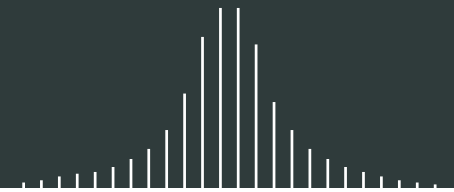
LSMBO

From wet lab to the Cloud

Alexandre BUREL



Université de Strasbourg

cnrs

IPHC Institut Pluridisciplinaire Hubert CURIEN STRASBOURG

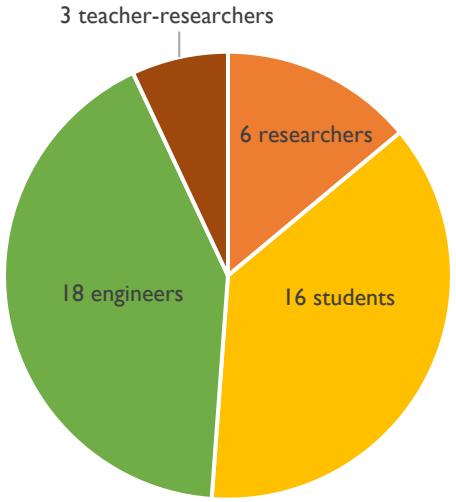
ProFI PROTEOMICS

Presentation of the LSMBO

- Laboratoire de Spectrométrie de Masse BioOrganique
  - Part of Institut Pluridisciplinaire Hubert Curien (UMR7178)
  - Under joint supervision of CNRS and University of Strasbourg
  - 12 Mass spectrometers

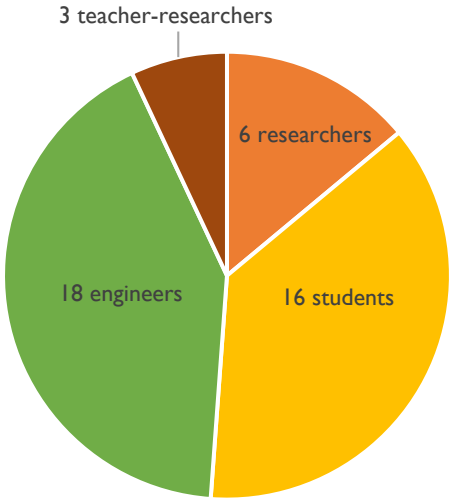
Presentation of the LSMBO

- Laboratoire de Spectrométrie de Masse BioOrganique
  - Part of Institut Pluridisciplinaire Hubert Curien (UMR7178)
  - Under joint supervision of CNRS and University of Strasbourg
  - 12 Mass spectrometers
- Who we are
  - Around 40 people
  - Half are permanent staff
  - A majority of chemists
  - But also biologists, pharmacists, bioinformaticians, etc.



Presentation of the LSMBO

- Laboratoire de Spectrométrie de Masse BioOrganique
  - Part of Institut Pluridisciplinaire Hubert Curien (UMR7178)
  - Under joint supervision of CNRS and University of Strasbourg
  - 12 Mass spectrometers
- Who we are
  - Around 40 people
  - Half are permanent staff
  - A majority of chemists
  - But also biologists, pharmacists, bioinformaticians, etc.
- What we do
  - Specialized in proteomics
  - Identification and quantification of proteins
  - Search for biomarkers
  - Characterization of therapeutic proteins

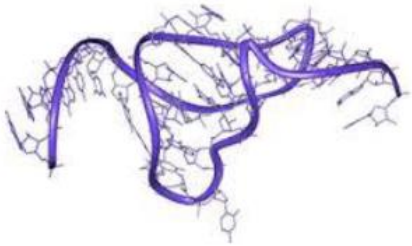


A few definitions

- Proteomics
  - Study of all the proteins in a cell, tissue or organism
  - Proteins are essential molecules for most functions in all living organisms
  - Genes are the source code for proteins

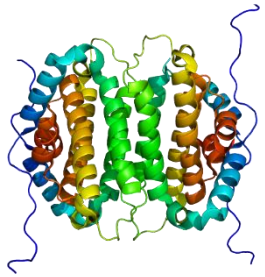
DNA

Transcription



RNA

Translation



Proteins

For humans:	~20 000 genes	~100 000 transcripts	~5 000 000 proteins

A few definitions

- Mass spectrometry
  - Technique used to measure the mass of molecules
  - Proteins are cut into small pieces called peptides
  - Each peptide is charged
  - Then accelerated through an flight tube
  - The flight time to reach the detector is measured

Bruker timsTOF Ultra

A few definitions

- Mass spectrometry
  - Technique used to measure the mass of molecules
  - Proteins are cut into small pieces called peptides
  - Each peptide is charged
  - Then accelerated through an flight tube
  - The flight time to reach the detector is measured
  - The mass of the peptide can now be deduced:

$$E = \frac{1}{2}mv^2$$

Bruker timsTOF Ultra

A few definitions

- Mass spectrometry
  - Technique used to measure the mass of molecules
  - Proteins are cut into small pieces called peptides
  - Each peptide is charged
  - Then accelerated through an flight tube
  - The flight time to reach the detector is measured
  - The mass of the peptide can now be deduced:

$$E = \frac{1}{2}mv^2$$

The mass we are looking for

Bruker timsTOF Ultra

A few definitions

- Mass spectrometry
  - Technique used to measure the mass of molecules
  - Proteins are cut into small pieces called peptides
  - Each peptide is charged
  - Then accelerated through an flight tube
  - The flight time to reach the detector is measured
  - The mass of the peptide can now be deduced:

$$E = \frac{1}{2}mv^2$$

Energy applied for the acceleration

The mass we are looking for

Bruker timsTOF Ultra

A few definitions

- Mass spectrometry
  - Technique used to measure the mass of molecules
  - Proteins are cut into small pieces called peptides
  - Each peptide is charged
  - Then accelerated through an flight tube
  - The flight time to reach the detector is measured
  - The mass of the peptide can now be deduced:

$$E = \frac{1}{2}mv^2$$

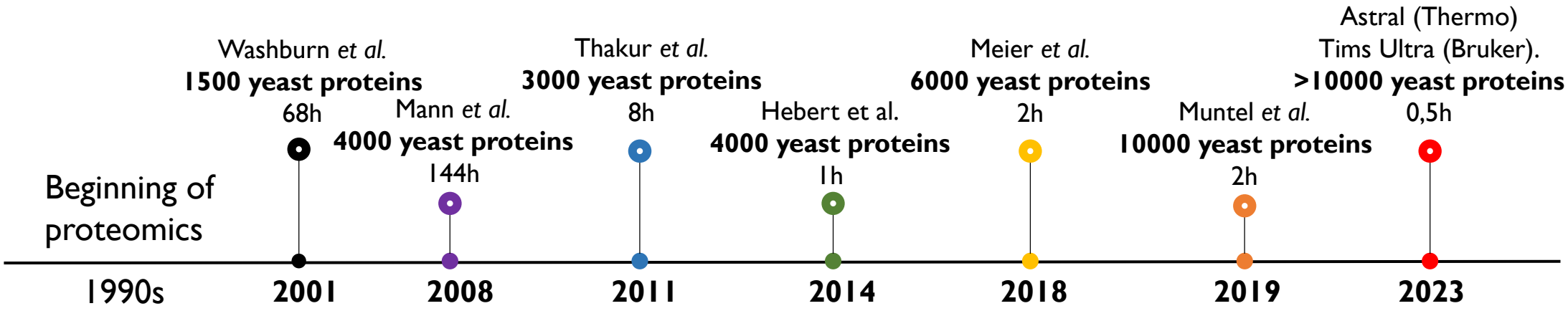
Energy applied for the acceleration

The mass we are looking for

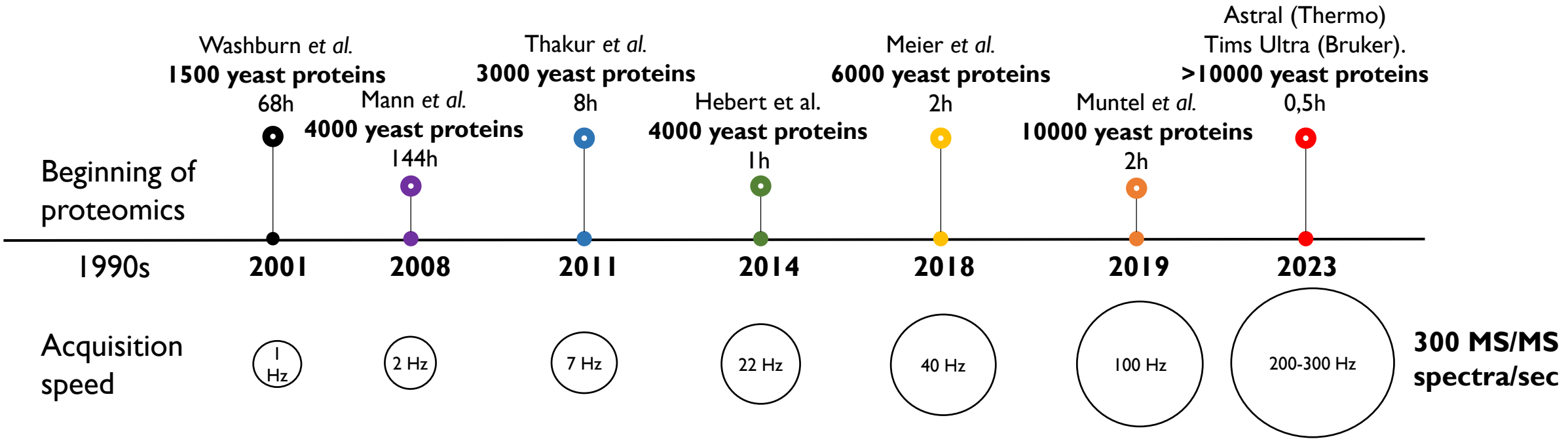
Velocity based on the time required to cover the flight tube

Bruker timsTOF Ultra

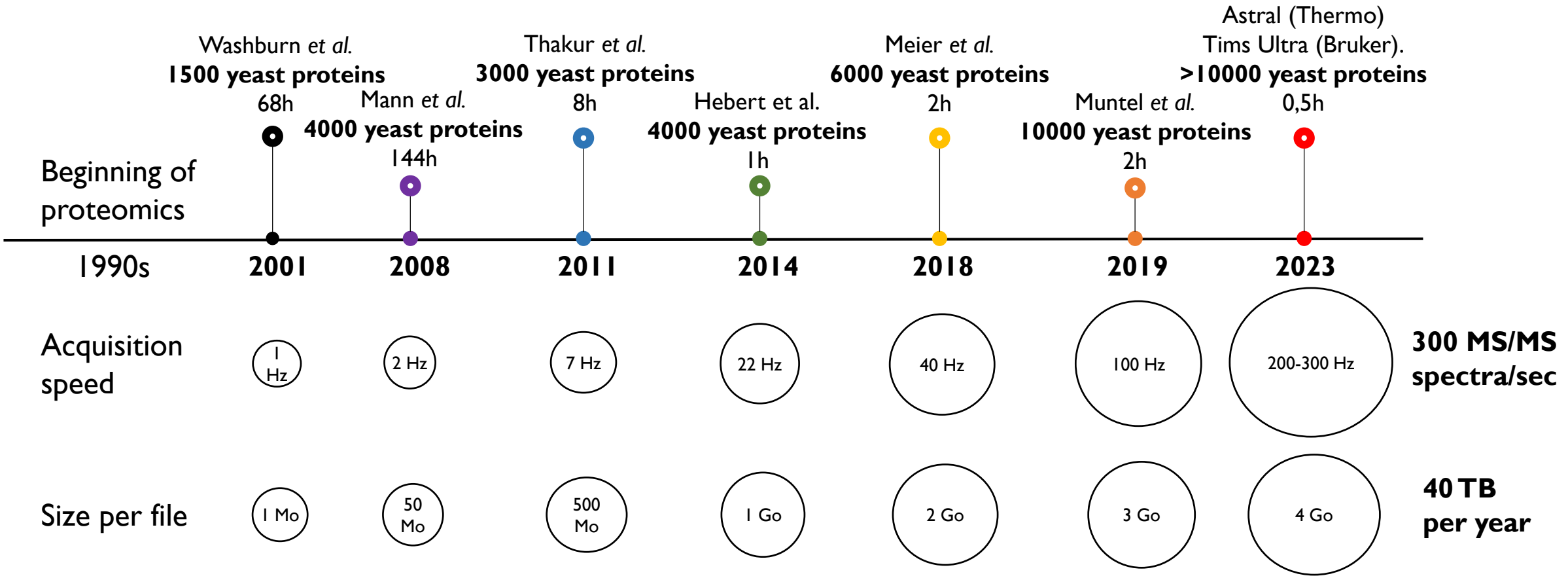
Improvements in mass spectrometry



Improvements in mass spectrometry



Improvements in mass spectrometry



Data processing

- Many software involved for different purposes
  - Identification & quantification (Proline, Maxquant, Dia-NN, etc)
  - Functional analysis (Kegg, GeneOntology, Reactome, etc.)
  - Statistical analysis (Prostar, MSqRob, R)
  - FAIR data practices (PRIDE, ProteomeXchange)

Data processing

- Many software involved for different purposes
  - Identification & quantification (Proline, Maxquant, Dia-NN, etc)
  - Functional analysis (Kegg, GeneOntology, Reactome, etc.)
  - Statistical analysis (Prostar, MSqRob, R)
  - FAIR data practices (PRIDE, ProteomeXchange)

- Data treatment can require huge resources and long computation time
  - Quantification compares all analyses at once
  - Can consume all the resources of one server for days
  - One process can block the server for everyone else

Data processing

- Many software involved for different purposes
  - Identification & quantification (Proline, Maxquant, Dia-NN, etc)
  - Functional analysis (Kegg, GeneOntology, Reactome, etc.)
  - Statistical analysis (Prostar, MSqRob, R)
  - FAIR data practices (PRIDE, ProteomeXchange)

- Data treatment can require huge resources and long computation time
  - Quantification compares all analyses at once
  - Can consume all the resources of one server for days
  - One process can block the server for everyone else
  - **Impossible to content 40 users with only local resources!!**

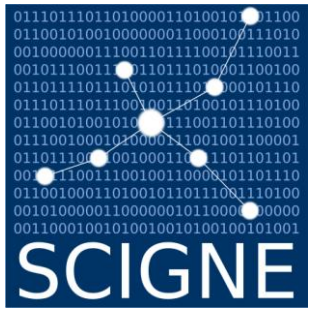
Data processing

- Many software involved for different purposes
  - Identification & quantification (Proline, Maxquant, Dia-NN, etc)
  - Functional analysis (Kegg, GeneOntology, Reactome, etc.)
  - Statistical analysis (Prostar, MSqRob, R)
  - FAIR data practices (PRIDE, ProteomeXchange)

- Data treatment can require huge resources and long computation time
  - Quantification compares all analyses at once
  - Can consume all the resources of one server for days
  - One process can block the server for everyone else
  - **Impossible to content 40 users with only local resources!!**

- The solution is in the Cloud!

The SCIGNE Platform



In a few words

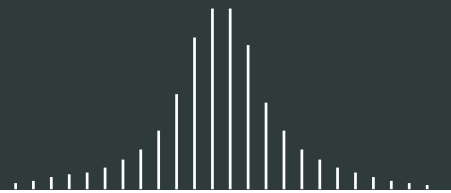
- SCIGNE – a platform offering compute and storage services hosted by IPHC
- Includes support to help researchers to manage and analyse large amounts of data in several scientific fields (physics, chemistry, biology and ecology)
- Labelised by IN2P3 and by the University of Strasbourg (CORTECS)
- Involved in several national and international scientific projects
- Managed by a team of 8 highly-skilled engineers

Expertise & Services

- Processing and analysis of large amounts of scientific data
- Computation reproducibility studies
- Data and software management plans, making the data FAIR
- Software development (including GPU & IA), source code opening
- IT Security & Green computing
- High-throughput computing, Cloud computing and scientific data management services

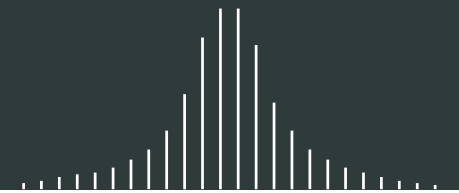
https://scigne.fr

The SCIGNE Platform



- Advantages
  - Plenty of resources available
  - Stable and powerful environment
  - The SCIGNE engineers are responsive and competent

The SCIGNE Platform



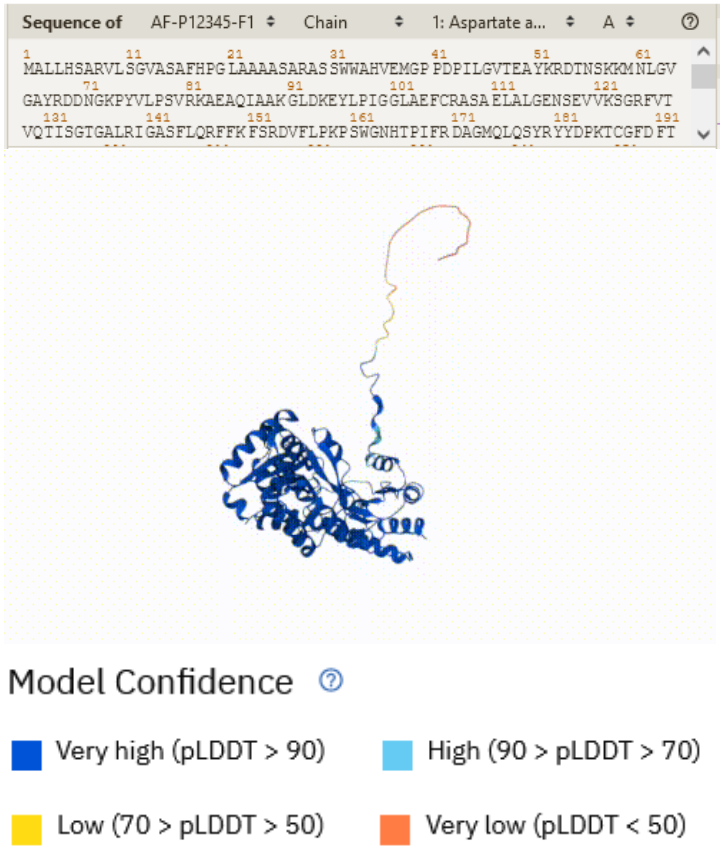
- Advantages
  - Plenty of resources available
  - Stable and powerful environment
  - The SCIGNE engineers are responsive and competent
- Drawbacks
  - Data transfer can take a while
  - Best fit for Linux command line software
  - Best fit for non-commercial software

The SCIGNE Platform

- Advantages
  - Plenty of resources available
  - Stable and powerful environment
  - The SCIGNE engineers are responsive and competent
- Drawbacks
  - Data transfer can take a while
  - Best fit for Linux command line software
  - Best fit for non-commercial software
- Software
  - Occasionally: Proline, Brownotate, NetMHCpan, etc.
  - On a daily basis: Alphafold, ionbot, Dia-NN, Galaxy

AlphaFold

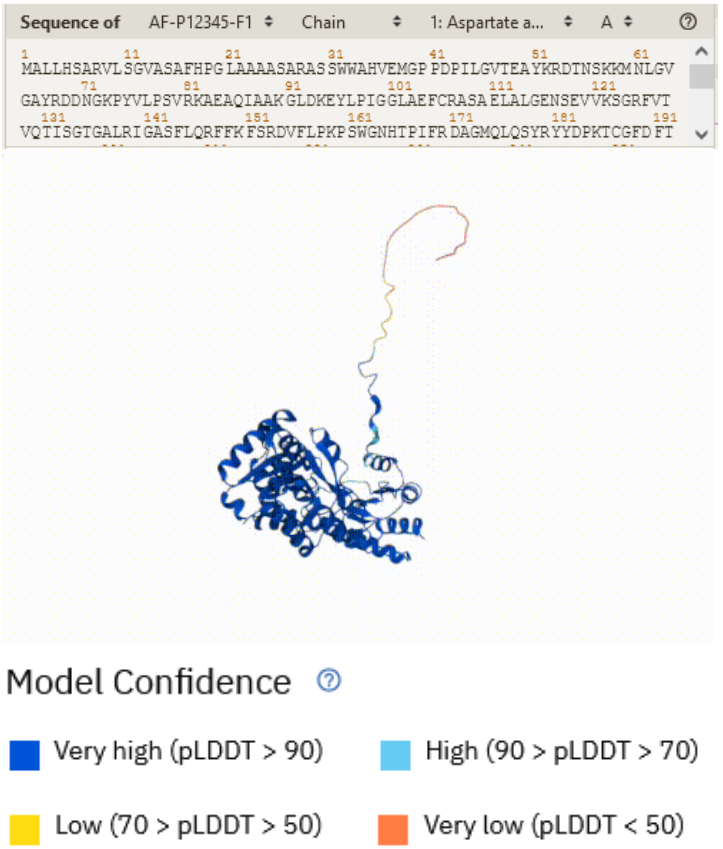
- What is it?
  - Developed by Google DeepMind
  - 2024 Chemistry Nobel Prize laureates
  - Predicts the 3D structure of a protein from its amino acid sequence



Jumper, J et al. Highly accurate protein structure prediction with AlphaFold. Nature (2021)
Varadi, M et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Research (2021)

AlphaFold

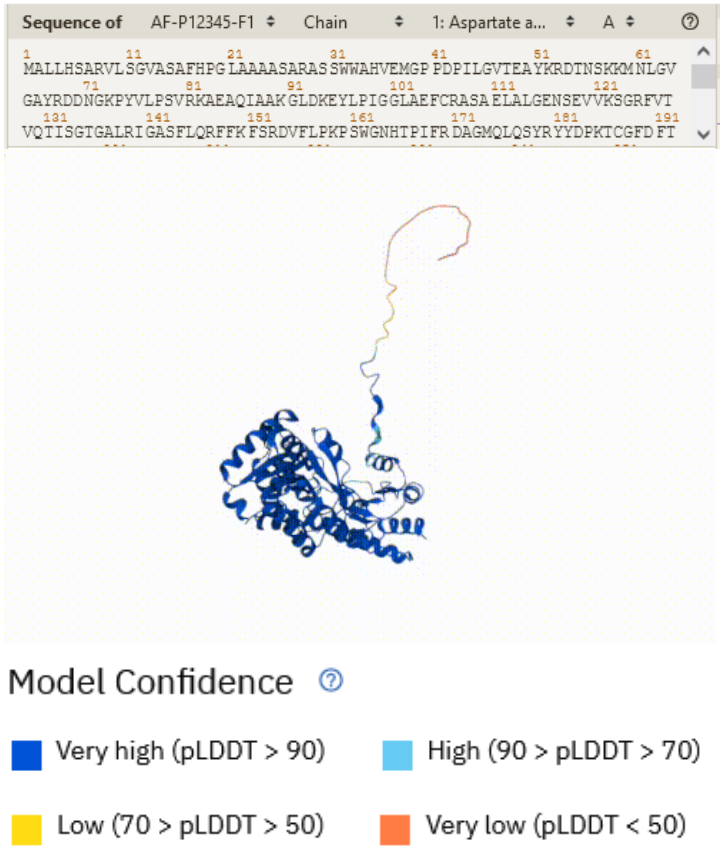
- What is it?
  - Developed by Google DeepMind
  - 2024 Chemistry Nobel Prize laureates
  - Predicts the 3D structure of a protein from its amino acid sequence
- Why is it important?
  - Characterization of proteins and protein complexes
  - Better understanding of the interactions with other molecules
  - Accelerate drug discovery
  - Understanding diseases linked to misfolded proteins



Jumper, J et al. Highly accurate protein structure prediction with AlphaFold. Nature (2021)
Varadi, M et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Research (2021)

AlphaFold

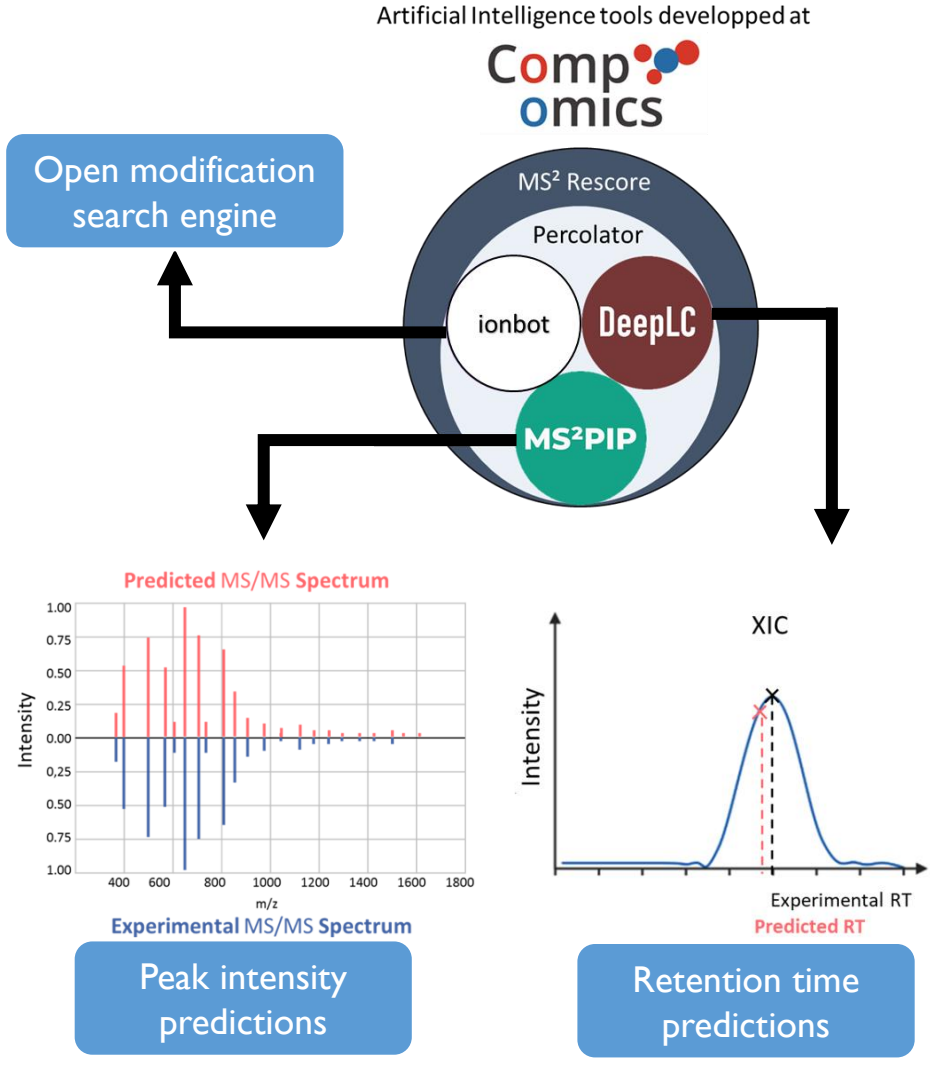
- What is it?
  - Developed by Google DeepMind
  - 2024 Chemistry Nobel Prize laureates
  - Predicts the 3D structure of a protein from its amino acid sequence
- Why is it important?
  - Characterization of proteins and protein complexes
  - Better understanding of the interactions with other molecules
  - Accelerate drug discovery
  - Understanding diseases linked to misfolded proteins
- Why using the Cloud?
  - Study of proteins not predicted yet
  - Not depend on the public server
  - Our VM: 8 CPU, 64GB RAM, 4TB



Jumper, J et al. Highly accurate protein structure prediction with AlphaFold. Nature (2021)
Varadi, M et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Research (2021)

ionbot

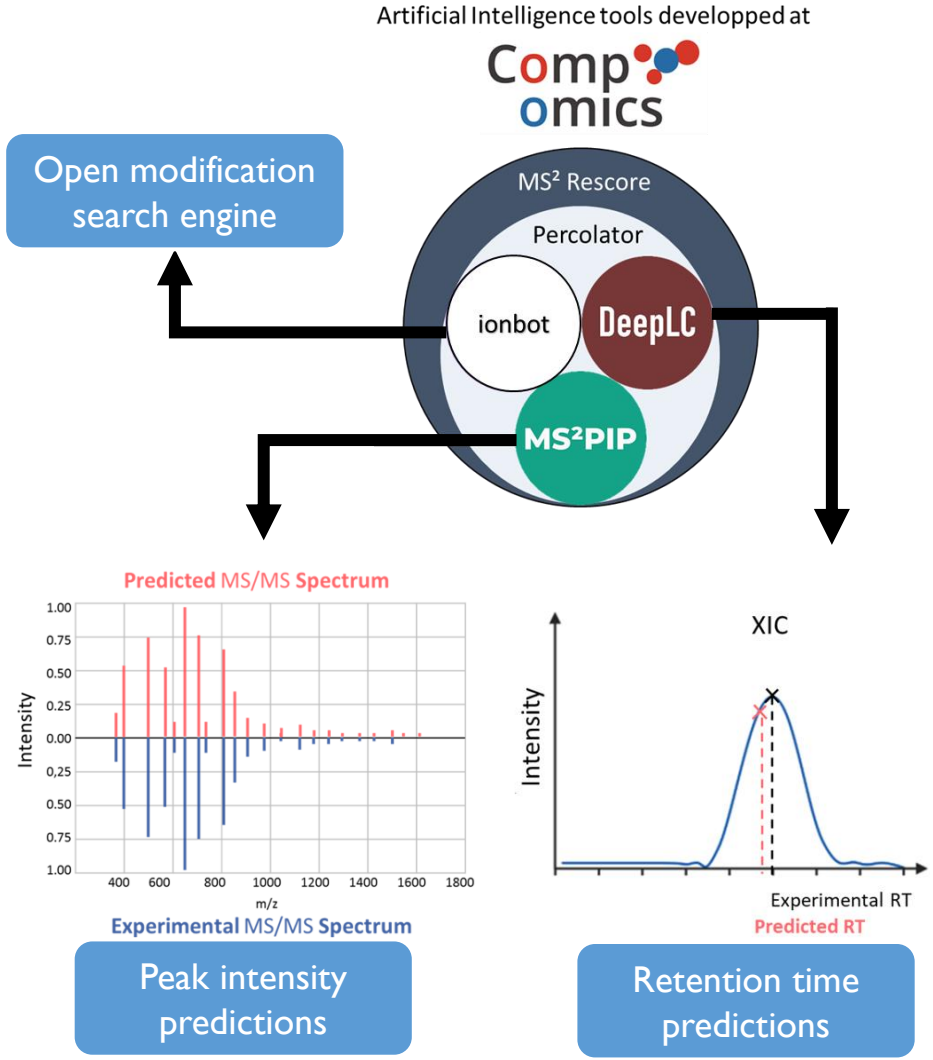
- What is it?
  - Developed by Compomics (VIB, Ghent University, BE)
  - Open modification search engine
  - Using Deep Learning to model the complex behaviour of peptide molecules in a mass spectrometer



MS²Rescore: C. Silva, Bioinformatics (2019), Declercq, MCP (2022)
MS²PIP: R. Gabriels, Nucleic Acids Research (2019)
DeepLC: R. Bouwmeester, Nature Methods (2021);
Percolator: L. Käll, Journal of Proteome Research (2009)

ionbot

- What is it?
  - Developed by Compomics (VIB, Ghent University, BE)
  - Open modification search engine
  - Using Deep Learning to model the complex behaviour of peptide molecules in a mass spectrometer
- Why is it important?
  - Improved identifications (including modifications) and confidence using prediction-based rescoring (0,1% FDR)

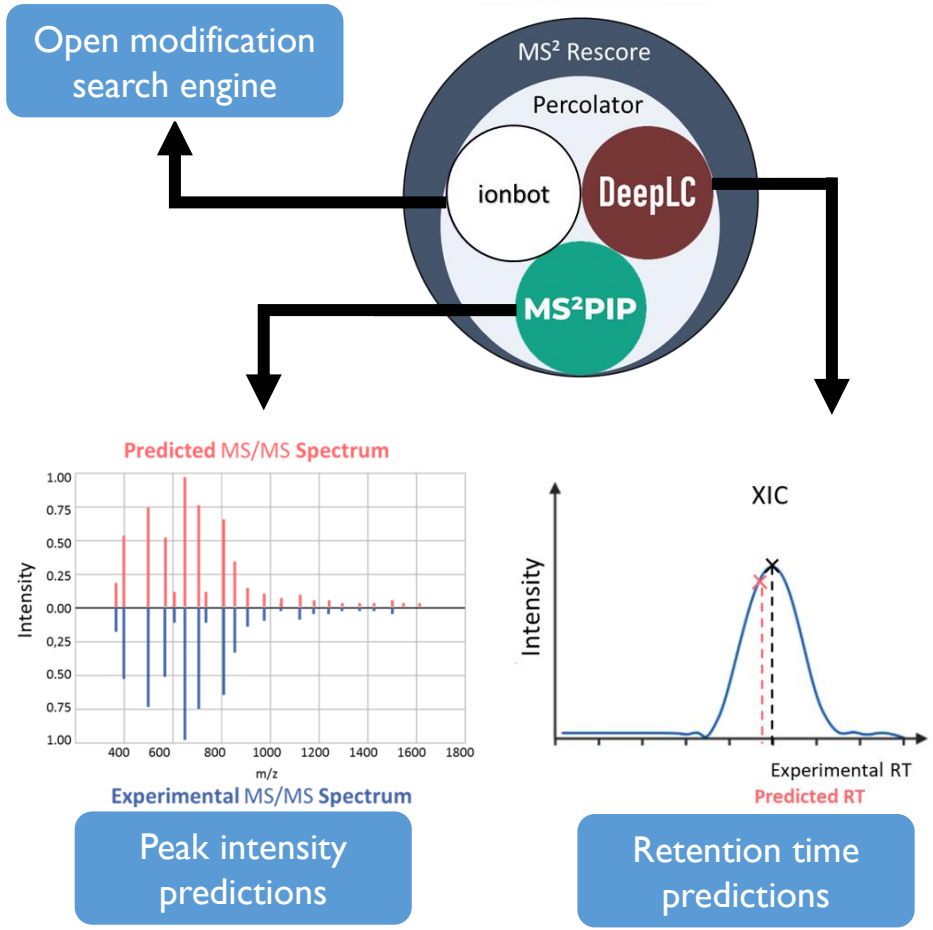


MS²Rescore: C. Silva, Bioinformatics (2019), Declercq, MCP (2022)
MS²PIP: R. Gabriels, Nucleic Acids Research (2019)
DeepLC: R. Bouwmeester, Nature Methods (2021);
Percolator: L. Käll, Journal of Proteome Research (2009)

ionbot

- What is it?
  - Developed by Compomics (VIB, Ghent University, BE)
  - Open modification search engine
  - Using Deep Learning to model the complex behaviour of peptide molecules in a mass spectrometer
- Why is it important?
  - Improved identifications (including modifications) and confidence using prediction-based rescoring (0,1% FDR)
- Why using the Cloud?
  - Need for important resources
  - Searches can take hours in some cases
  - Our VM: 16CPU, 64GB RAM, 1TB

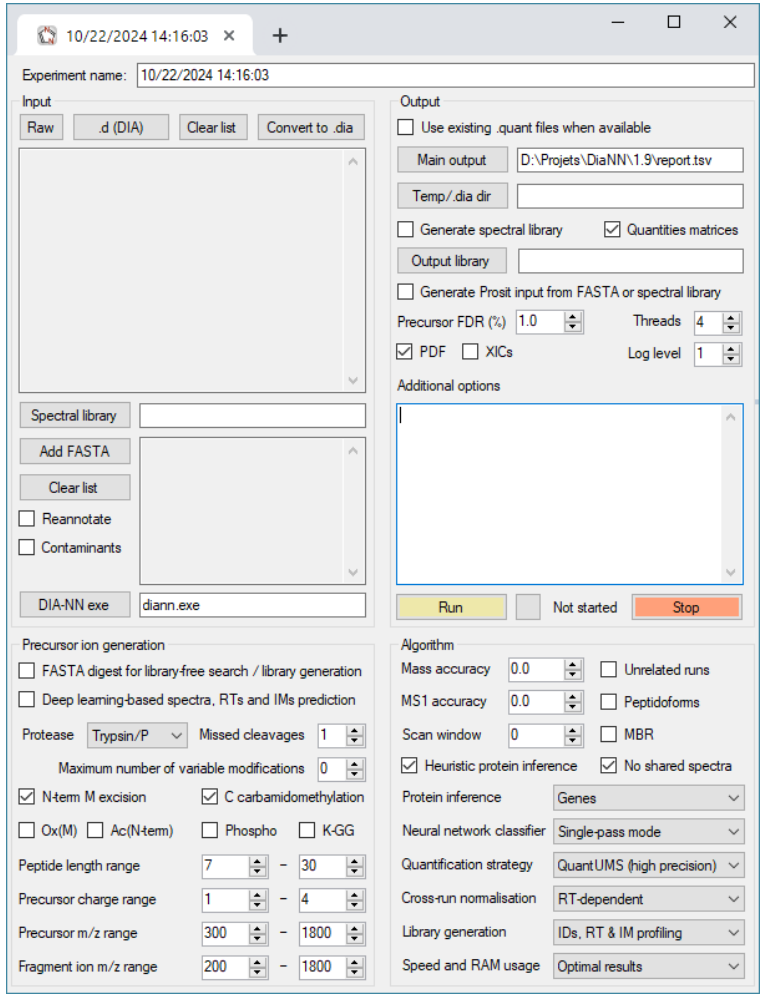
Artificial Intelligence tools developped at Compomics



MS²Rescore: C. Silva, Bioinformatics (2019), Declercq, MCP (2022)
MS²PIP: R. Gabriels, Nucleic Acids Research (2019)
DeepLC: R. Bouwmeester, Nature Methods (2021);
Percolator: L. Käll, Journal of Proteome Research (2009)

Dia-NN

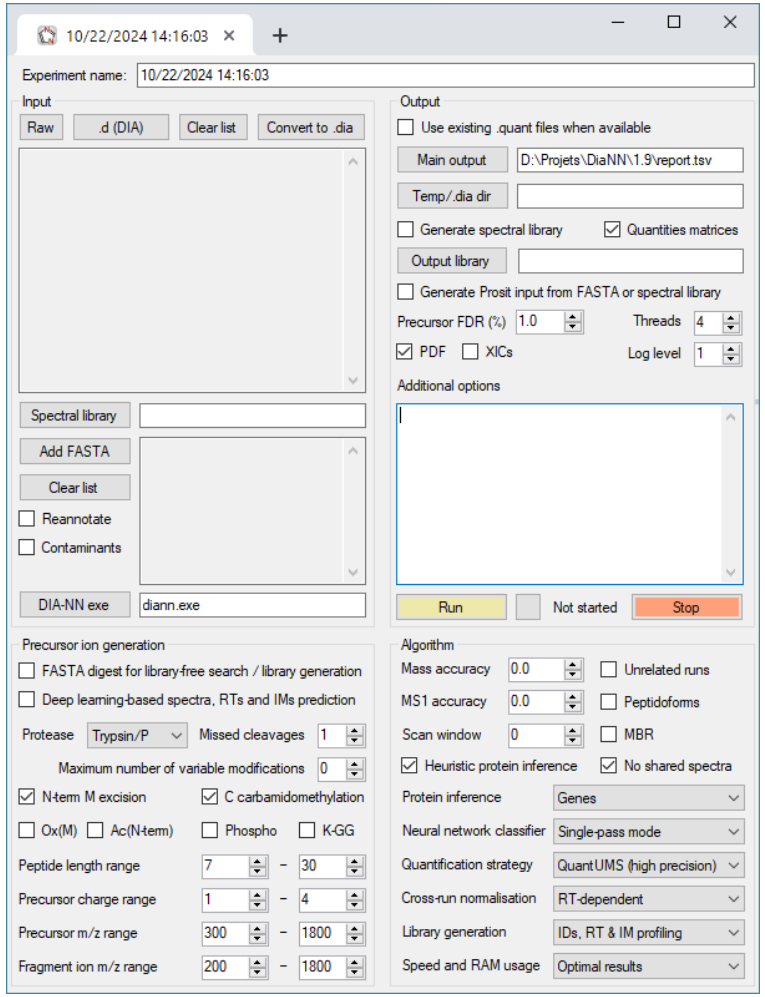
- What is it?
  - Developed by Vadim Demichev (Charité Universitätsmedezin Berlin, DE)
  - Quantify large-scale experiments using neural networks



Demichev et al. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput (Nature Methods, 2020)

Dia-NN

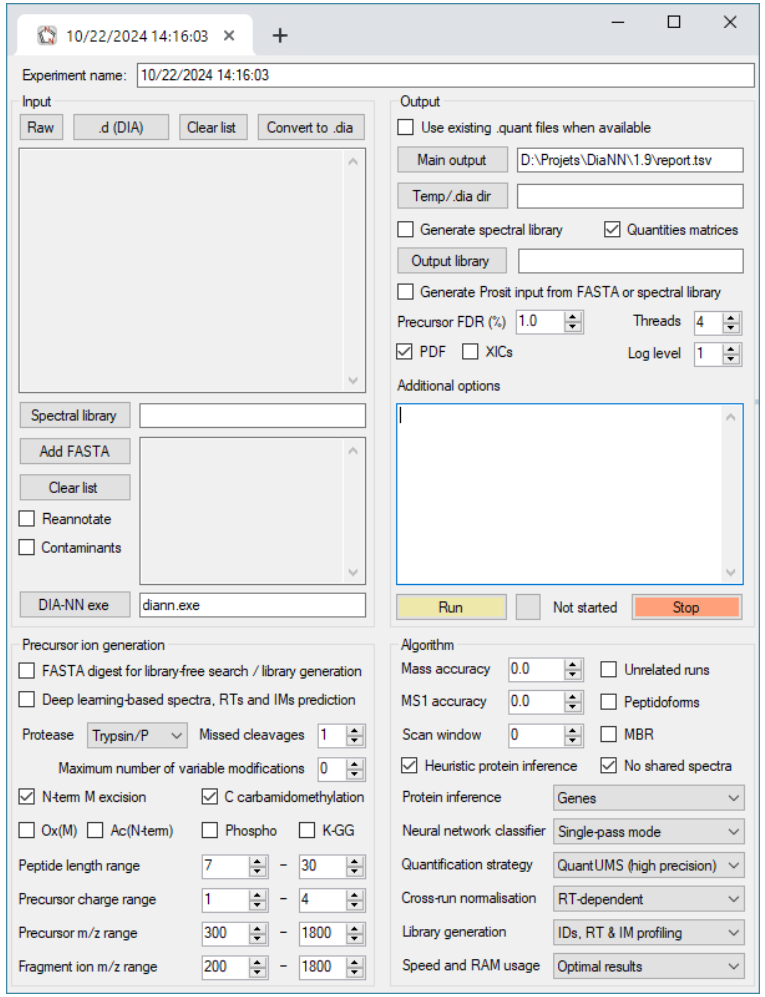
- What is it?
  - Developed by Vadim Demichev (Charité Universitätsmedezin Berlin, DE)
  - Quantify large-scale experiments using neural networks
- Why is it important?
  - Free alternative to commercial solutions, with similar results
  - One of the most widely used software and one of the most demanding



Demichev et al. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput (Nature Methods, 2020)

Dia-NN

- What is it?
  - Developed by Vadim Demichev (Charité Universitätsmedezin Berlin, DE)
  - Quantify large-scale experiments using neural networks
- Why is it important?
  - Free alternative to commercial solutions, with similar results
  - One of the most widely used software and one of the most demanding
- Why using the Cloud?
  - Computation can take weeks on our local servers
  - 2 VMs with 64 CPU and 128GB RAM
  - More VMs are about to be created with a in-house software solution to dispatch the jobs automatically



Demichev et al. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput (Nature Methods, 2020)

Galaxy

- What is it?
  - A Galaxy instance
  - Used to deploy a series of in-house tools developed for diverse purposes

The Galaxy Community , The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update (*Nucleic Acids Research)*
https://galaxyproject.org/

Galaxy

- What is it?
  - A Galaxy instance
  - Used to deploy a series of in-house tools developed for diverse purposes
- Why is it important?
  - Generating protein databanks (Uniprot, NCBI)
  - Searching online databases (Kegg, Gene Ontology, etc.)
  - Run local tools (blastp, fasta36, etc.)

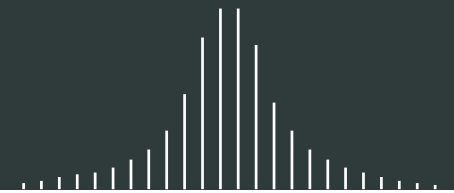
The Galaxy Community , The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update (*Nucleic Acids Research)*
https://galaxyproject.org/

Galaxy

- What is it?
  - A Galaxy instance
  - Used to deploy a series of in-house tools developed for diverse purposes
- Why is it important?
  - Generating protein databanks (Uniprot, NCBI)
  - Searching online databases (Kegg, Gene Ontology, etc.)
  - Run local tools (blastp, fasta36, etc.)
- Why using the Cloud?
  - Convenient to keep these tools in one place
  - Easier maintenance
  - Access from home
  - Our VM: 16CPU, 32Go RAM, 1To

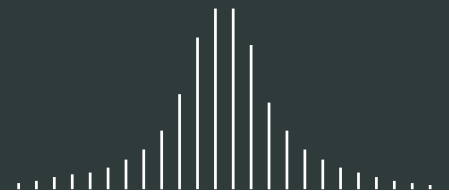
The Galaxy Community , The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update (*Nucleic Acids Research)*
https://galaxyproject.org/

Conclusion



- Proteomics and Mass spectrometry generate huge quantities of data
- Local servers are not able to process such data
- We use the SCIGNE platform everyday

Conclusion



- Proteomics and Mass spectrometry generate huge quantities of data
- Local servers are not able to process such data
- We use the SCIGNE platform everyday
- Side note
  - Most of the recent tools are using Machine Learning
  - Only possible with a large amount of public data to train the models
  - Open Science and FAIR data practices are important!

Conclusion

- Proteomics and Mass spectrometry generate huge quantities of data
- Local servers are not able to process such data
- We use the SCIGNE platform everyday
- Side note
  - Most of the recent tools are using Machine Learning
  - Only possible with a large amount of public data to train the models
  - Open Science and FAIR data practices are important!
- Perspectives
  - Development of Cumulus to dispatch the jobs on different VM on the Cloud
  - RSync client to automate data transfer
  - Will be used for Dia-NN, other software will follow
  - https://github.com/stars/AlexandreBurel/lists/cumulus

Acknowledgments

Directors : Christine Carapito, Sarah Cianferani