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

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

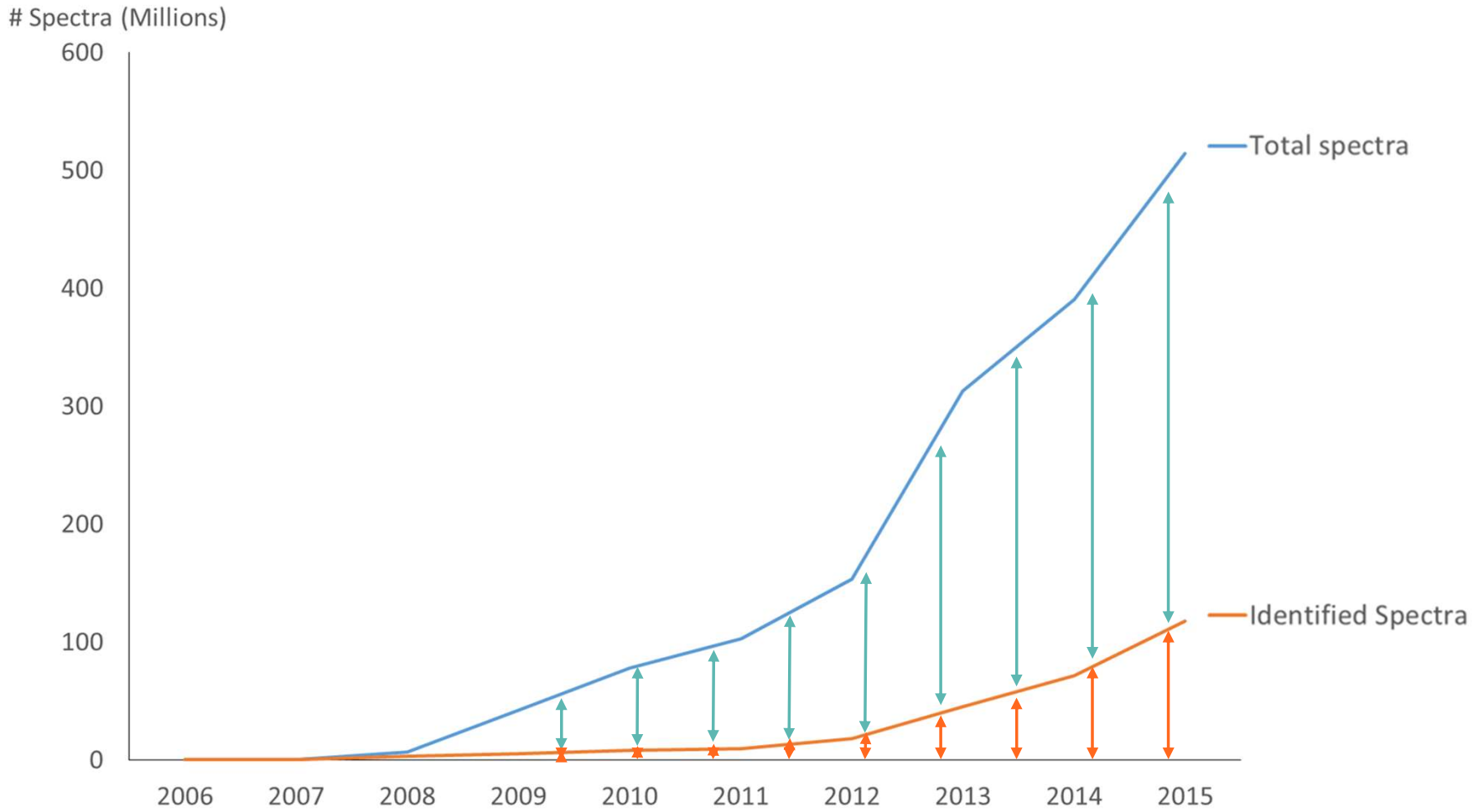
Four types of data re-use

Re-using available data to build machine learning models

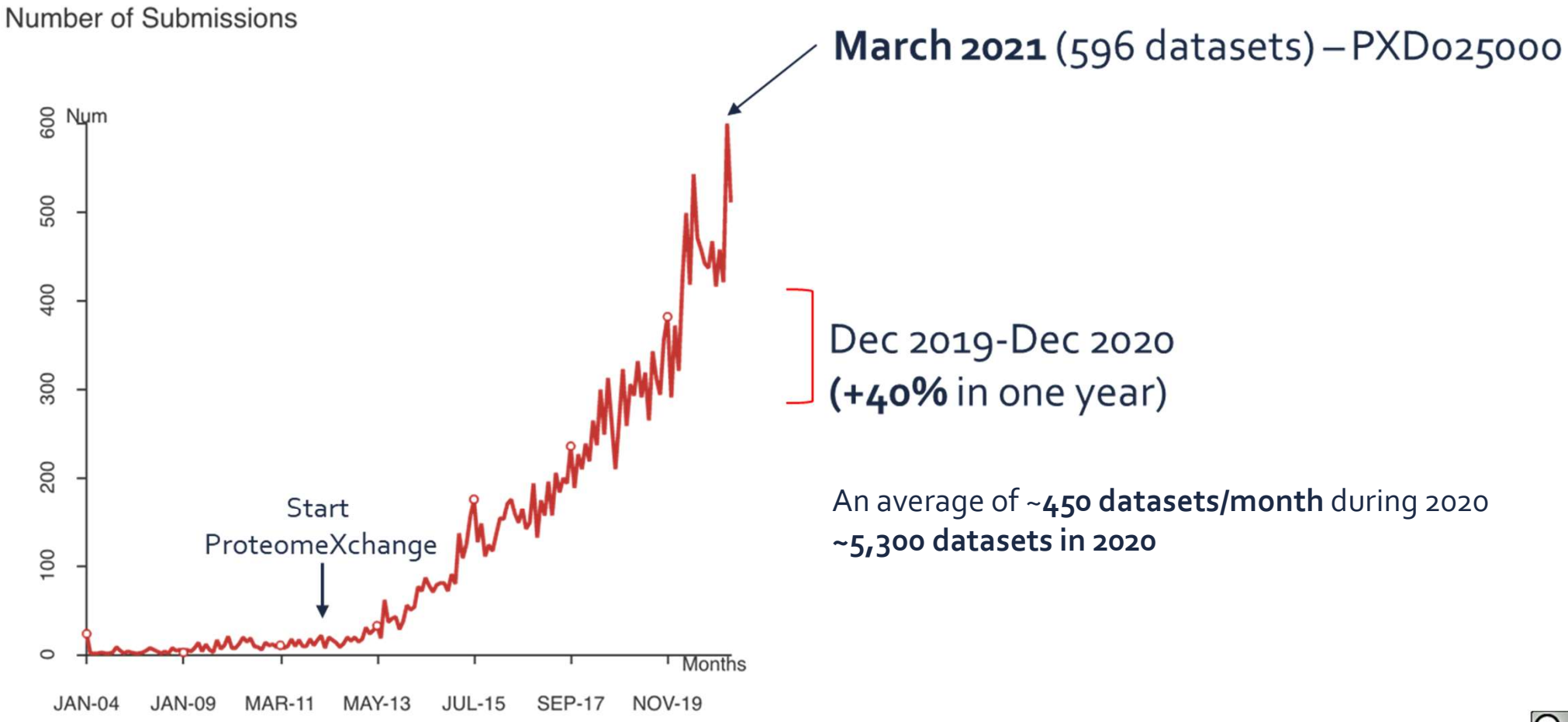
Reprocessing data with new models for new insights

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

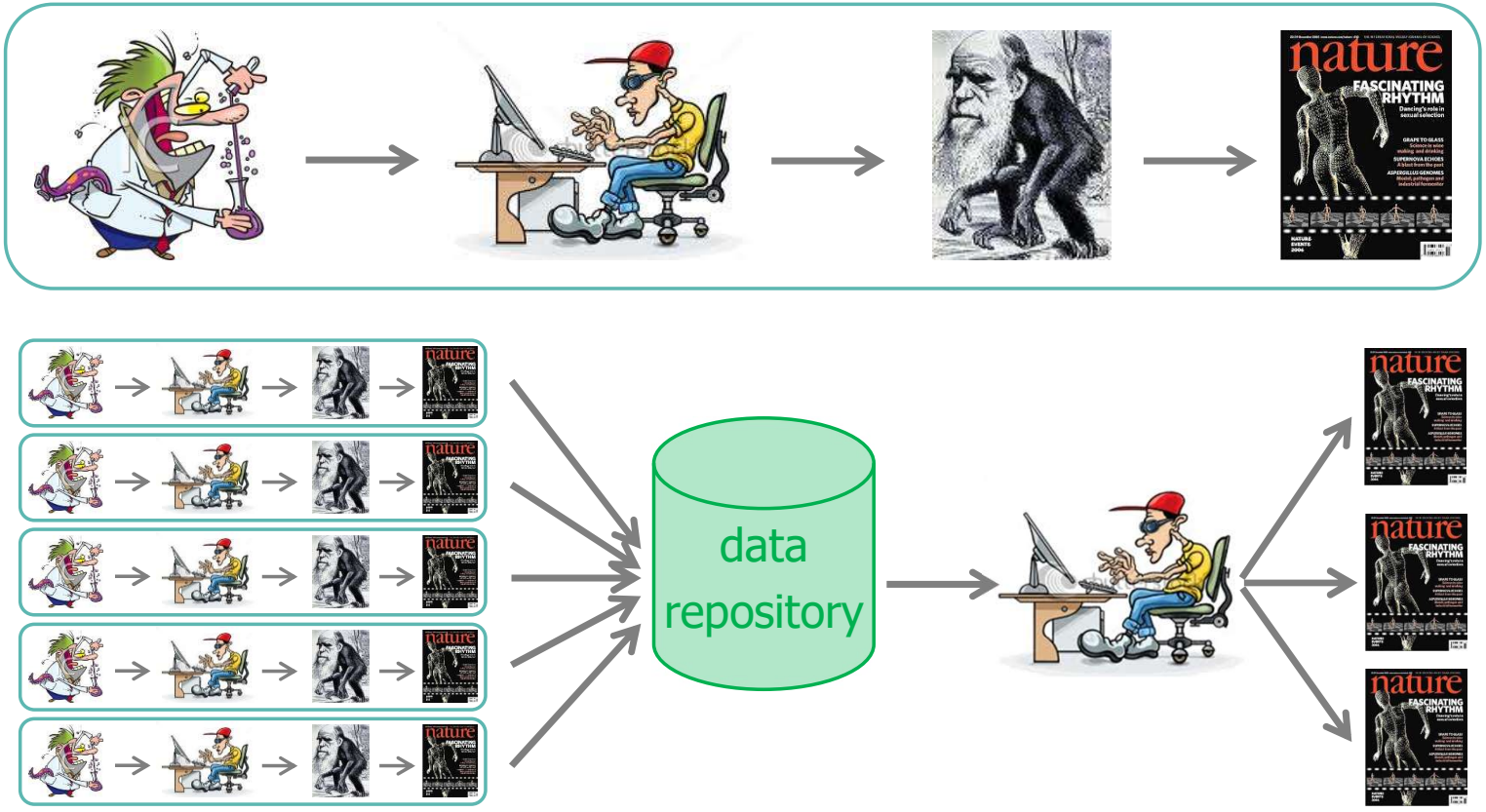


Most of our data is also high throughput, meaning there is lots of data available!

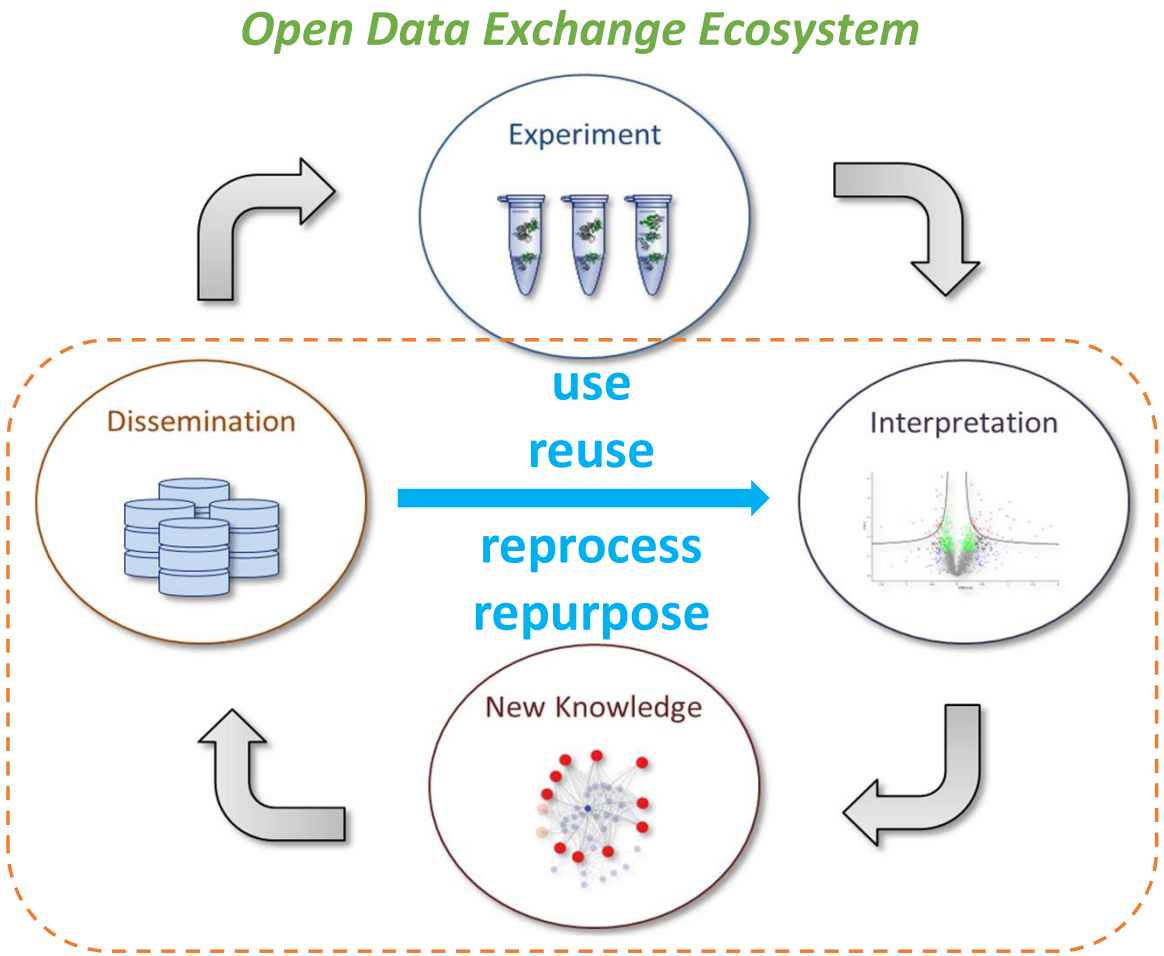


Slide courtesy of Dr. Juan Antonio Vizcaíno, Proteomics Team Leader, EMBL-EBI

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



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

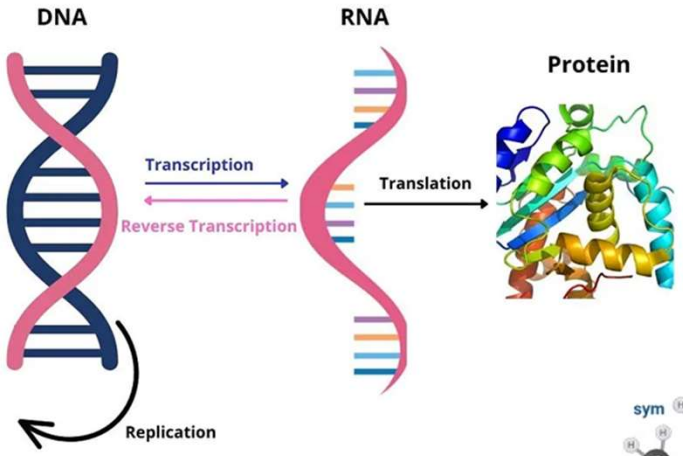
Re-using available data to build machine learning models

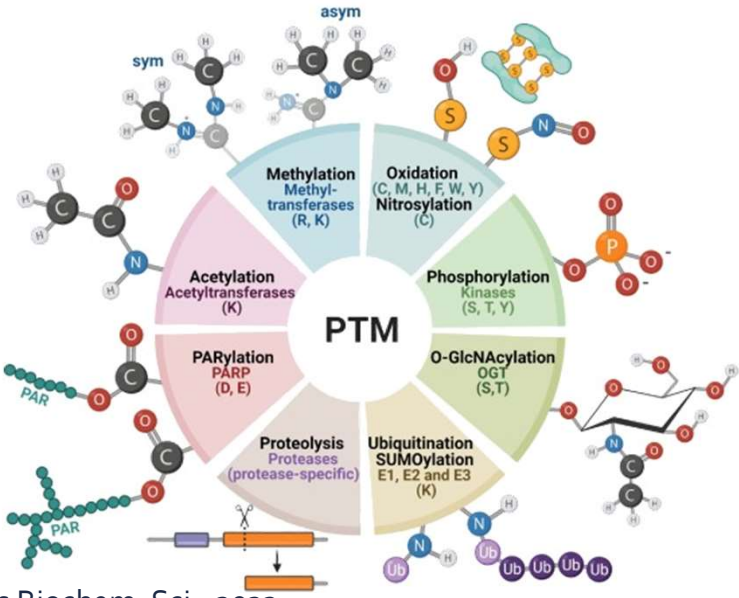
Reprocessing data with new models for new insights

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

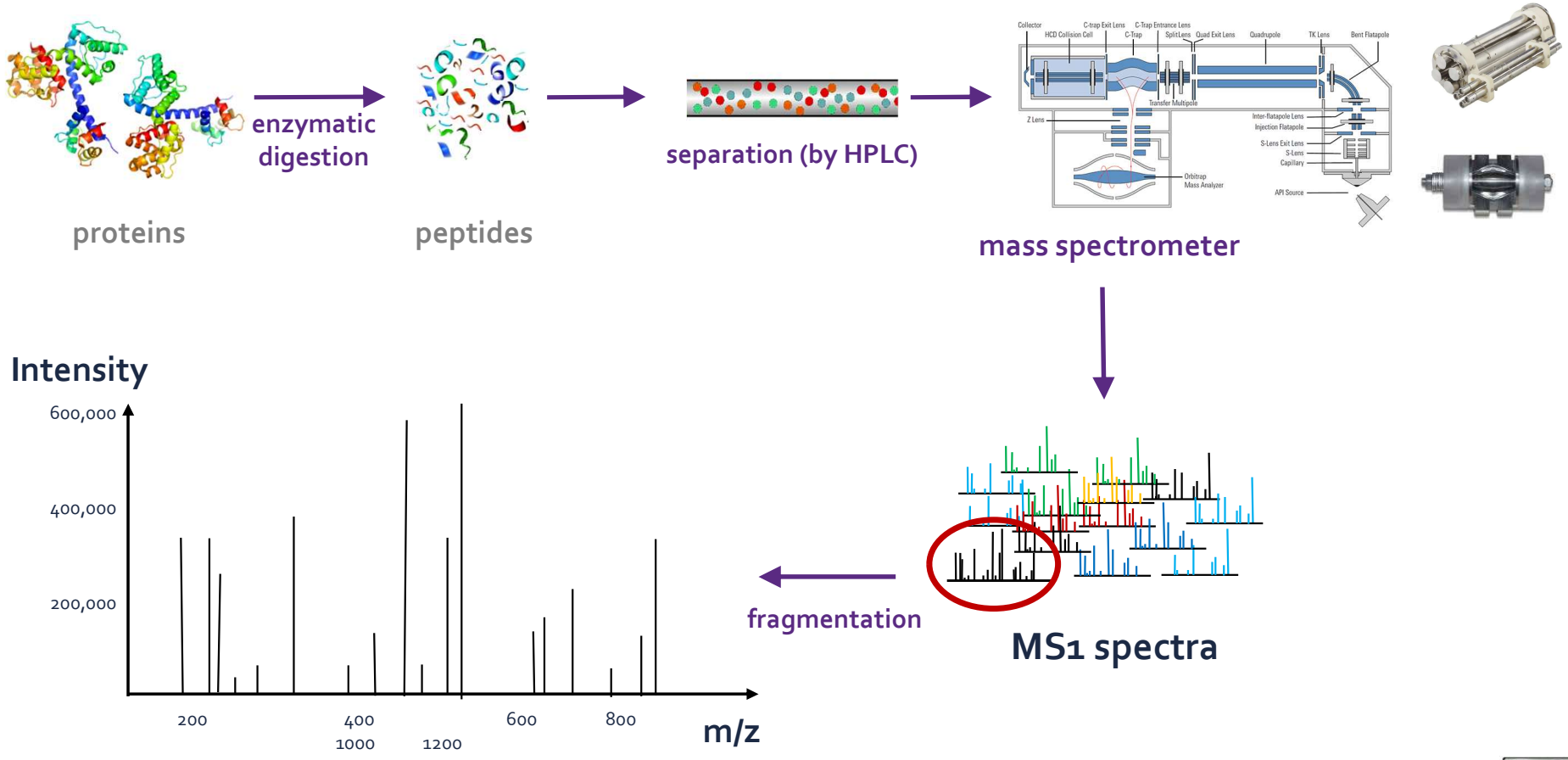
Gael McGill
https://gaelmcgill.artstation.com/projects/PmoJL1



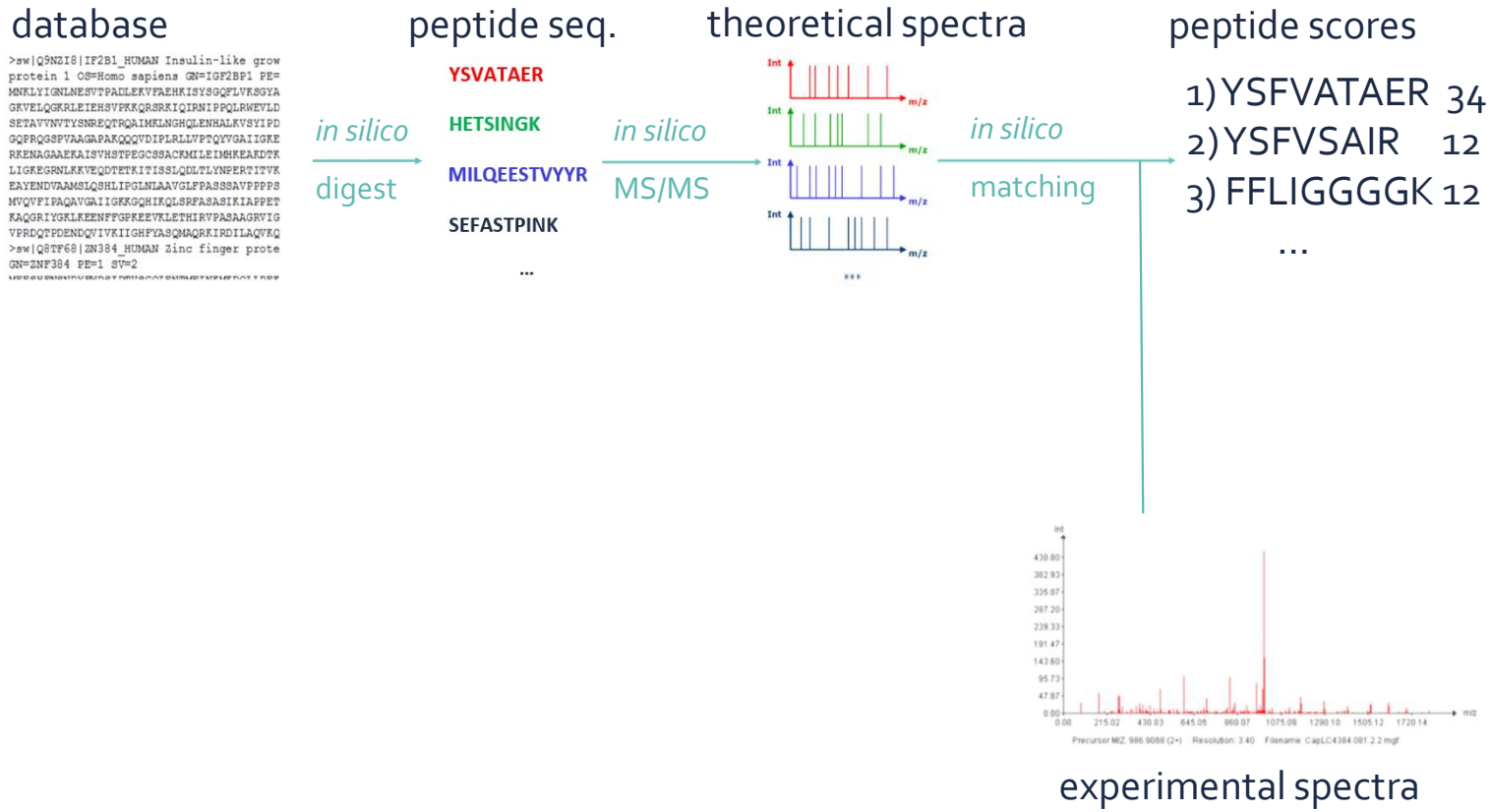


Sternburg, Trends Biochem. Sci., 2022

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



Verheggen, Adv Exp Med Biol, 2016

Why should we be re-using data?

The weird and wonderful world of proteomics

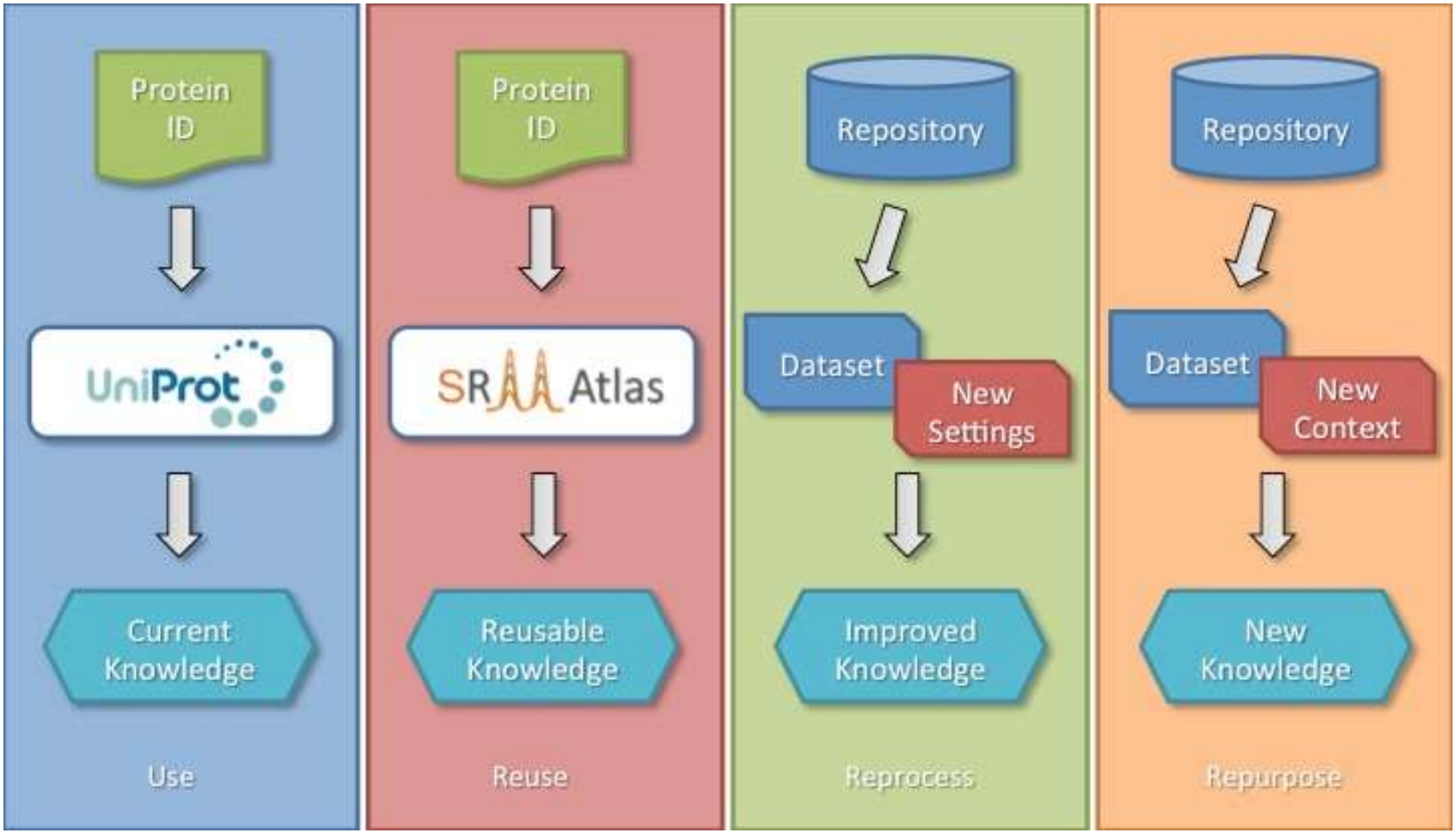
Four types of data re-use

Re-using available data to build machine learning models

Reprocessing data with new models for new insights

Repurposing large-scale data for new knowledge

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



Vaudel, Proteomics, 2016

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!

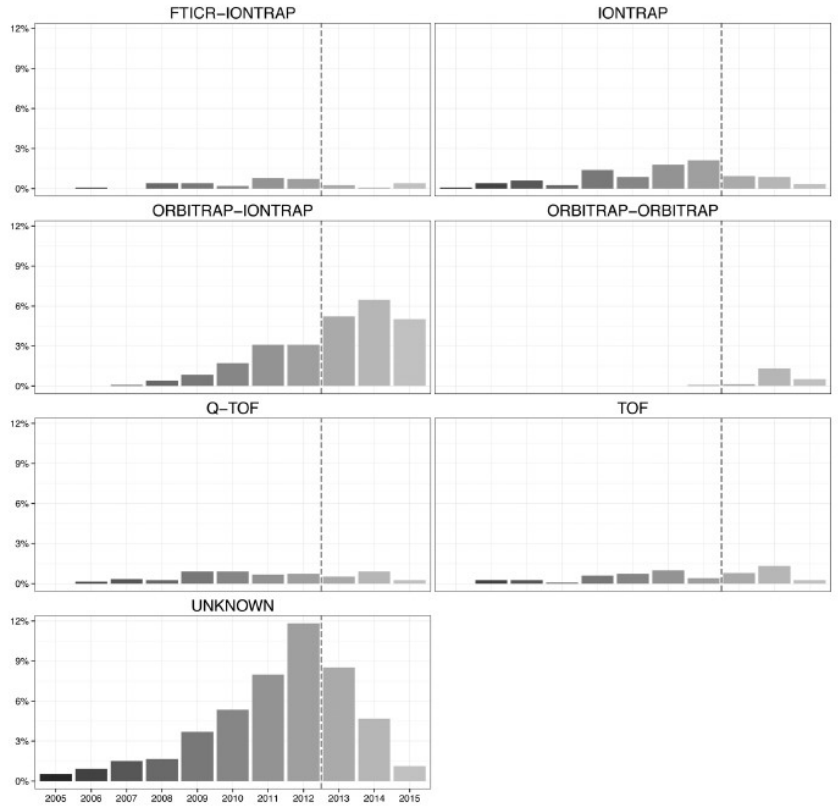
Look what I found in here!

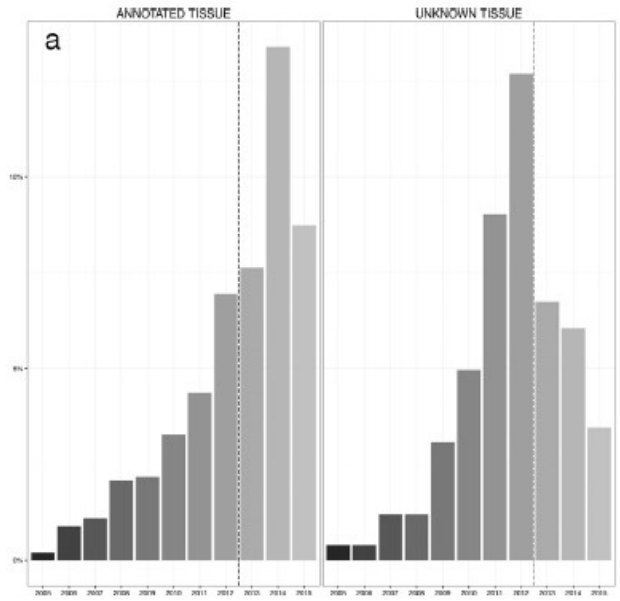
Your data is so useful!

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

https://hero.fandom.com/wiki/Good_Cop

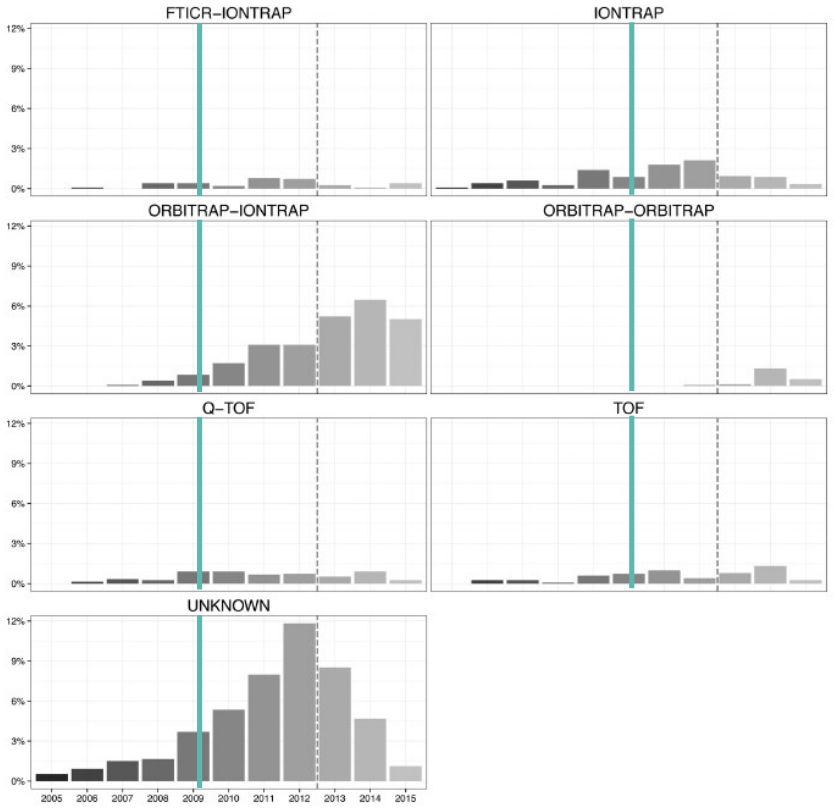
Metadata is often the key issue, as it requires manual work

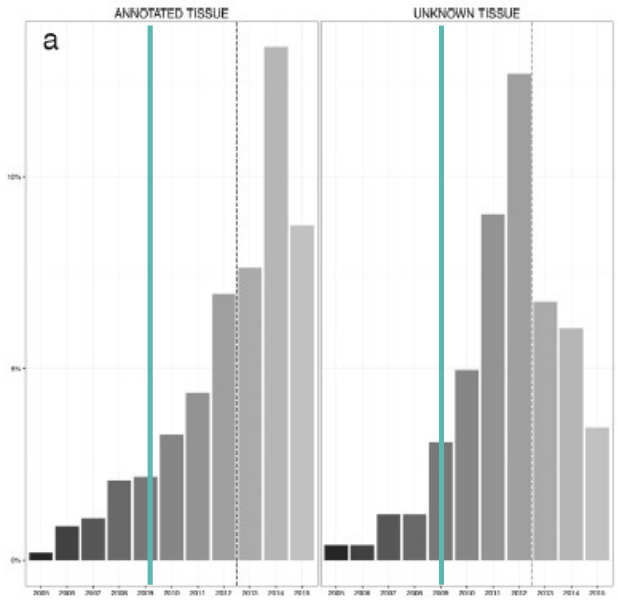




Verheggen, EuPA Open Proteomics, 2015

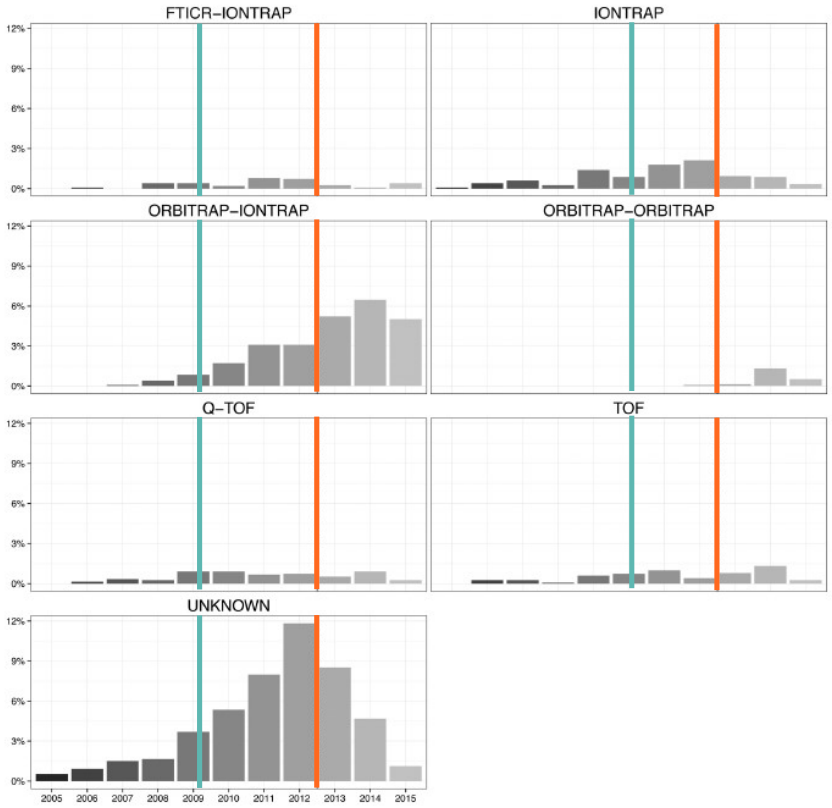
Even user-friendly submission tools cannot correct for a lack of elementary motivation

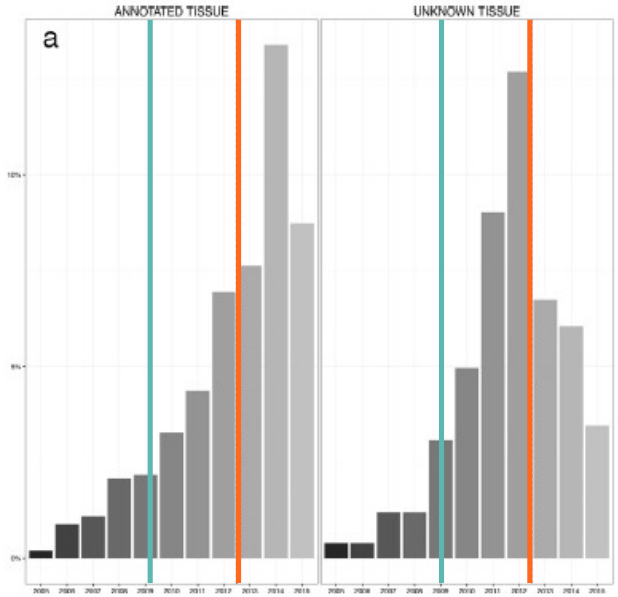




Verheggen, EuPA Open Proteomics, 2015

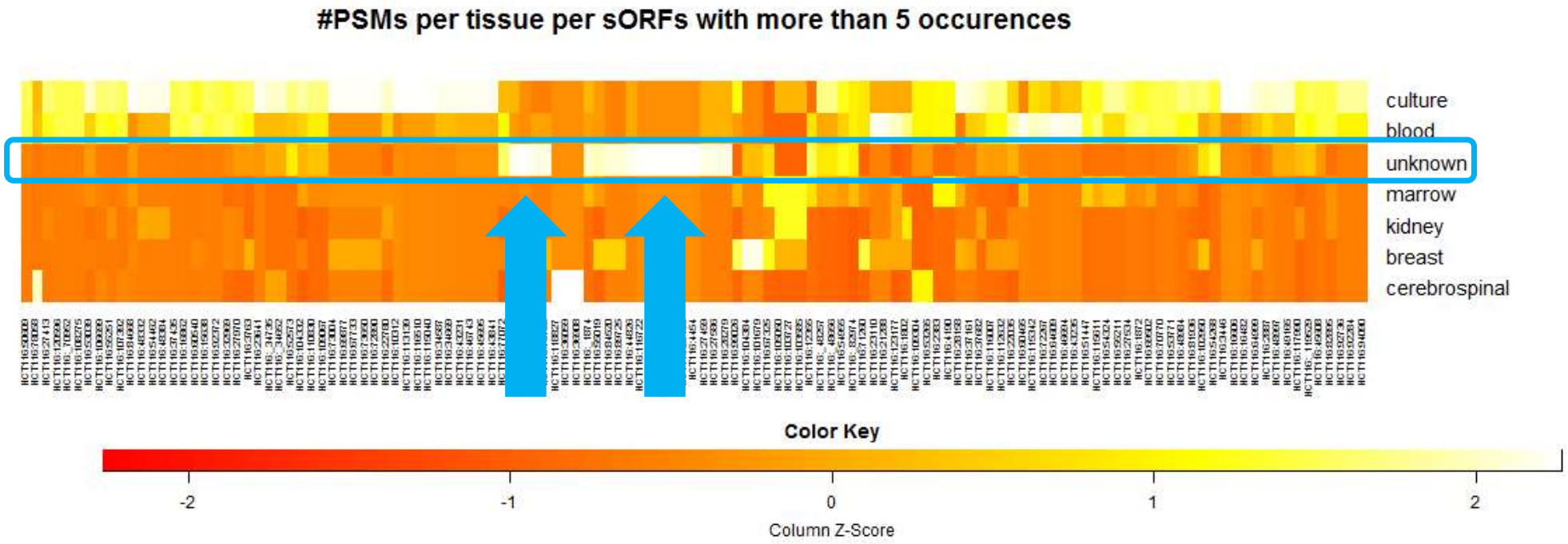
Manual curation of submissions, equivalent to restrictive policing, does help





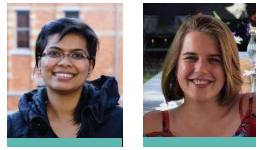
Verheggen, EuPA Open Proteomics, 2015

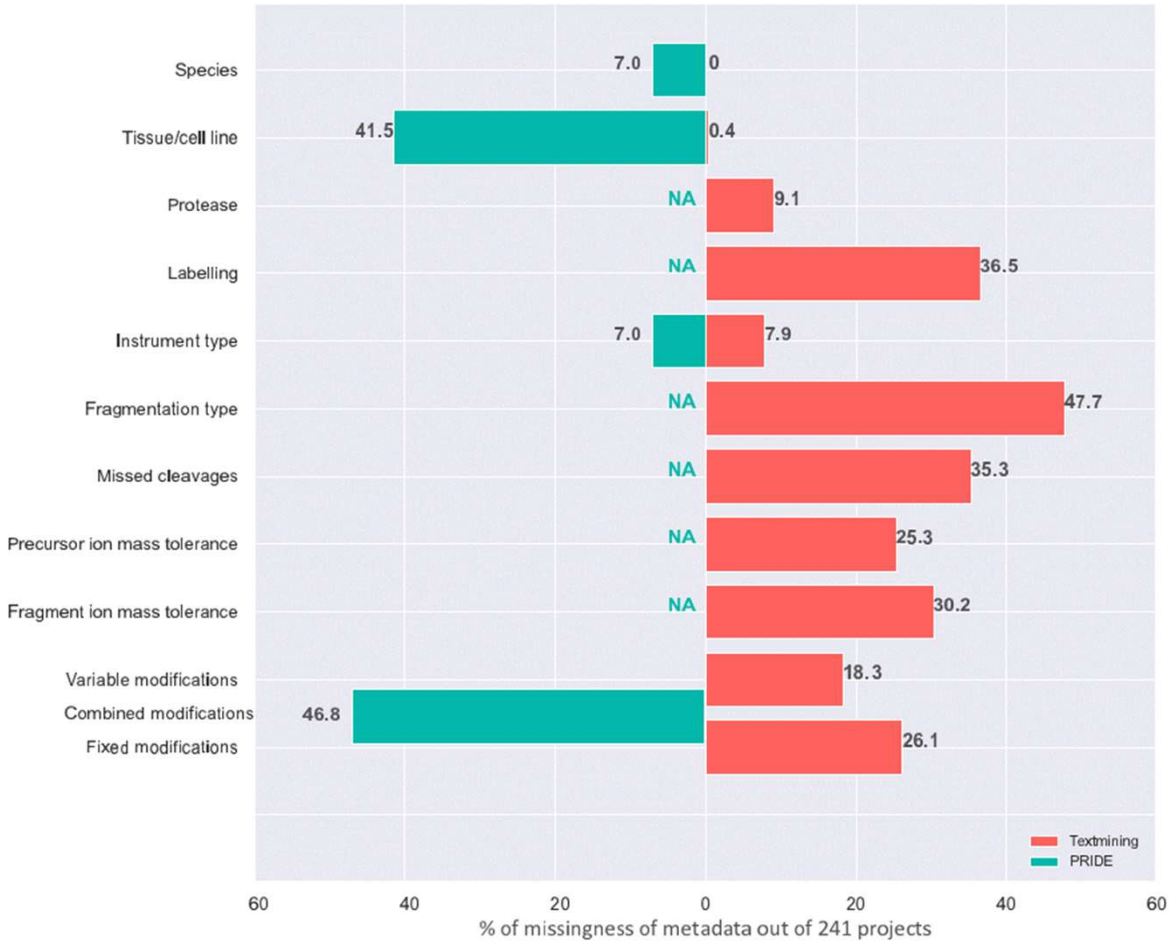
Missing metadata becomes pretty annoying when people successfully re-use your data



Olexiouk, NAR, 2016

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





Claeys, Nature Communications, 2023

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)

nature

Proteomics Standards Initiative

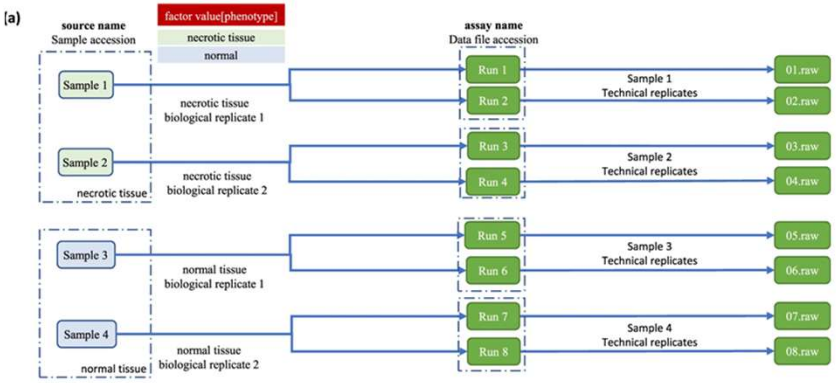
efo

CELL ONTOLOGY

OLS

PRIDE
PRoteomics IDEntifications Database

UNIMOD protein modifications for mass spectrometry



(b)

	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

Claeys, Nature Communications, 2023

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

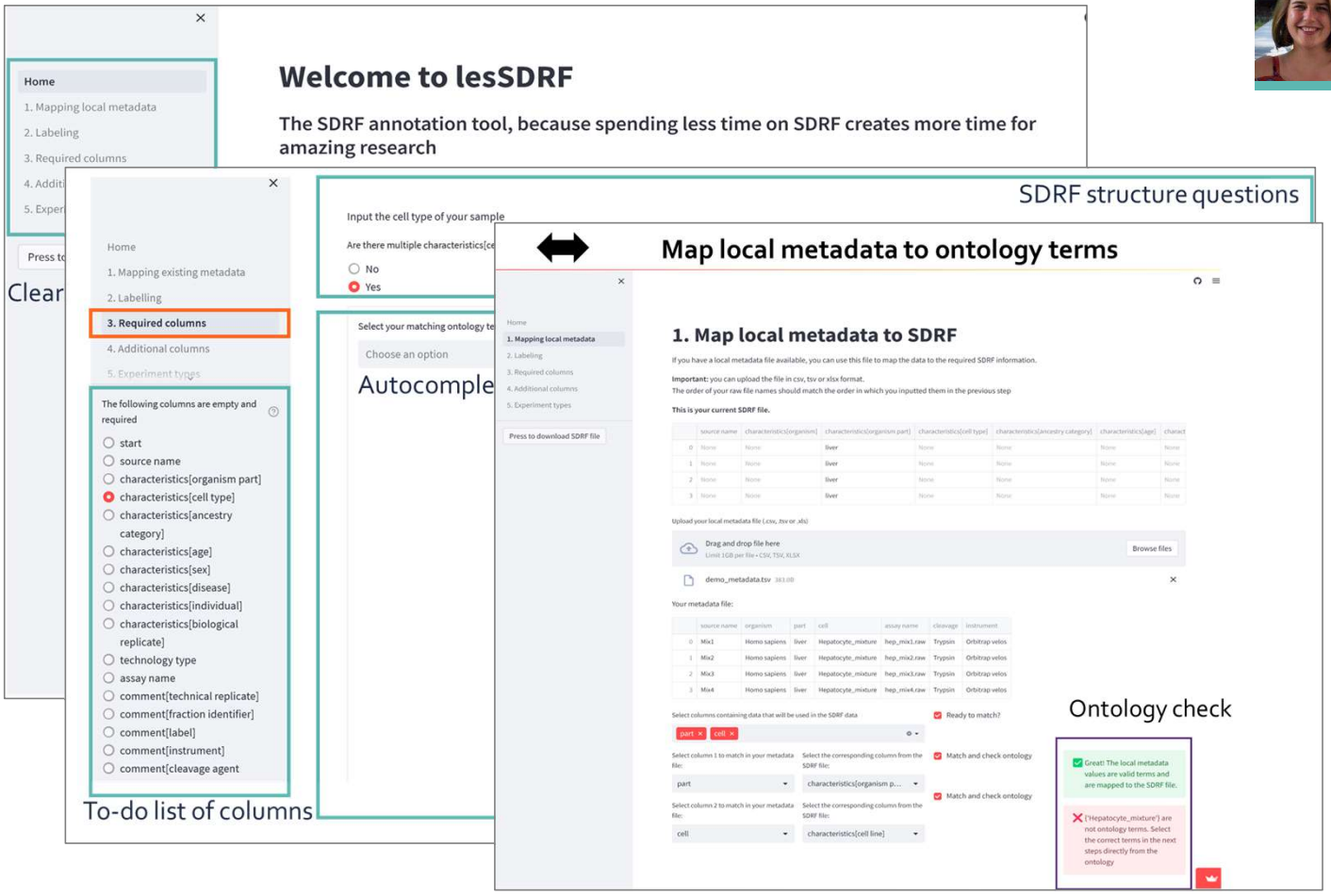
Ontologies

SDRF guidelines

Local metadata

lesSDRF

freely available
open source
no install required

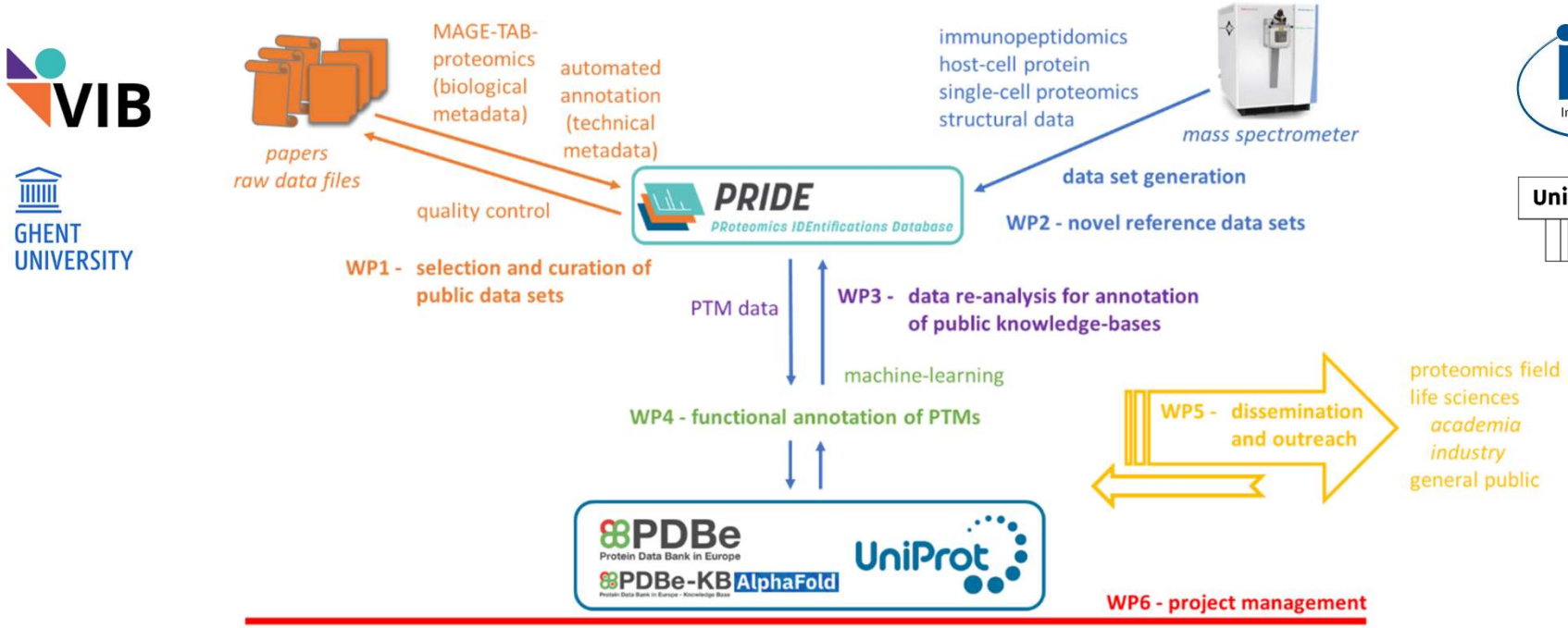


https://github.com/compomics/lesSDRF
Claeys, Nature Communications, 2023

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.



Why should we be re-using data?

The weird and wonderful world of proteomics

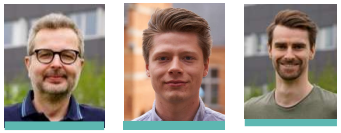
Four types of data re-use

Re-using available data to build machine learning models

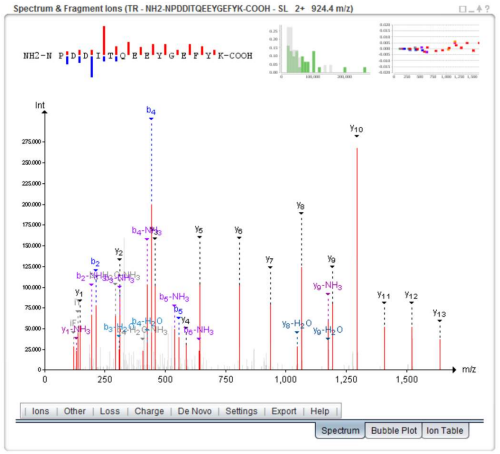
Reprocessing data with new models for new insights

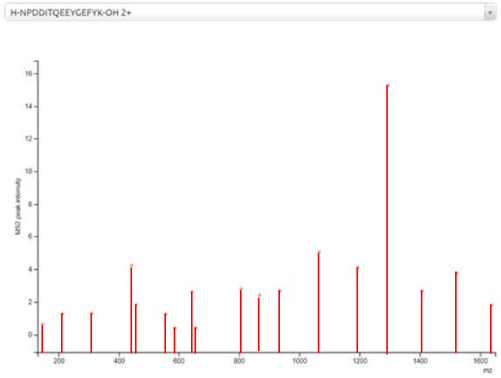
Repurposing large-scale data for new knowledge

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

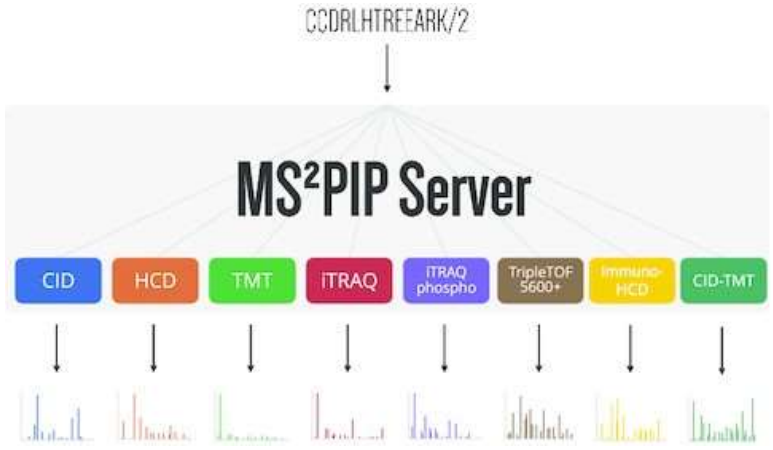


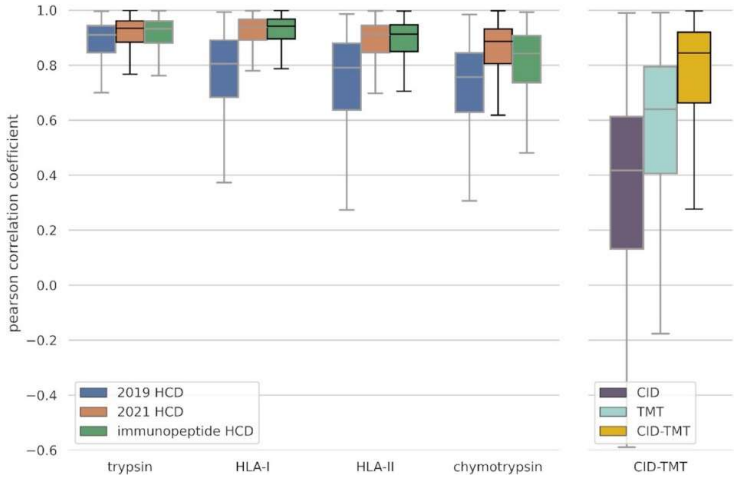
Vaudel, Nat. Biotech., 2015





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

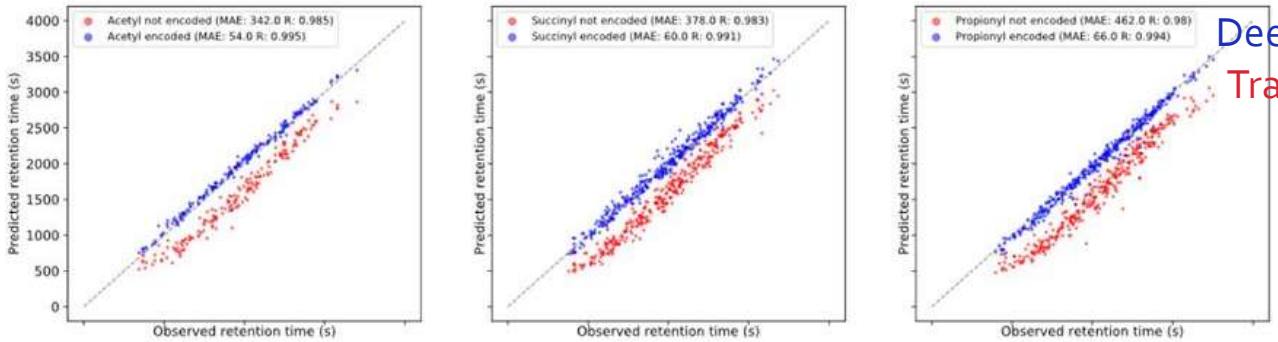




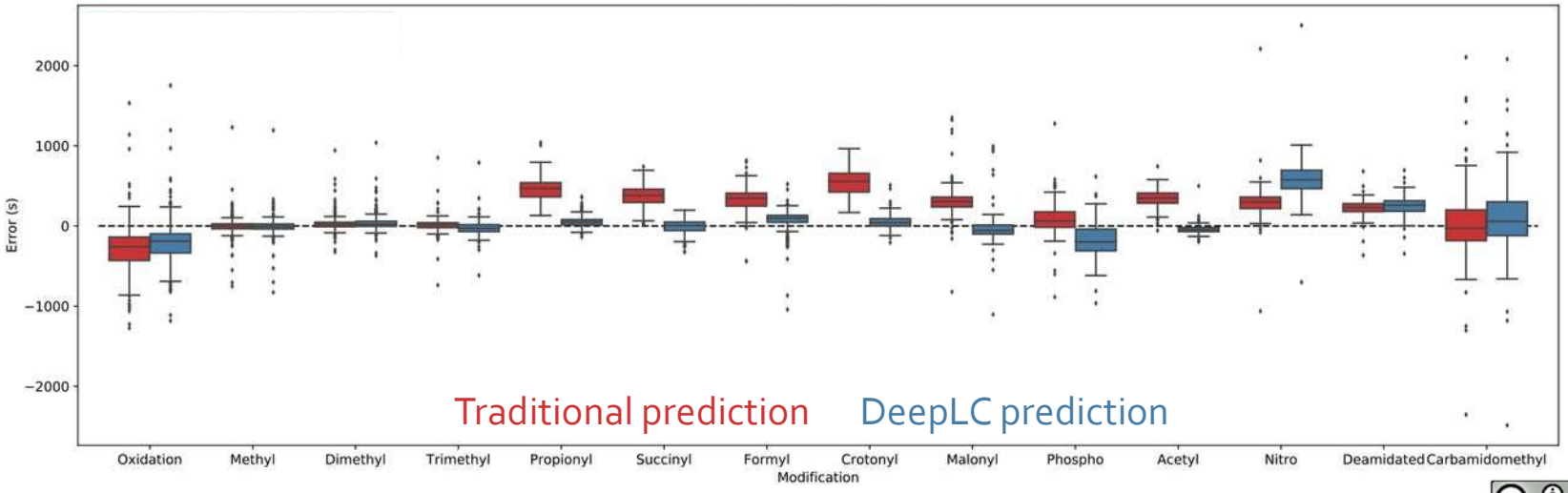
https://iomics.ugent.be/ms2pip
Declercq, Nucleic Acids Research, 2023

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

acetylation succinylation propionylation



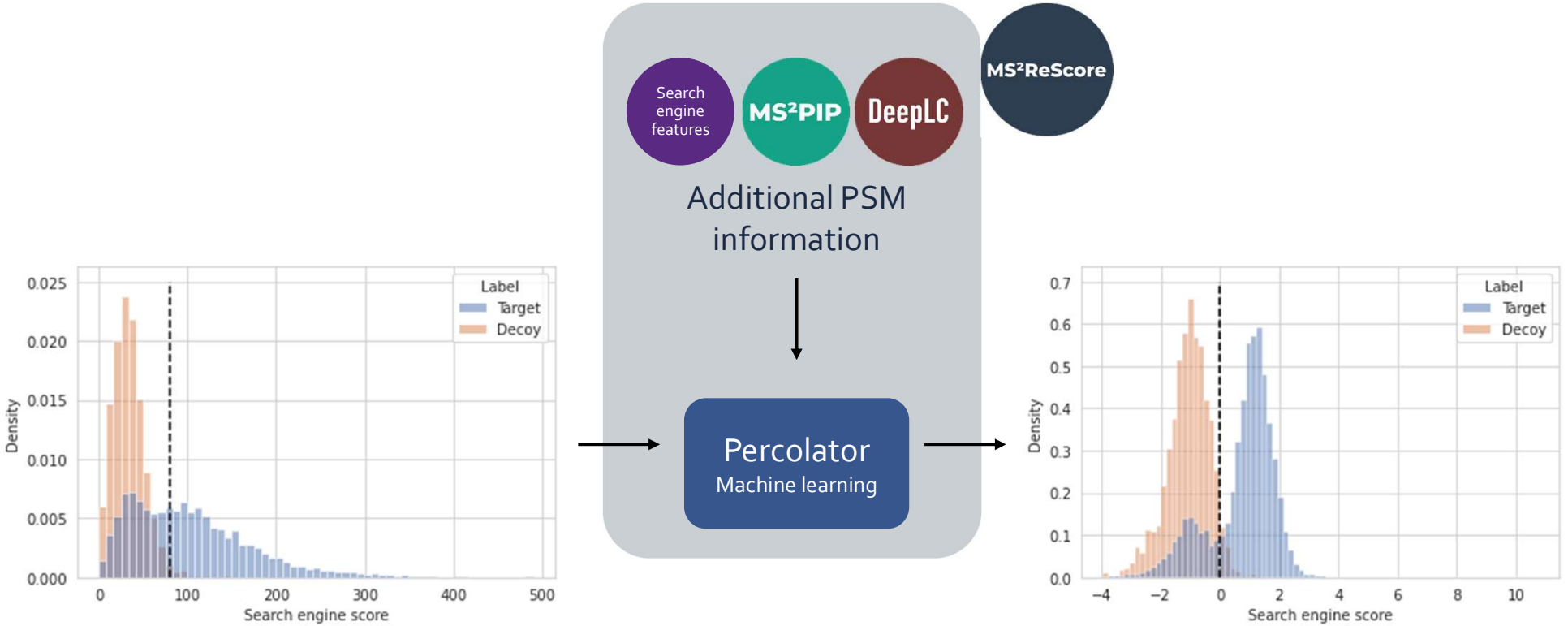
DeepLC prediction
Traditional prediction



Bouwmeester, Nature Methods, 2021

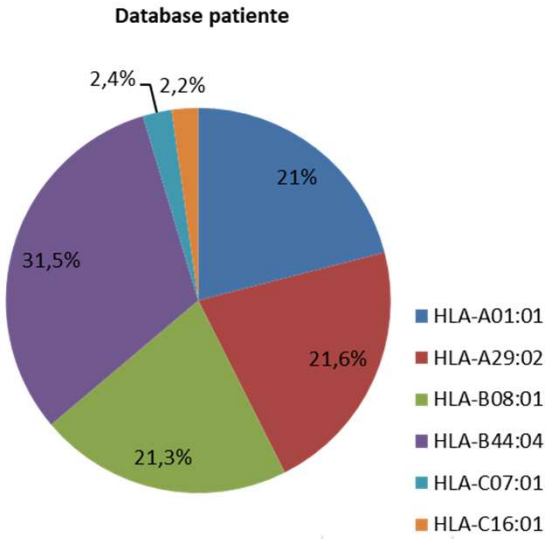
MS²Rescore makes use of these machine learning predictions to rescore identifications for improved sensitivity

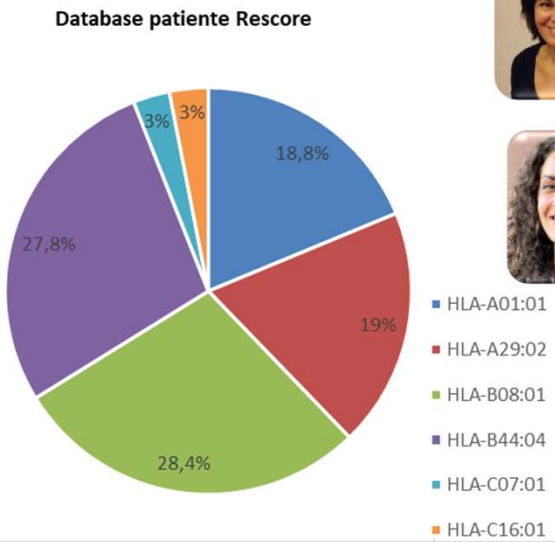
https://github.com/compomics/ms2rescore



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

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

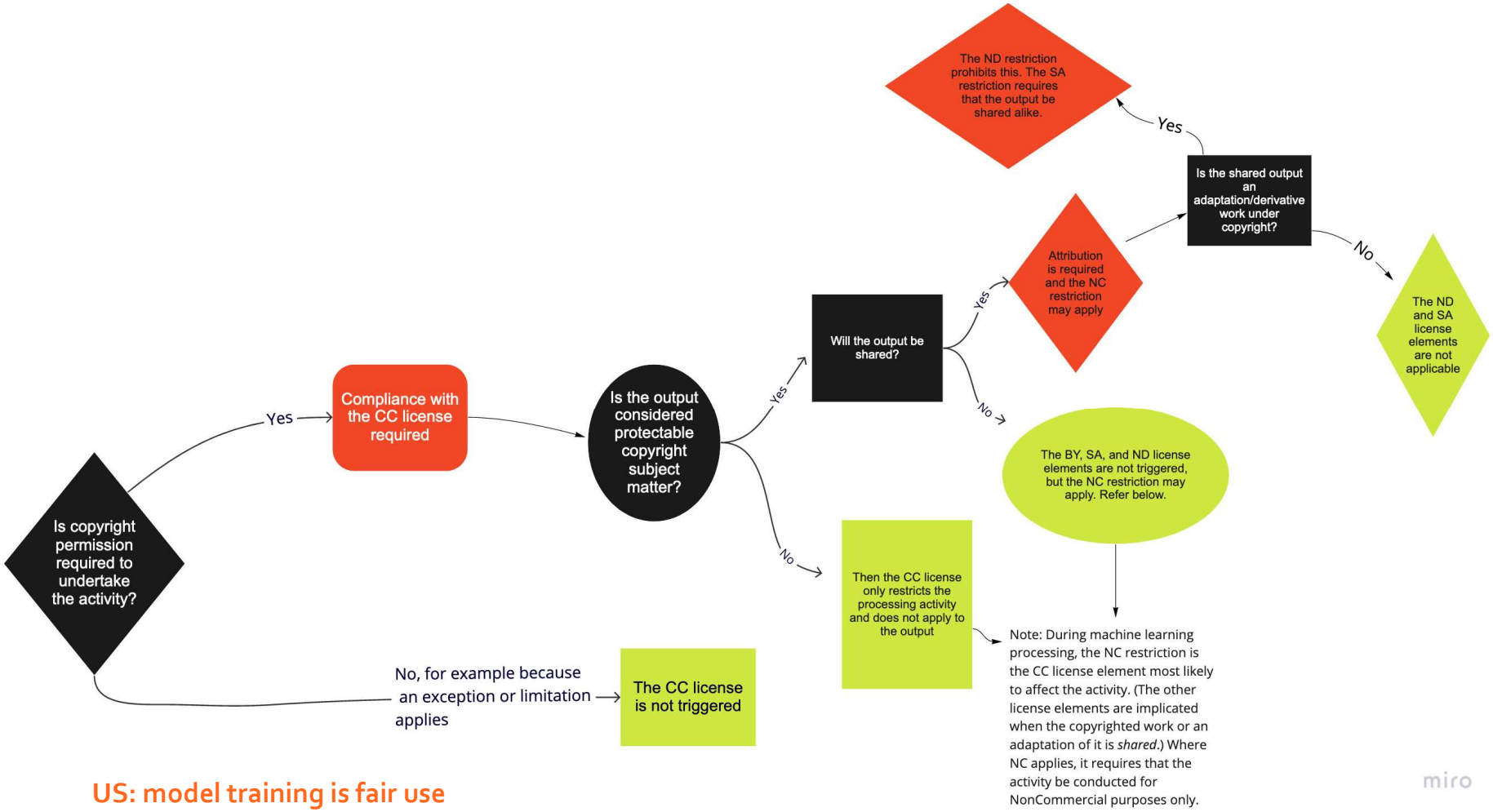
Analyse affinités

	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

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

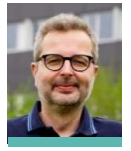
Four types of data re-use

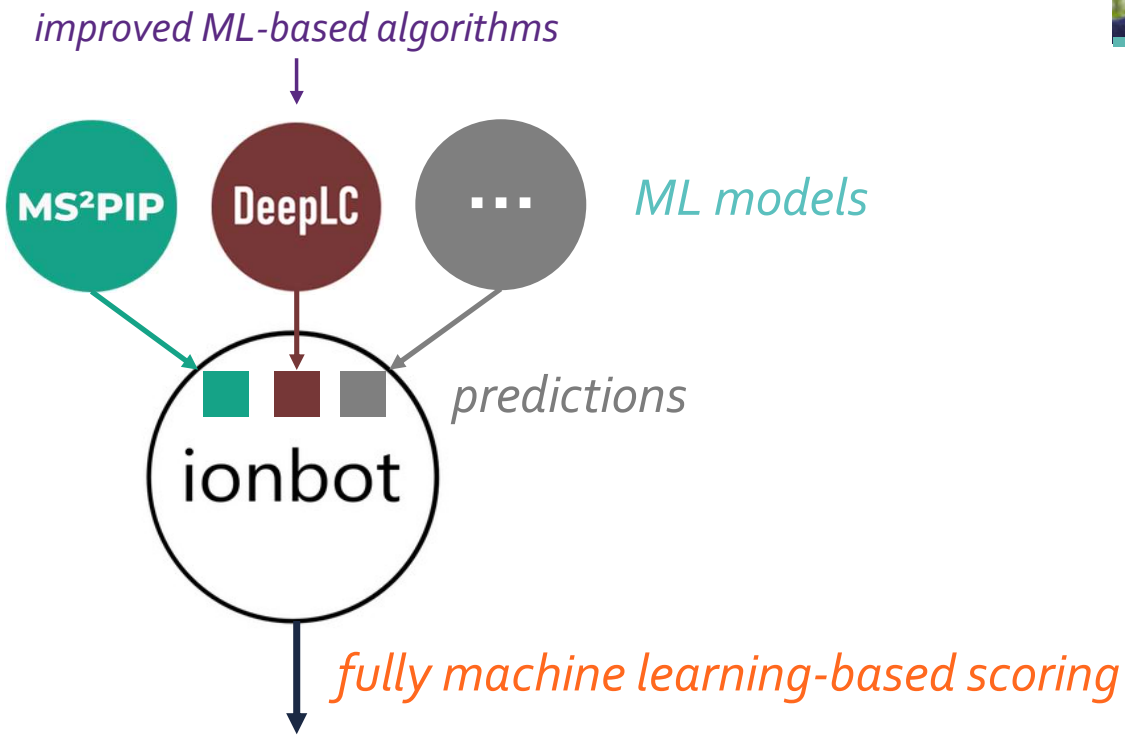
Re-using available data to build machine learning models

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

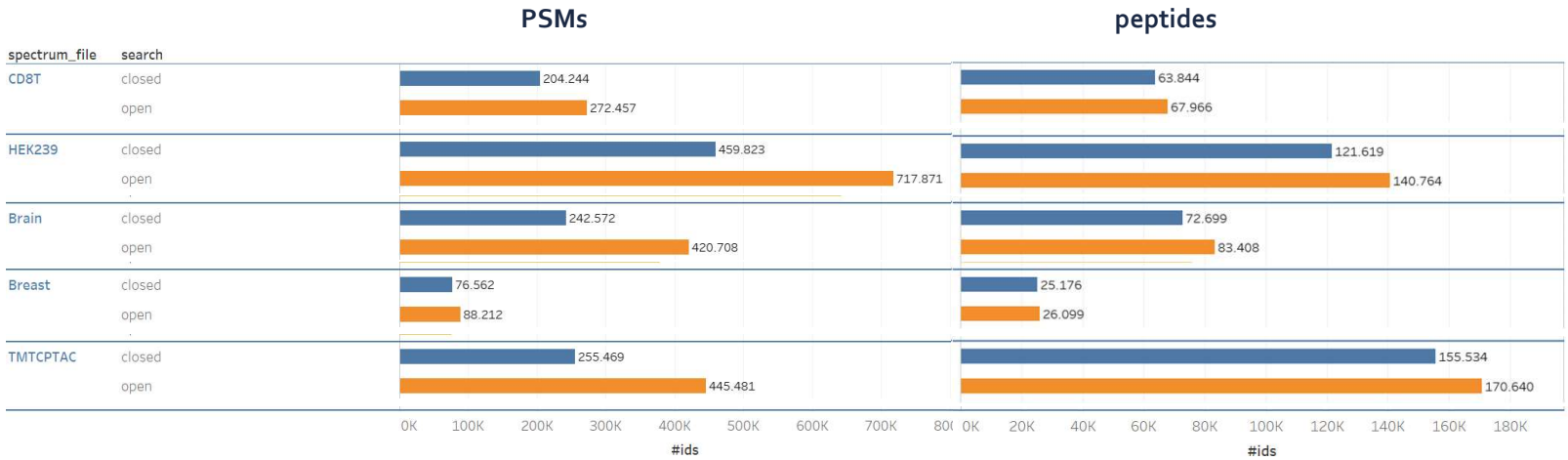


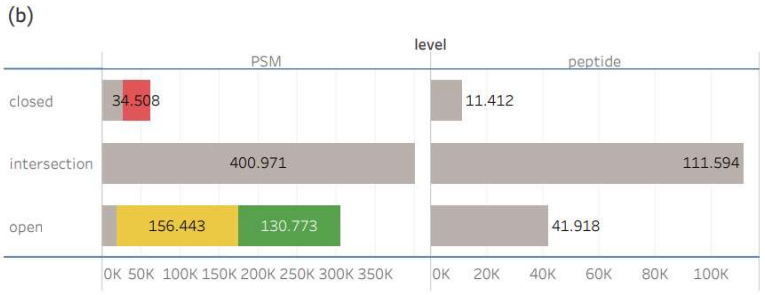


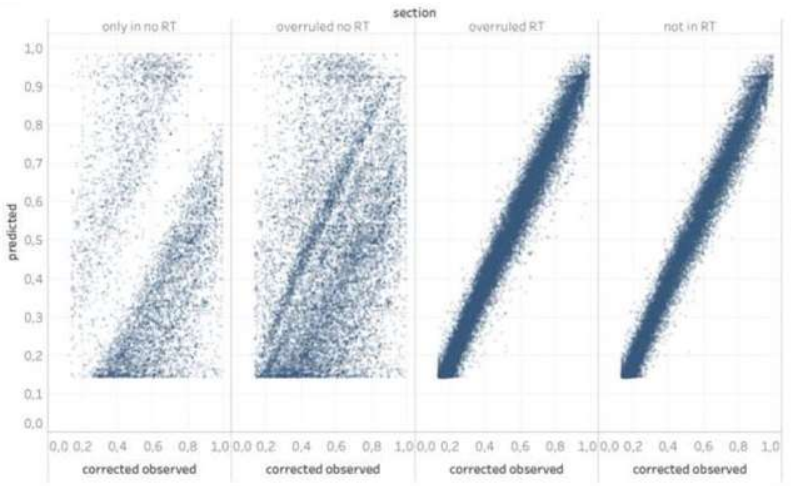
sensitive and highly accurate
identification of (modified) peptides

https://ionbot.cloud
Degroeve, https://www.biorxiv.org/content/10.1101/2021.07.02.450686v2

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

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

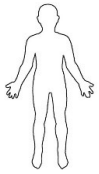


277 **mouse data sets**
appr. 600 million spectra
14.5 k raw files
from PRIDE

ionbot

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

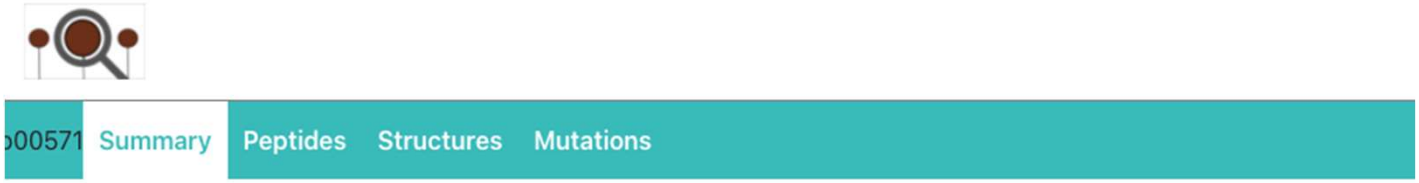
539 **human data sets**
appr. 924 million spectra
25 k raw files
from PRIDE



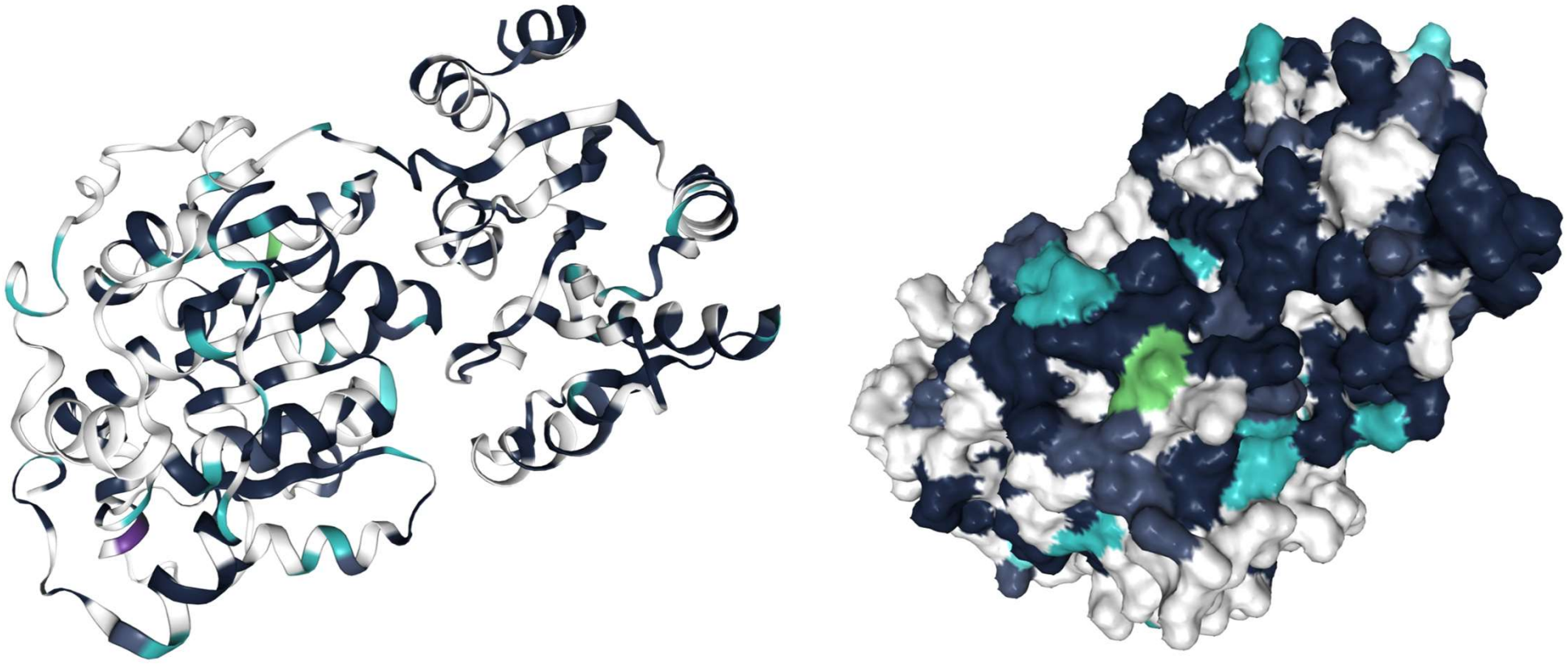
ionbot

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

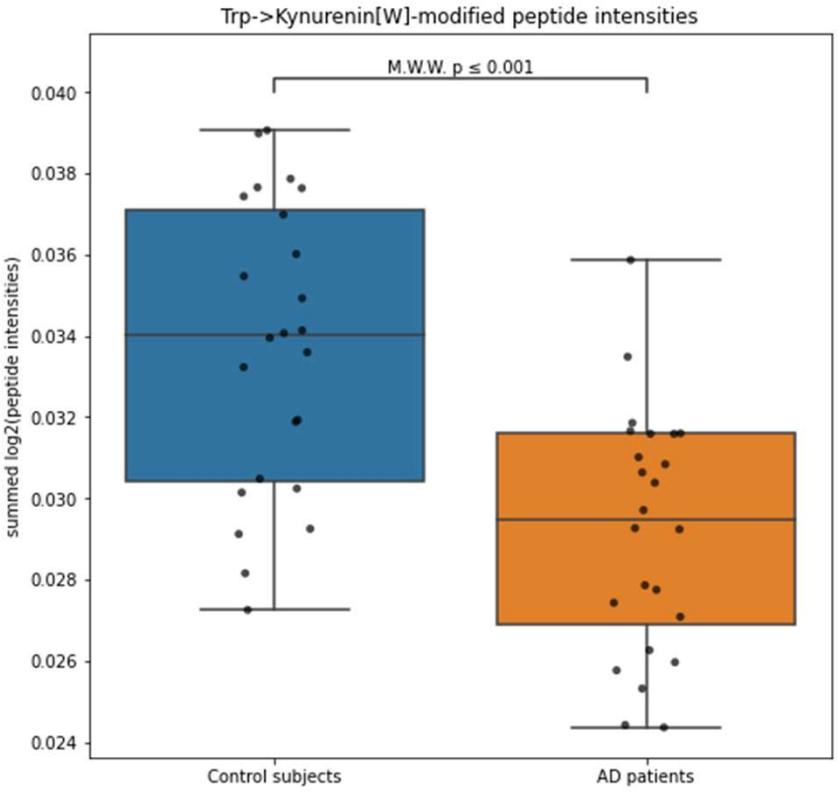


The 3D structure view drives home the point that proteins have tremendous modification potential

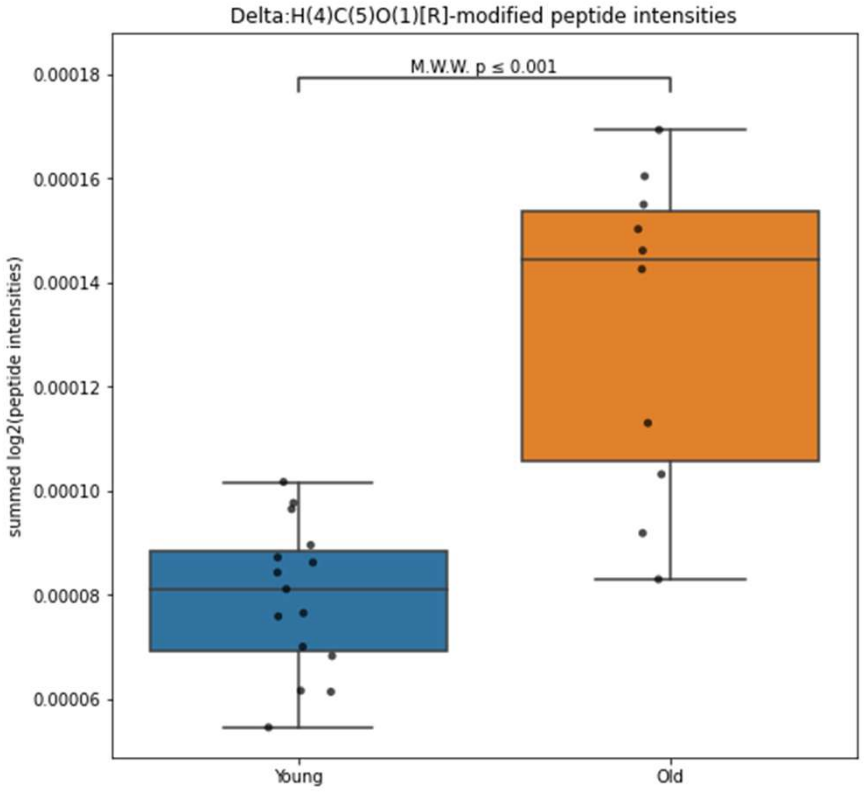


ATP-dependent RNA helicase DDX3X – O00571 – PDB 2I4I

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



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

De Vreese, *in preparation*

Why should we be re-using data?

The weird and wonderful world of proteomics

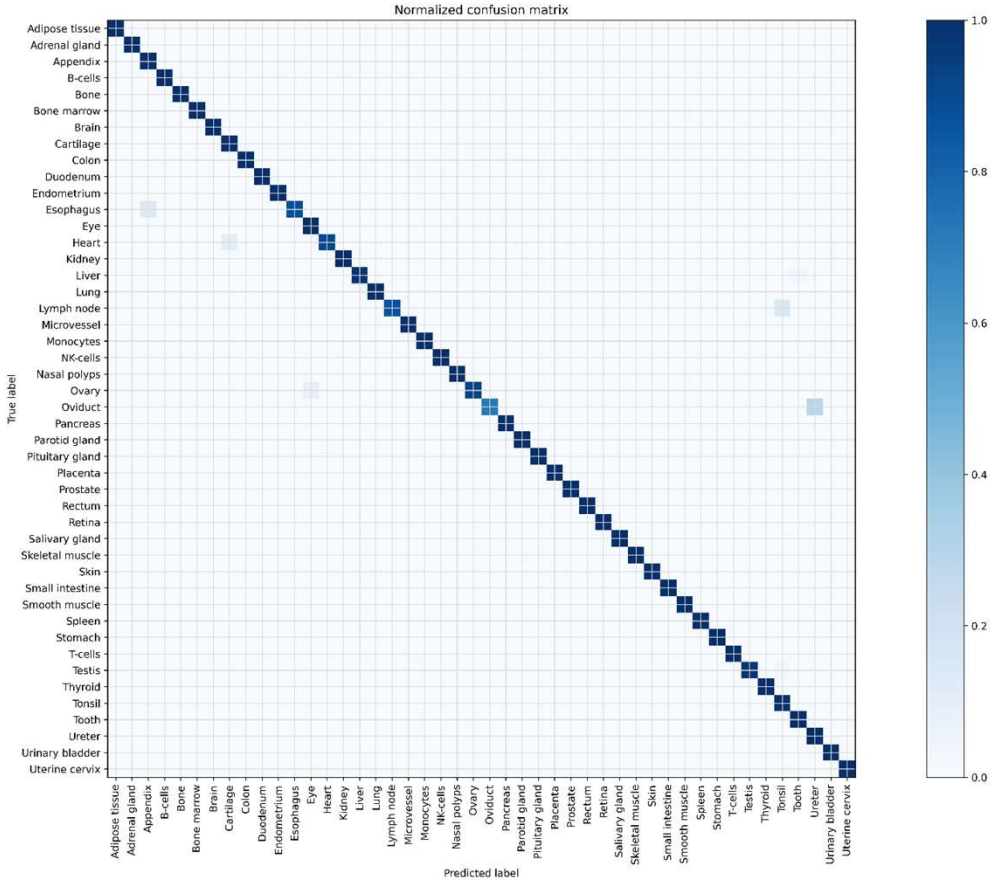
Four types of data re-use

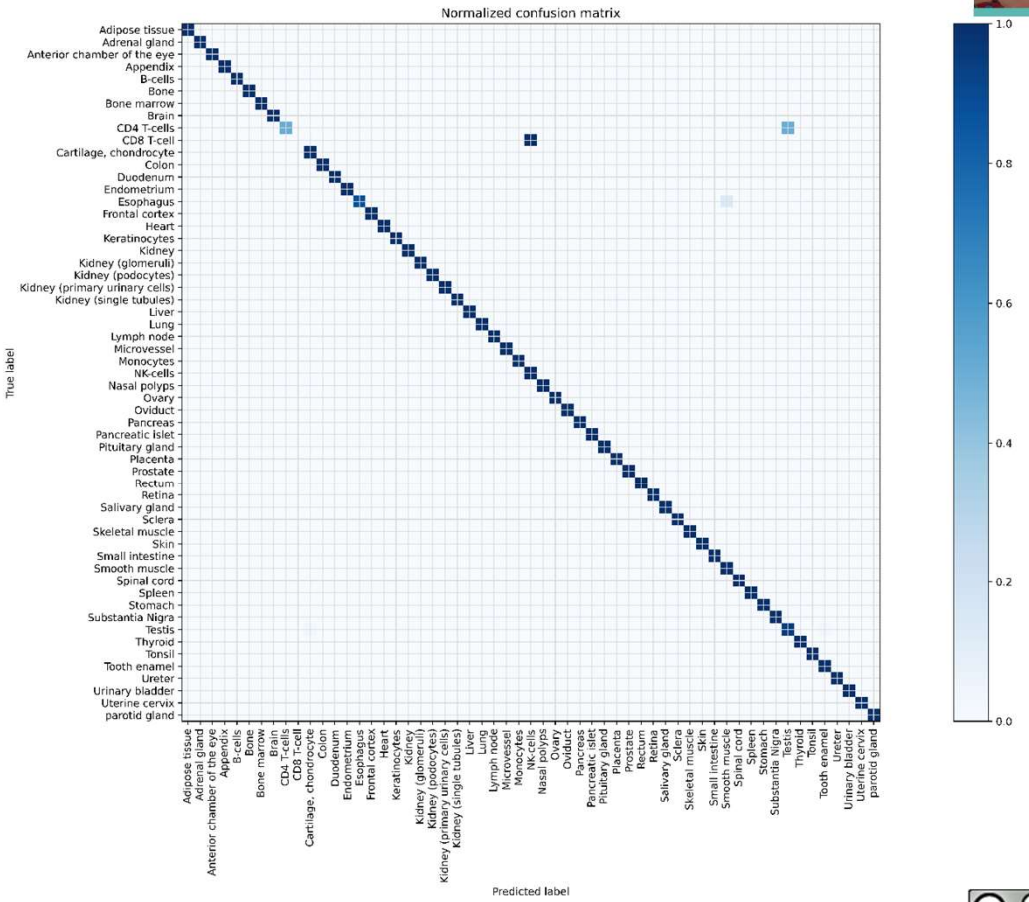
Re-using available data to build machine learning models

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





Claeys, J. Prot. Res., 2023

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

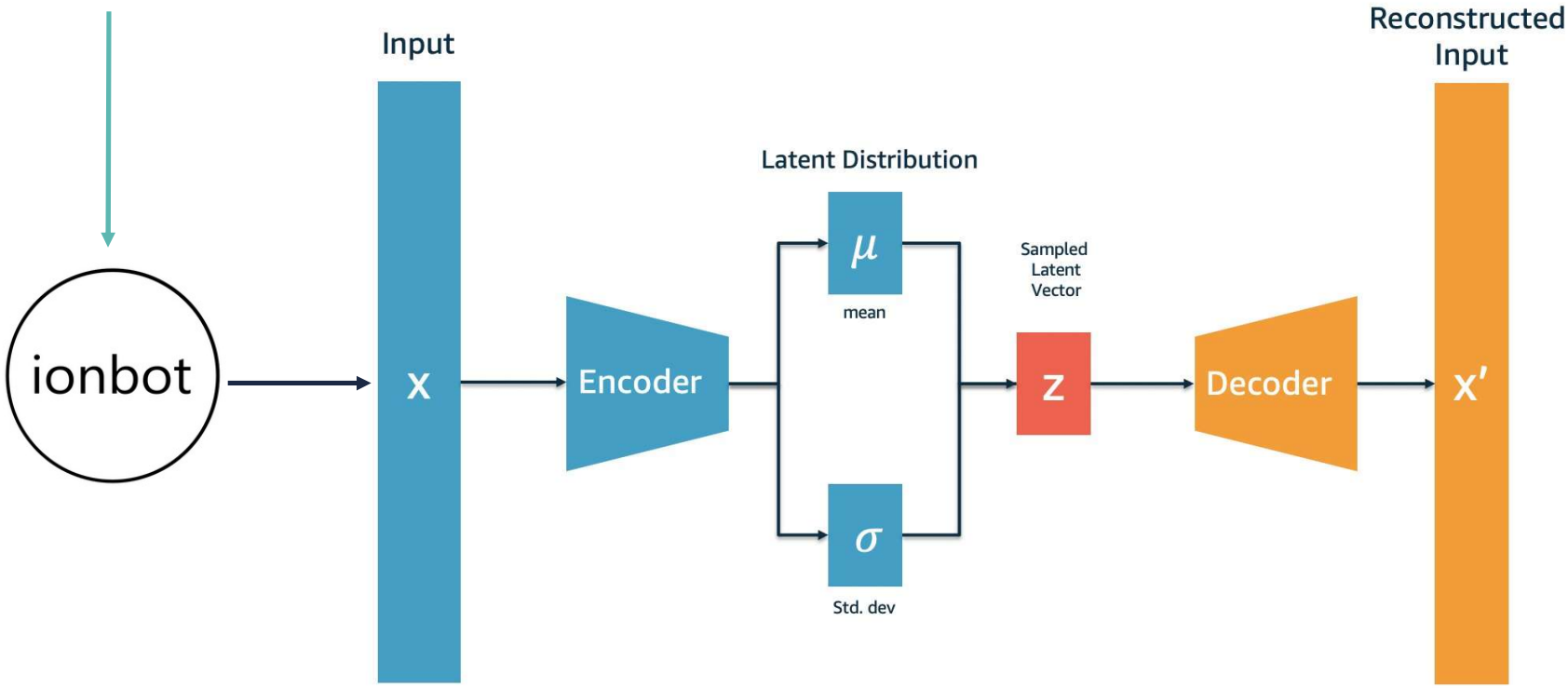
Lung was predicted with 53.6% probability

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

Claeys, J. Prot. Res., 2023

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

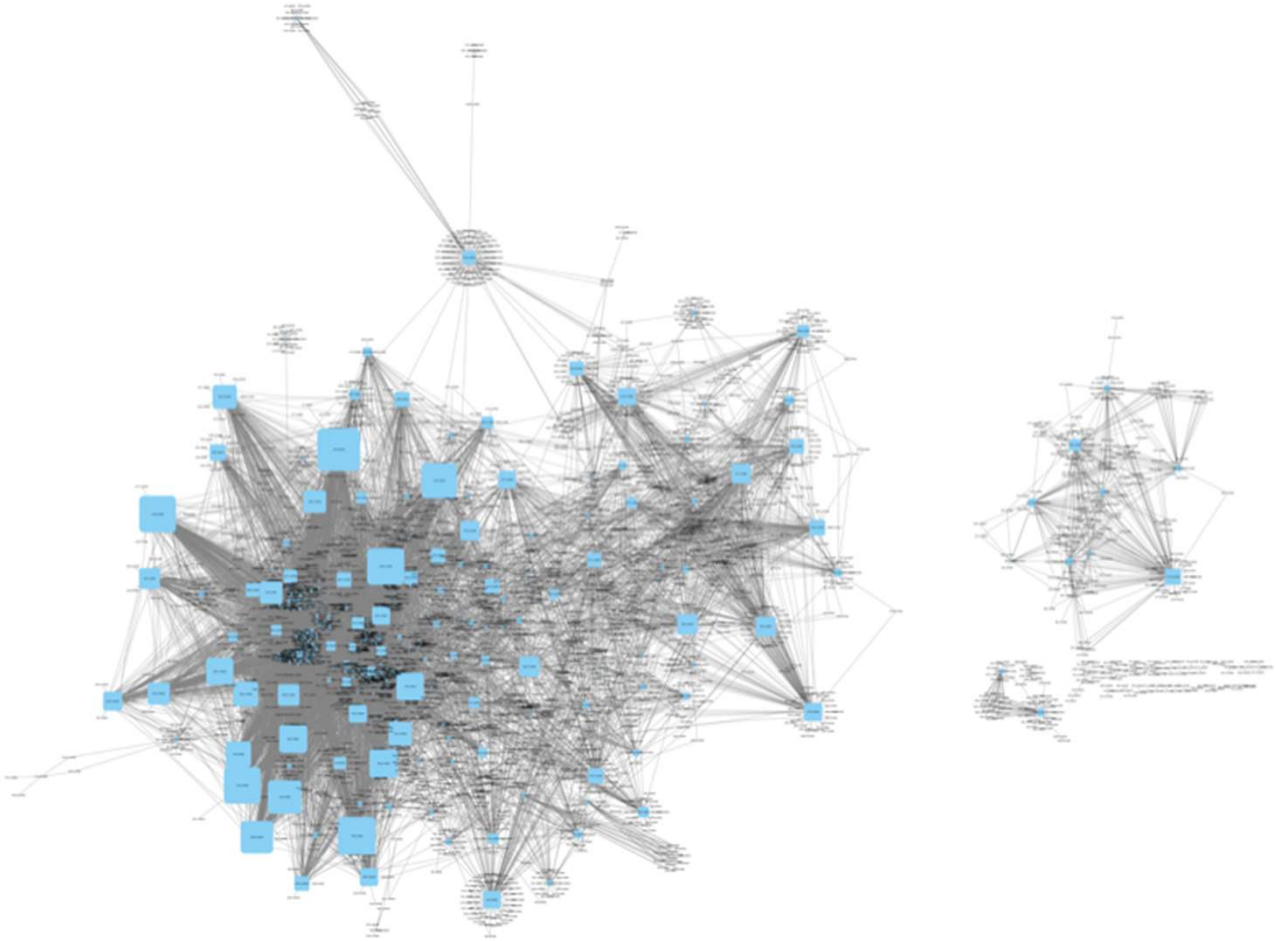
PRIDE Archive
PRoteomics IDEntifications Database



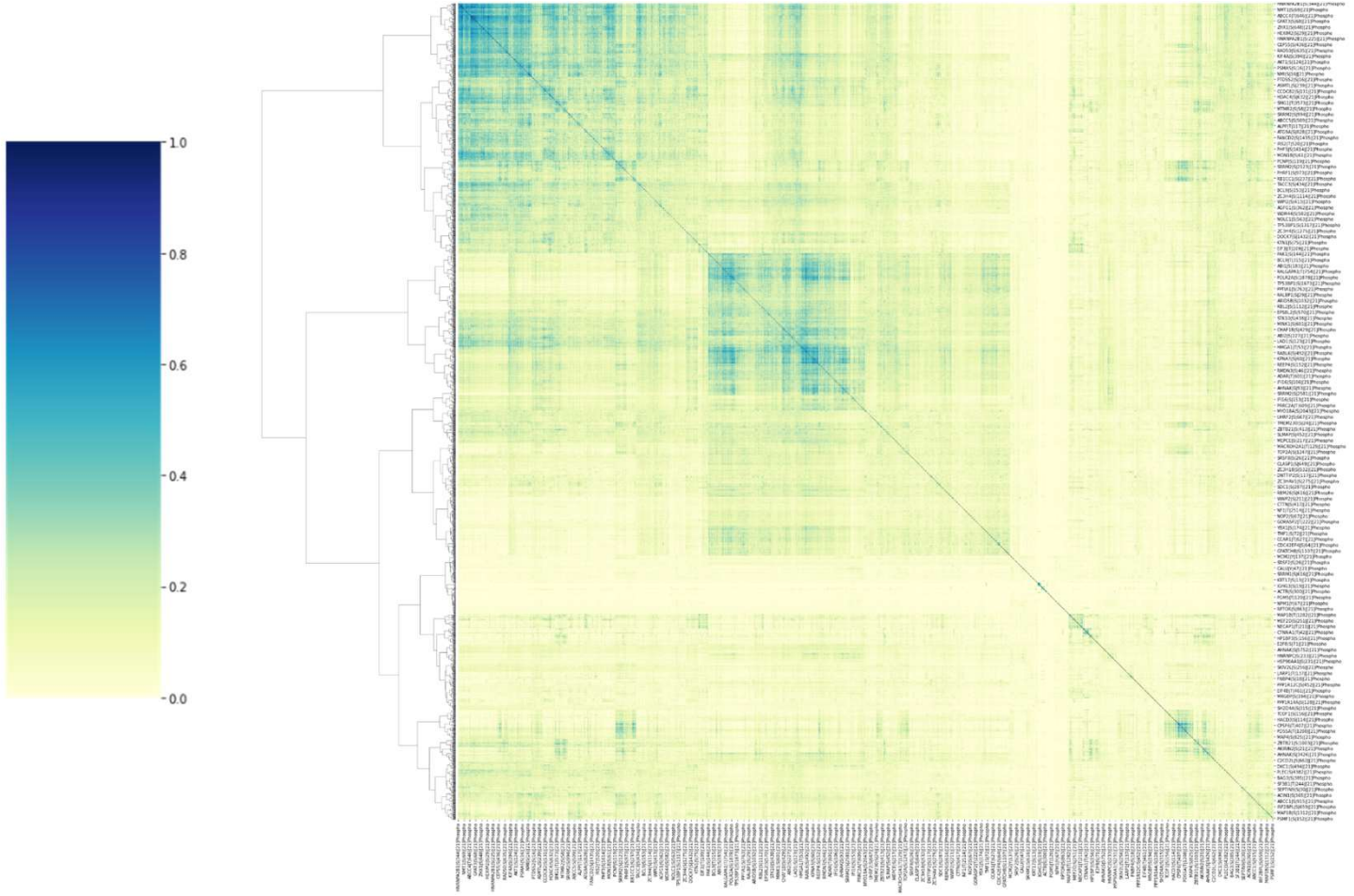
STRING

Koutrouli, Bioinformatics, 2024

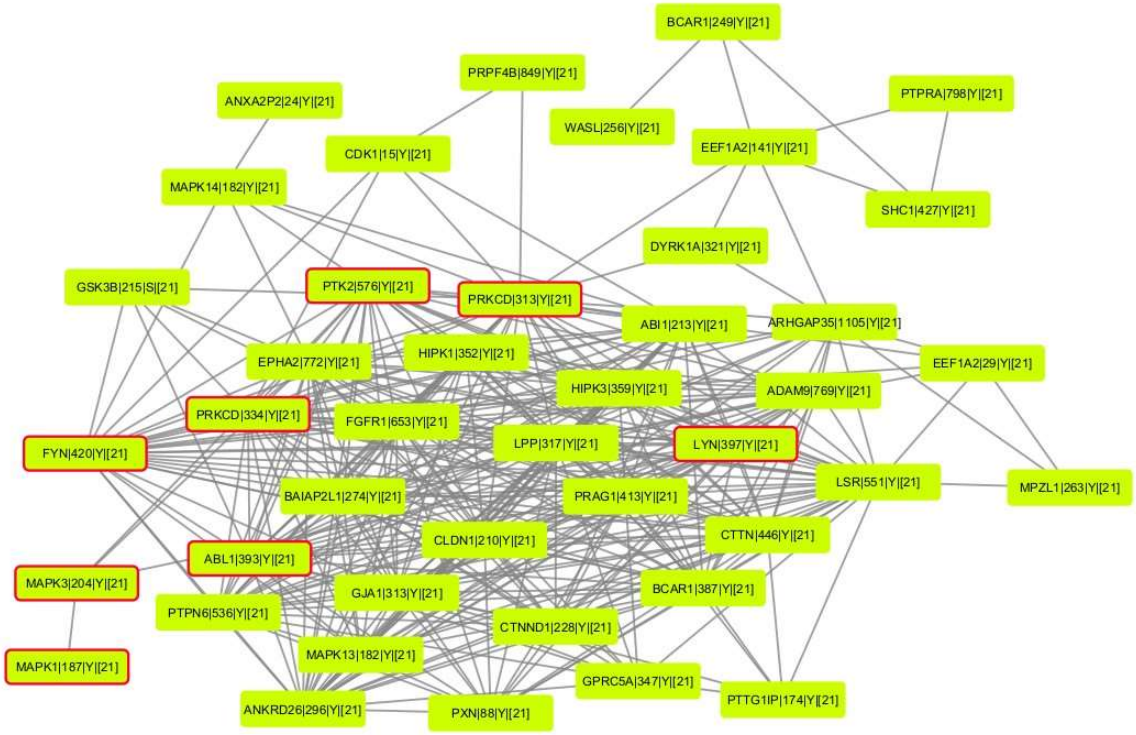
These associations have already been used to uncover the clientele of protein chaperones (prof. Joost Schymkowitz)

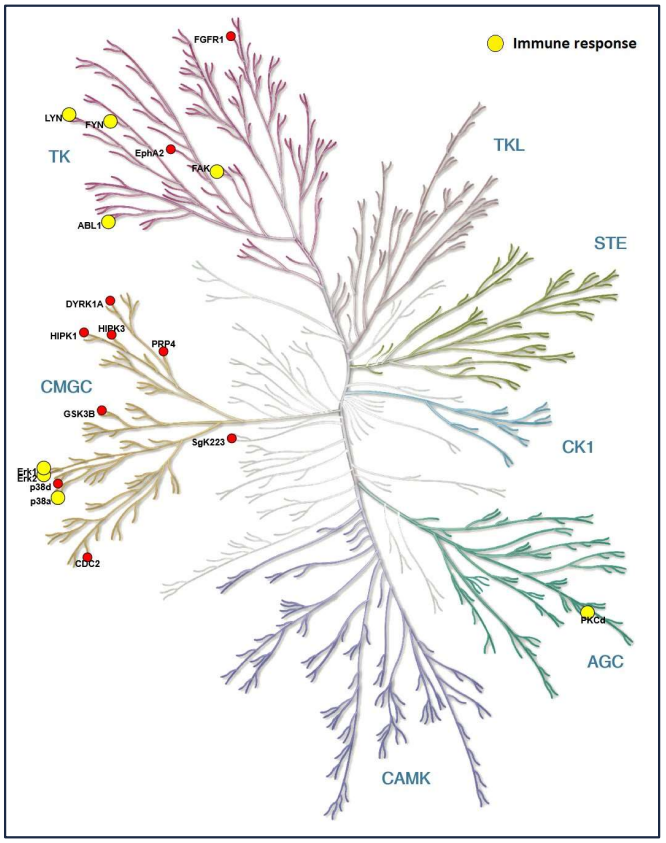


A VAE map of proteome-wide phosphosite co-occurrence reveals some interesting patterns, including kinase clusters

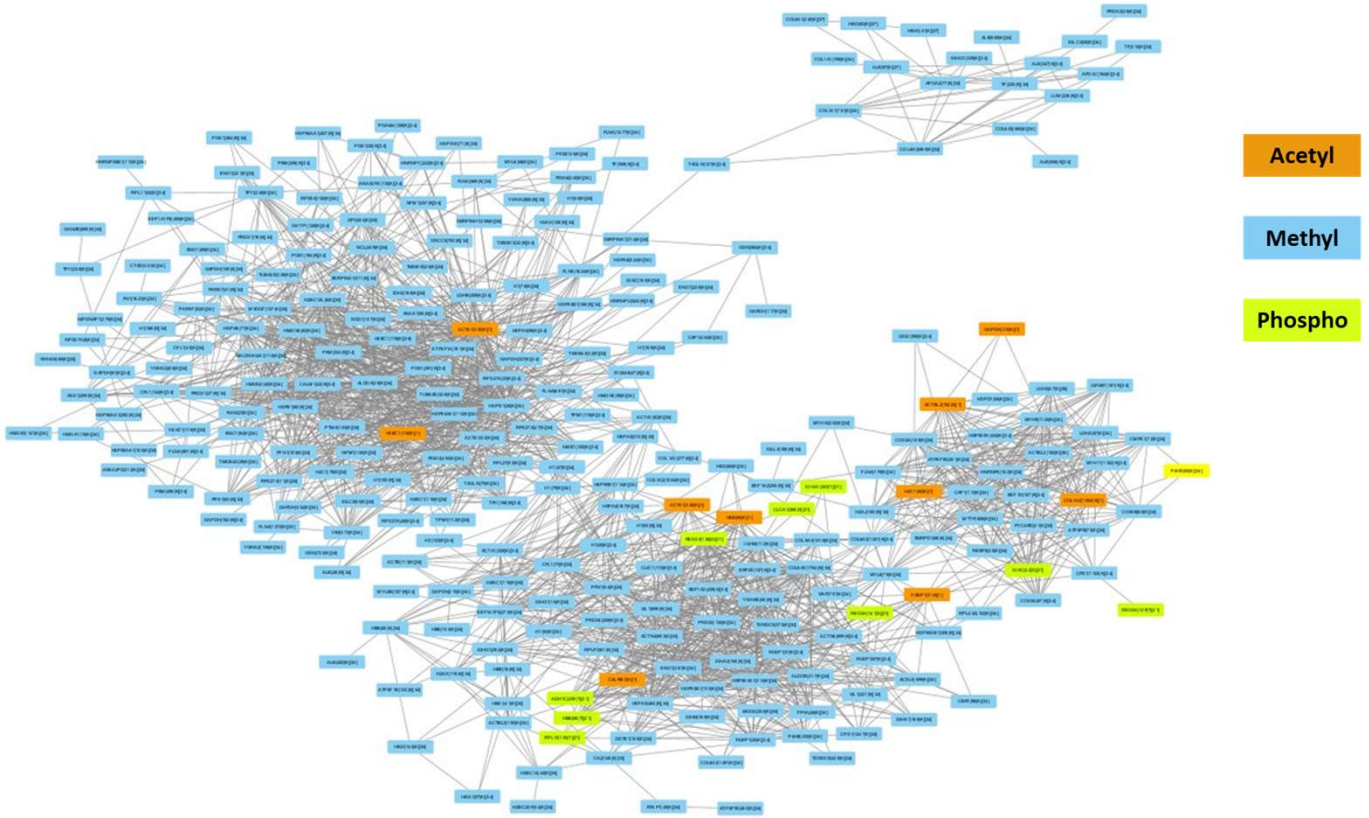


These phospho clusters contain connected components, here shown for tyrosine kinases and their key targets

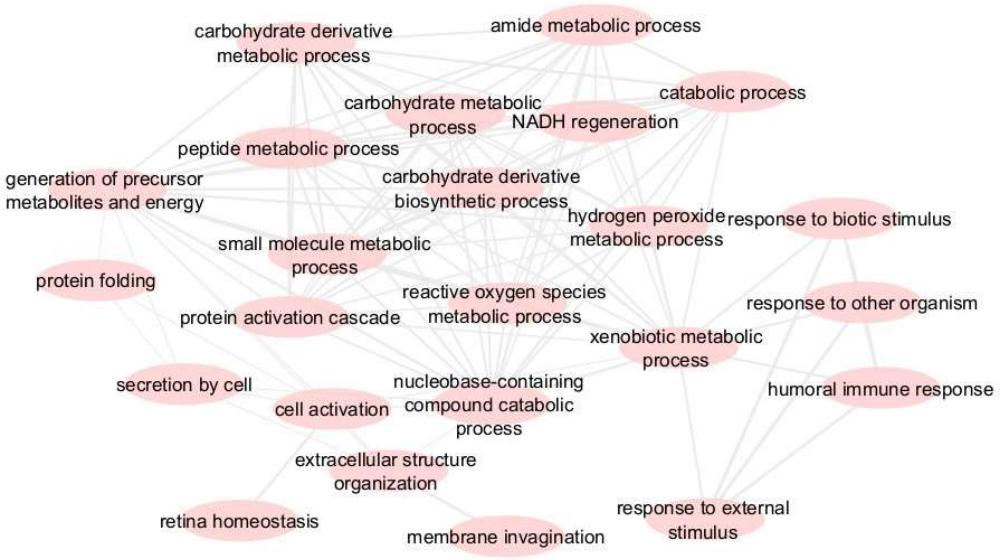


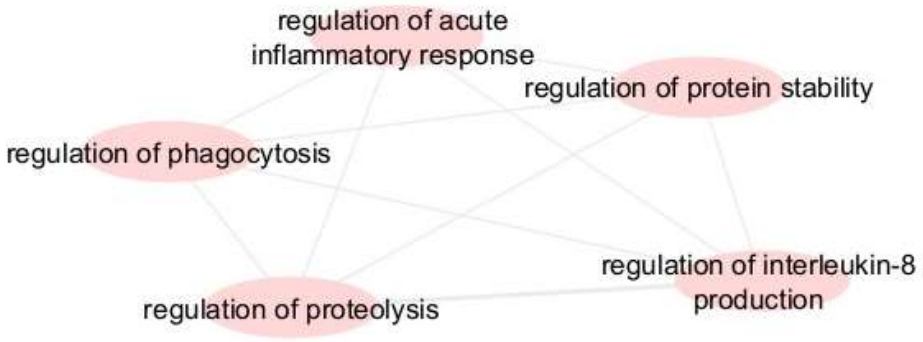


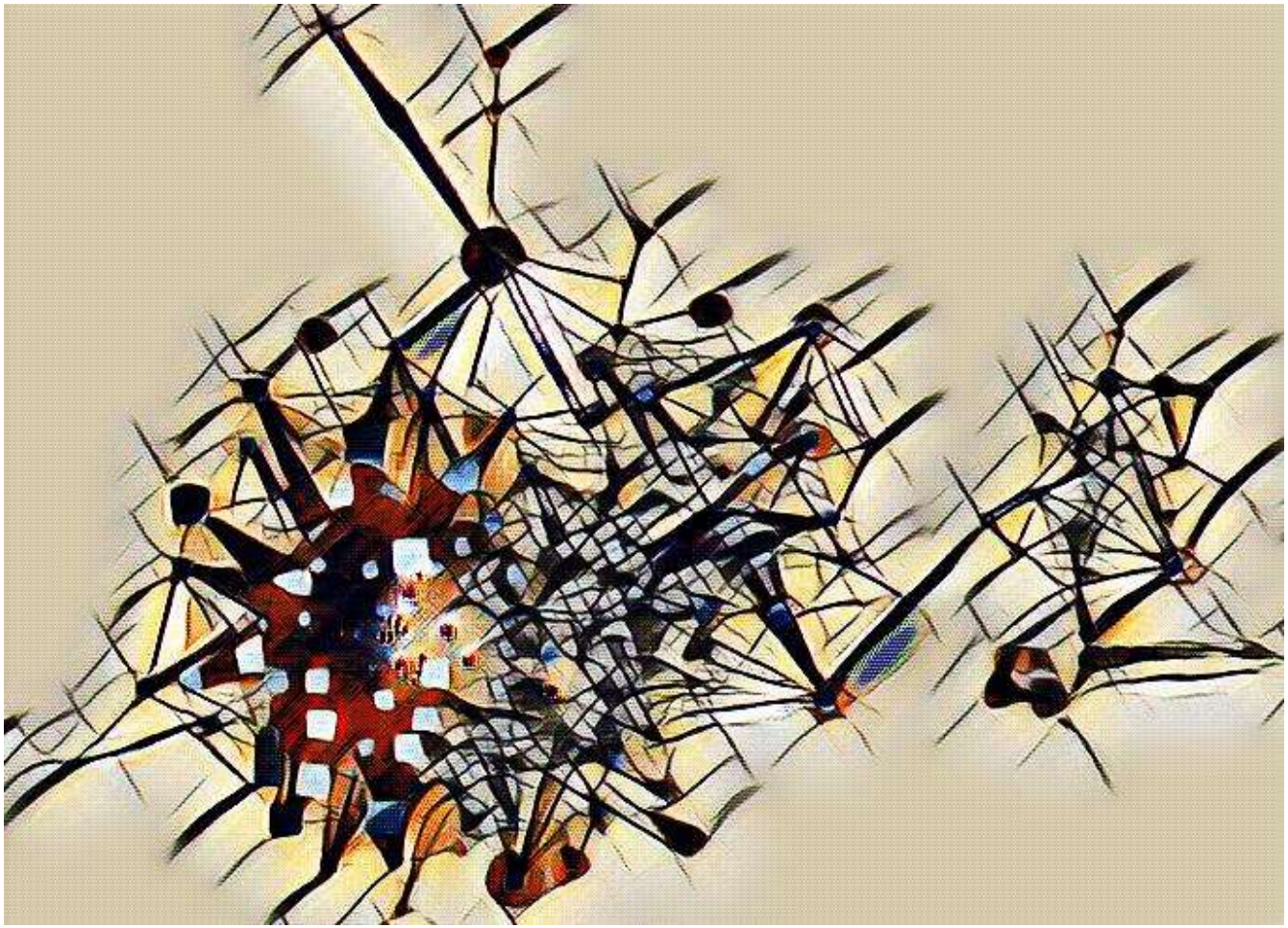
We can also analyse several modifications together, here methylation, acetylation, and phosphorylation



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







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









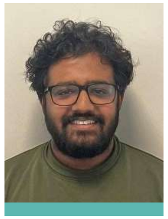










Comp
omics





VIB
science meets life

GHENT
UNIVERSITY






www.compomics.com

compomics.github.io