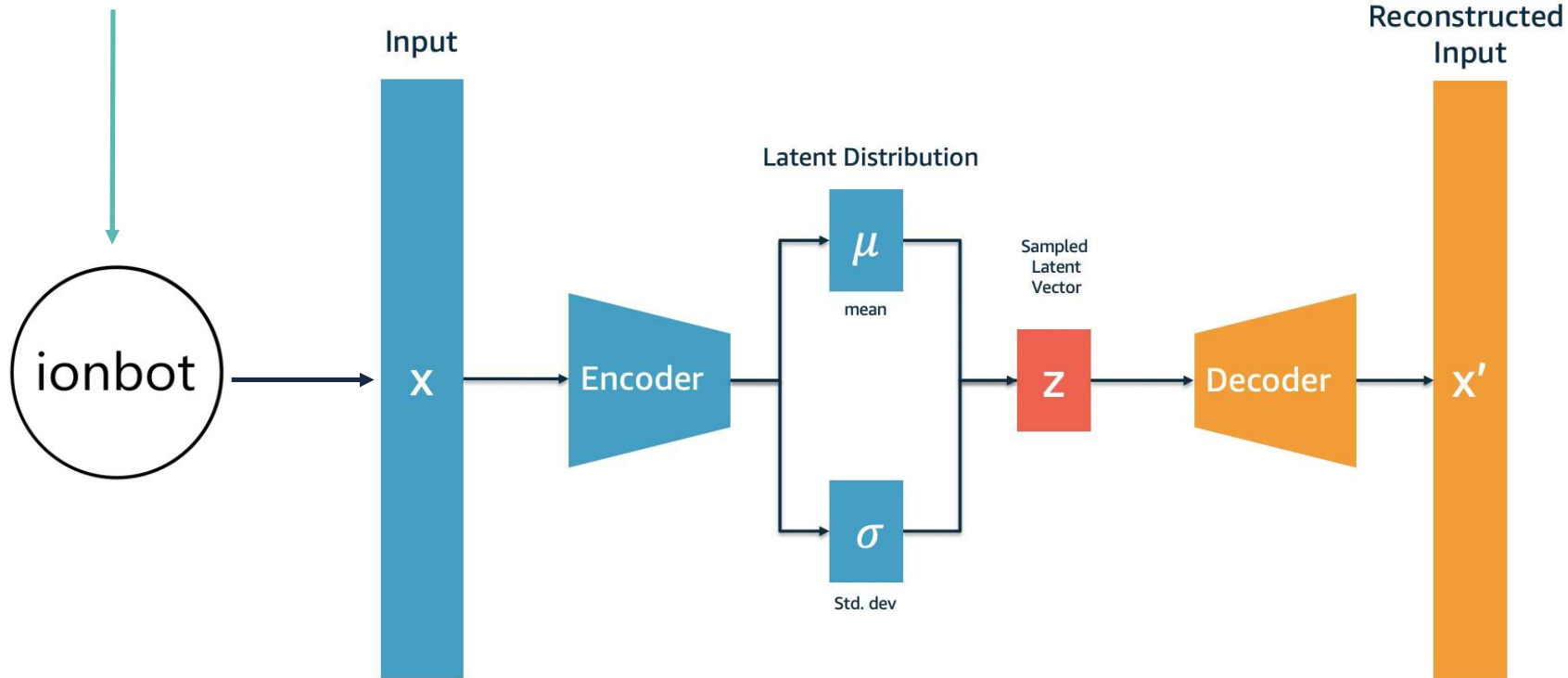


Lung was predicted with 53.6% probability

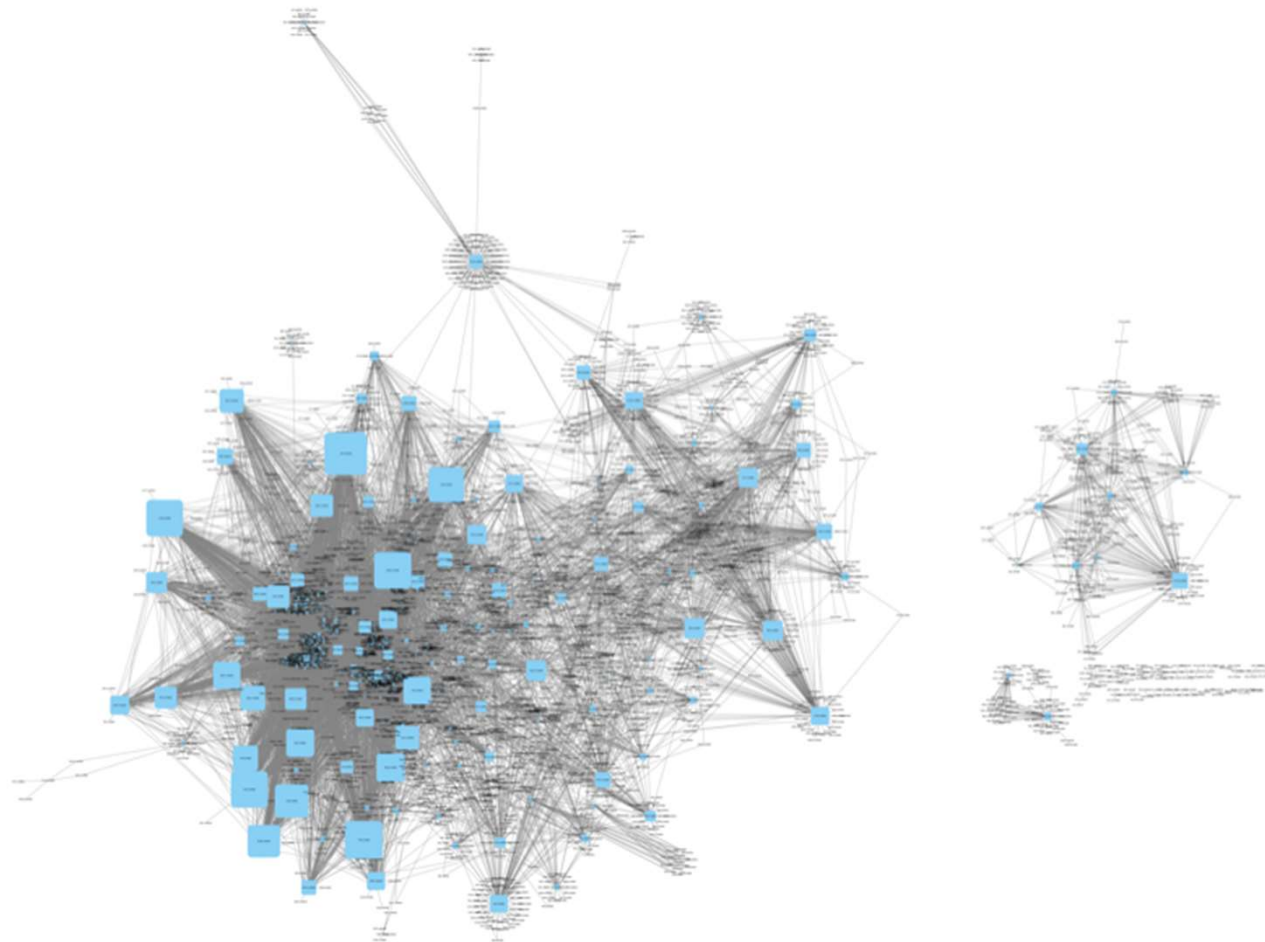


Protein	NSAF	Protein name	Observation (HPA, Bgee, Pubmed)
O75891	NaN	Cytosolic 10-formyltetrahydrofolate dehydrogenase	Low to no expression in lung
A6NMZ7	0.0001364	Collagen alpha-6(VI) chain	Fibroblast specific
P23141	0.001089	Liver carboxylesterase 1	Enriched in lung
P13796	0.0004924	Plastin-2	Immune system specific, observed in lung cancer
Q96RW7	NaN	Hemicentin-1	Fibroblast specific
P35625	NaN	Metalloproteinase inhibitor 3	Fibroblast and ECM specific
P04114	0.0004059	Apolipoprotein B-100	Liver specific, higher levels observed in long obstruction and cancer

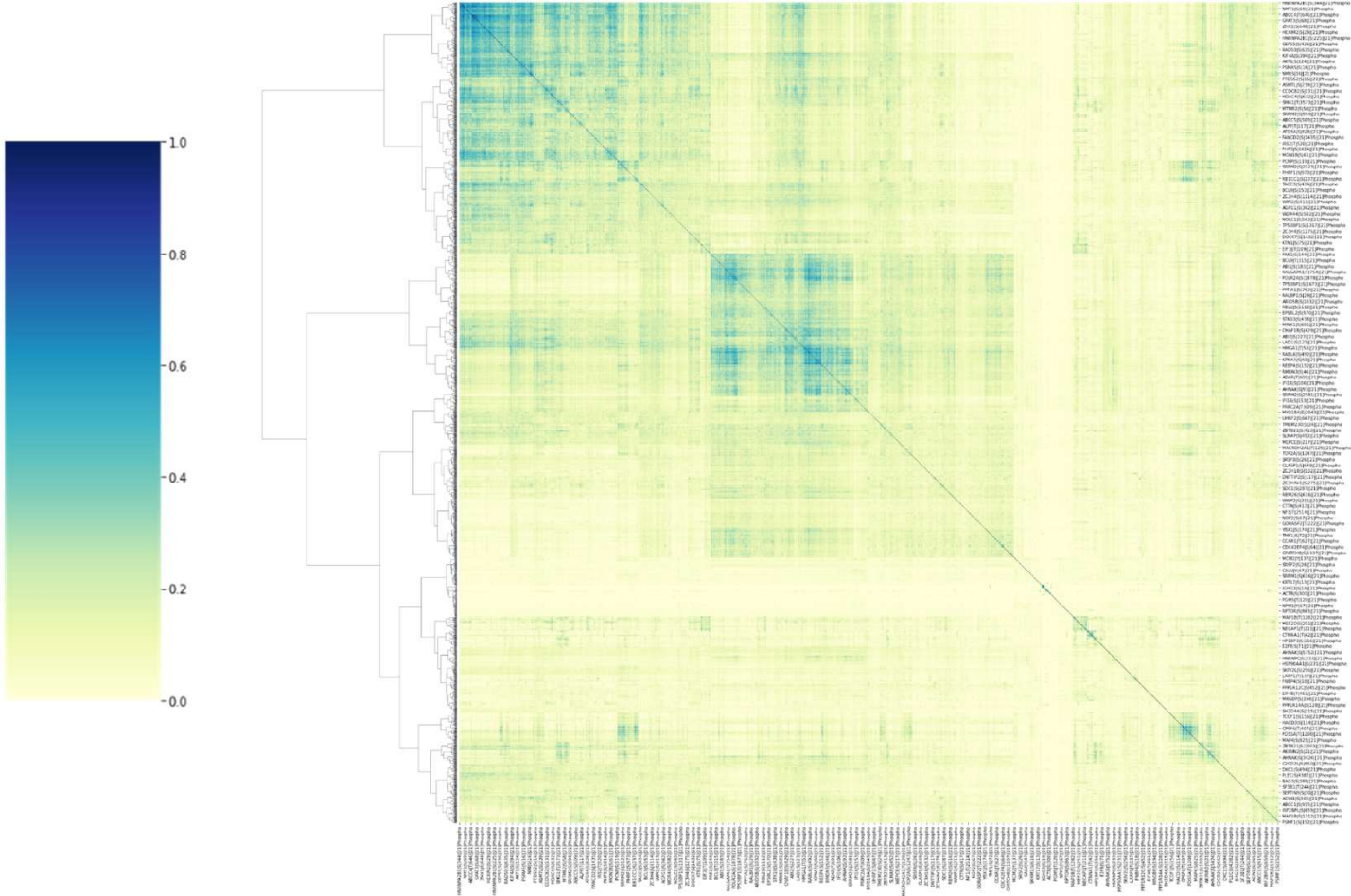
Mikaela Koutrouli, from the LJ Jensen Lab, visited our group to find protein associations from these data



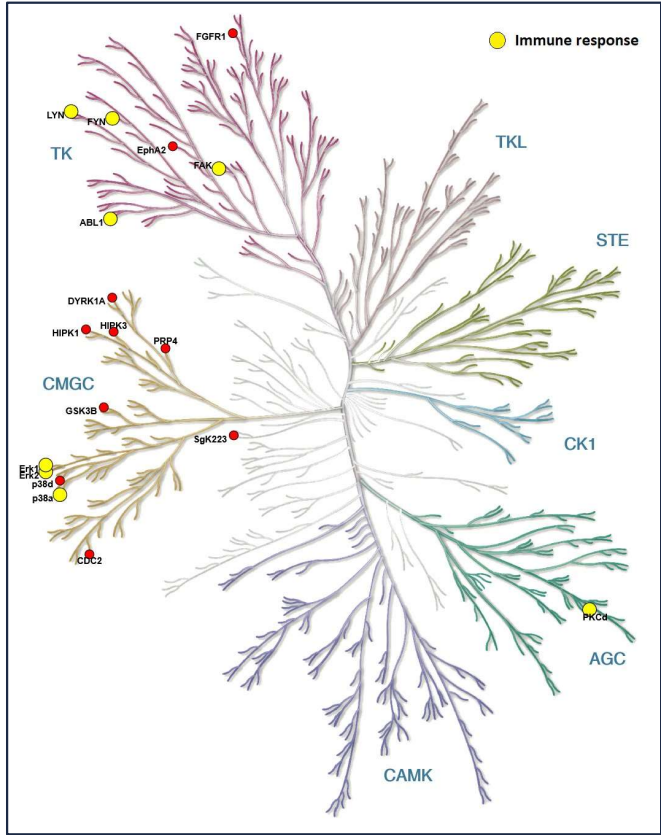
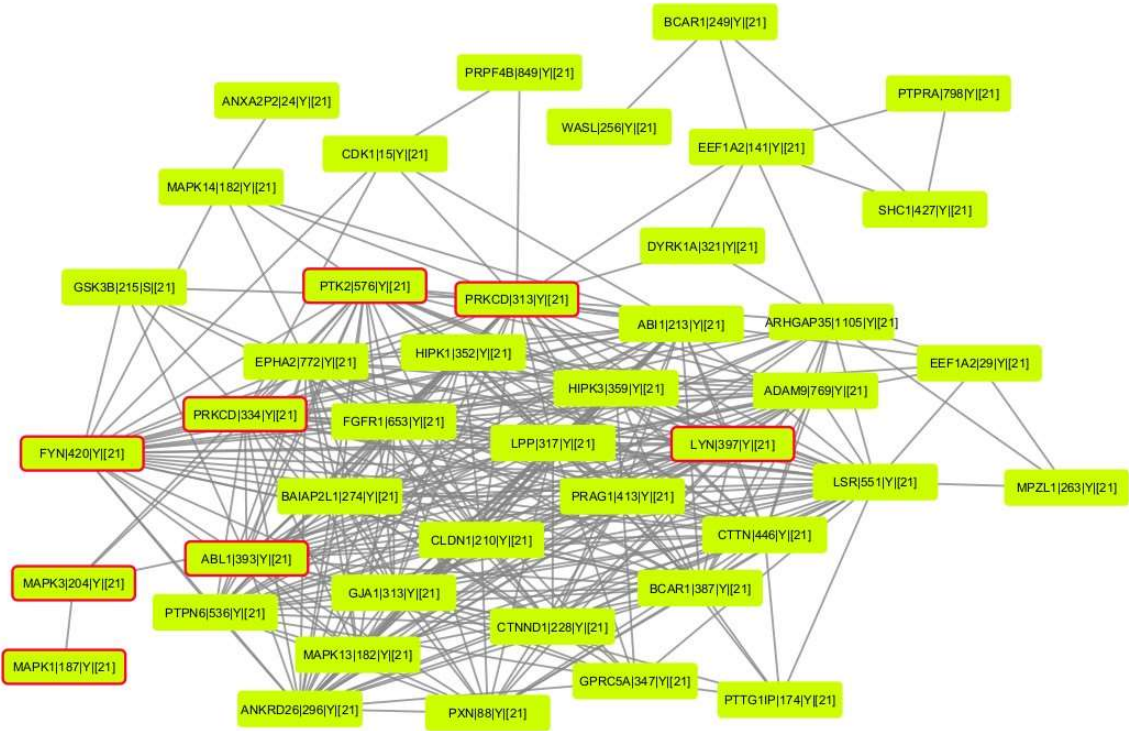
These associations have already been used to uncover the clientele of protein chaperones (prof. Joost Schymkowitz)



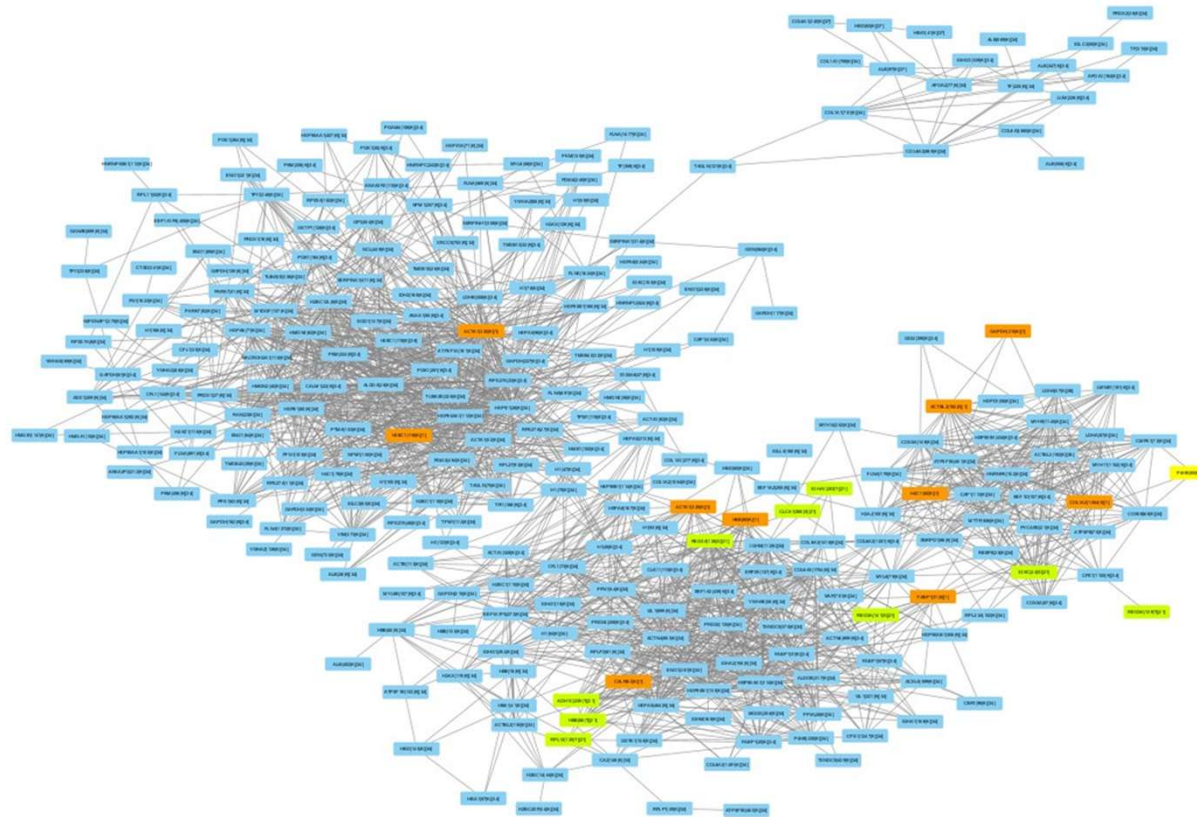
A VAE map of proteome-wide phosphosite co-occurrence reveals some interesting patterns, including kinase clusters



These phospho clusters contain connected components, here shown for tyrosine kinases and their key targets

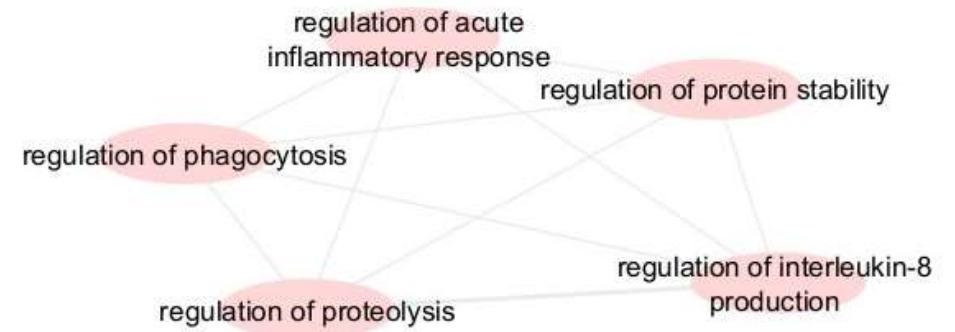
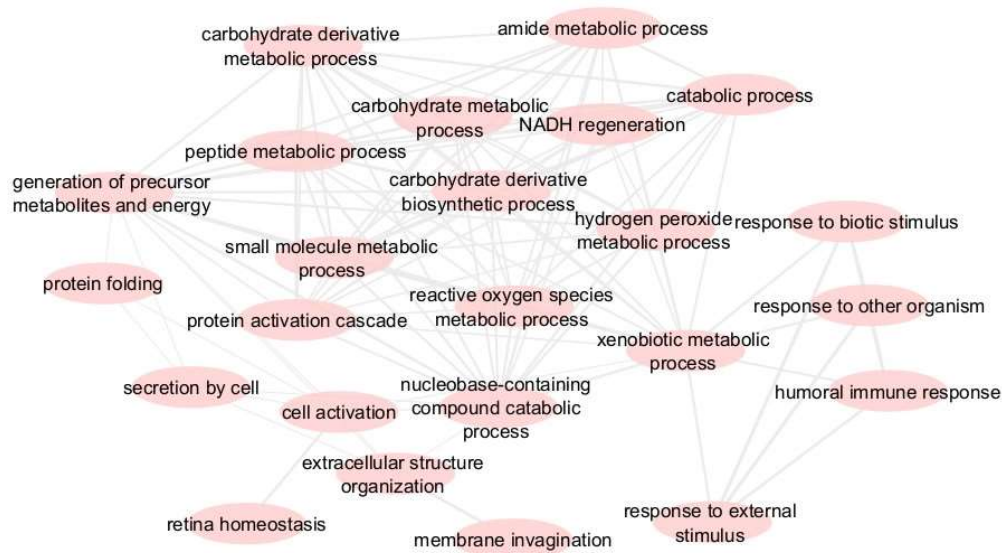


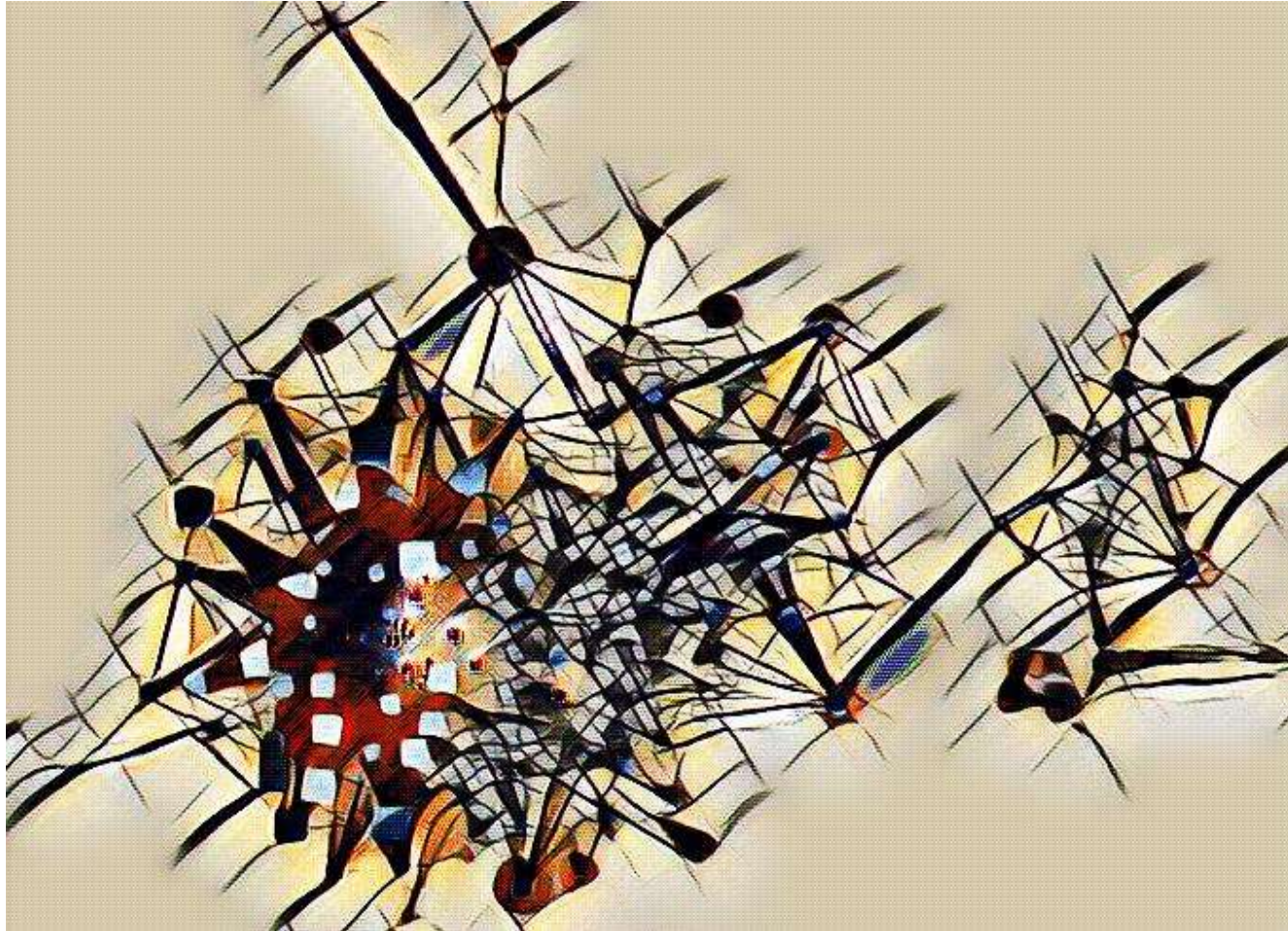
We can also analyse several modifications together, here methylation, acetylation, and phosphorylation



- Acetyl
- Methyl
- Phospho

A GO analysis of this giant component reveals strong links with innate immunity and its metabolic support





<https://www2.lunapic.com/editor/?action=kandinsky>



Comp
omics



www.compomics.com
compomics.github.io