MUNI | RECETOX

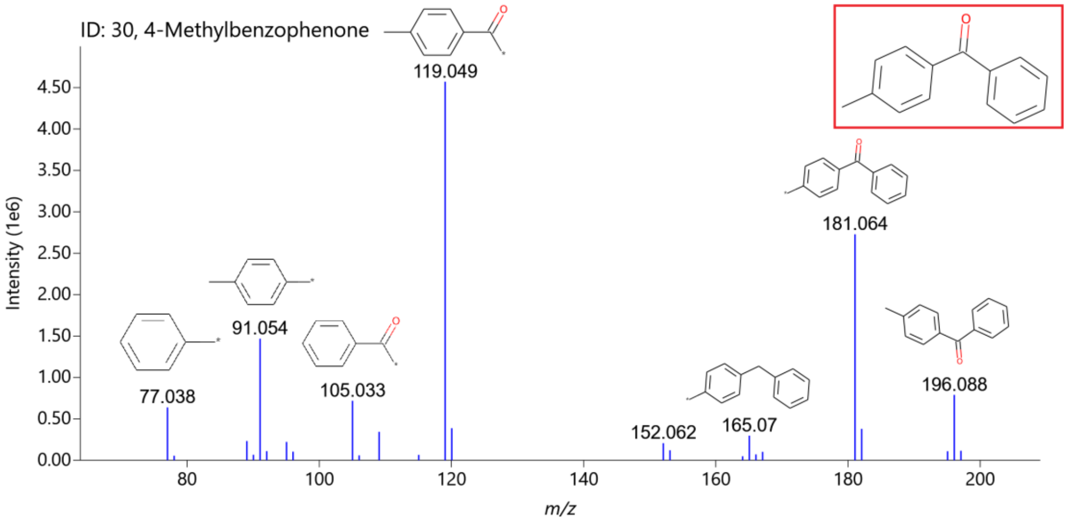
# Galaxy Pipeline & Tool Development for Processing Gas Chromatography – Mass Spectrometry Data

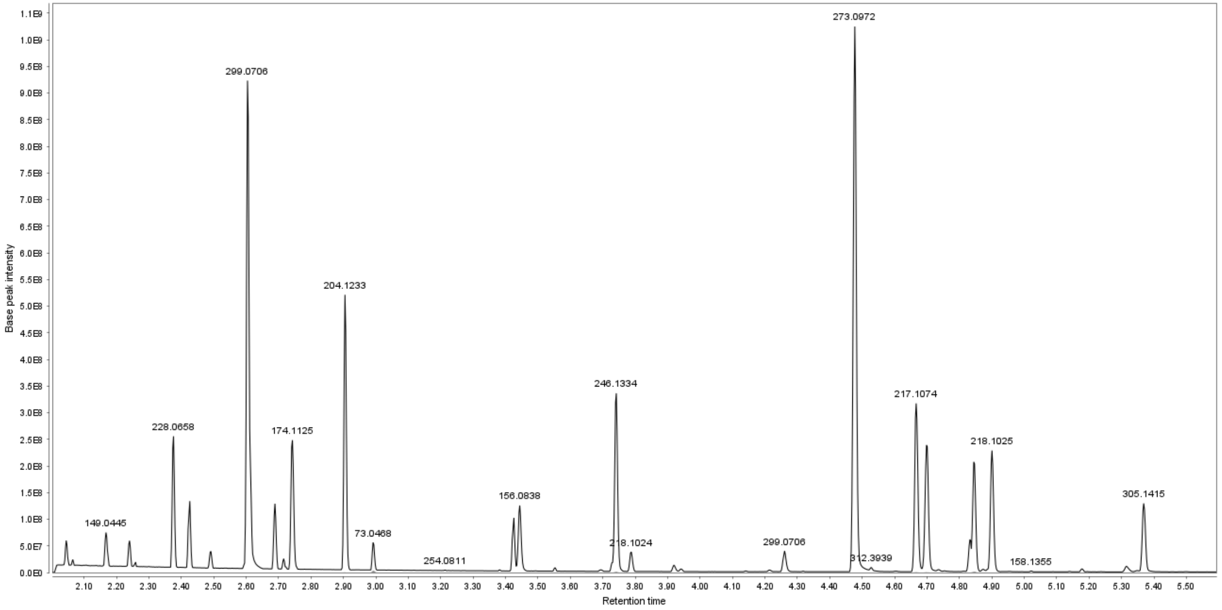
14.09.2021

Helge Hecht

# Overview

1. Introduction
2. State Of The Art
3. Problem Statement
4. Methods
5. Results
6. Summary
7. Future Work
8. Acknowledgements



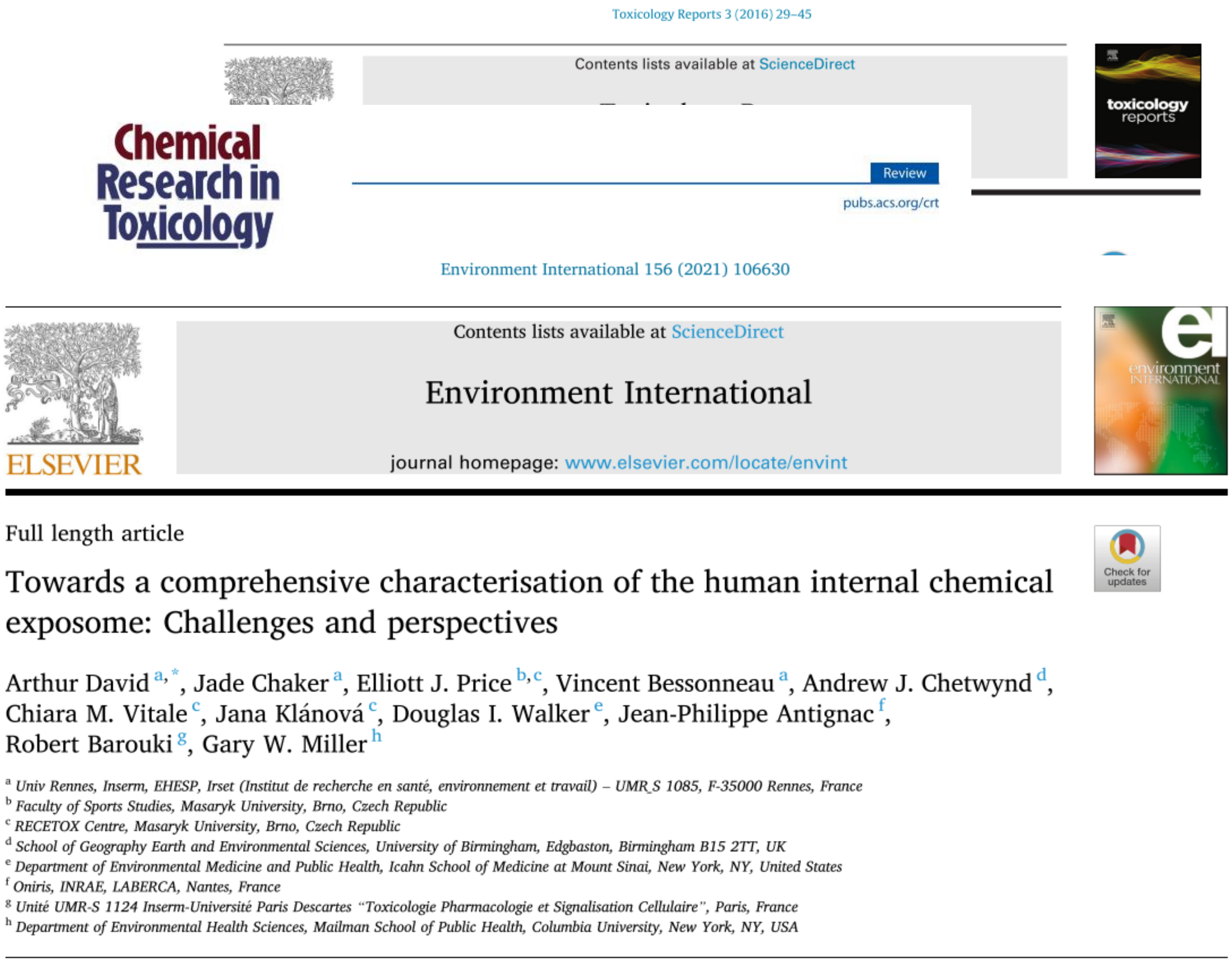


# Introduction

„This is presented **as a call to the international environmental health research community** to champion this effort and work together in this common goal.“ (10.1016/j.toxrep.2015.11.009)

„... have facilitated the detection of tens of thousands of ions, **metabolite identification remains one of the biggest challenges** of available analytical methods.“ (10.1021/acs.chemrestox.6b00179)

„...(iii) the lack of **automation of the annotation/identification** process.“ (10.1016/j.envint.2021.106630)



Toxicology Reports 3 (2016) 29–45

Contents lists available at ScienceDirect

toxicology reports

Chemical Research in Toxicology

Review

pubs.acs.org/crt

Environment International 156 (2021) 106630

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

ELSEVIER

environment INTERNATIONAL

Check for updates

Full length article

Towards a comprehensive characterisation of the human internal chemical exposome: Challenges and perspectives

Arthur David [a,*], Jade Chaker [a], Elliott J. Price [b,c], Vincent Bessonneau [a], Andrew J. Chetwynd [d], Chiara M. Vitale [c], Jana Klánová [c], Douglas I. Walker [e], Jean-Philippe Antignac [f], Robert Barouki [g], Gary W. Miller [h]

[a] Univ Rennes, Inserm, EHESP, Irset (Institut de recherche en santé, environnement et travail) – UMR_S 1085, F-35000 Rennes, France
[b] Faculty of Sports Studies, Masaryk University, Brno, Czech Republic
[c] RECETOX Centre, Masaryk University, Brno, Czech Republic
[d] School of Geography Earth and Environmental Sciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK
[e] Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York, NY, United States
[f] Oniris, INRAE, LABERCA, Nantes, France
[g] Unité UMR-S 1124 Inserm-Université Paris Descartes "Toxicologie Pharmacologie et Signalisation Cellulaire", Paris, France
[h] Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University, New York, NY, USA

# State of the Art

# State of the Art – GUI-based Tools

**Good**

- easy to use
- work well as standalone tools

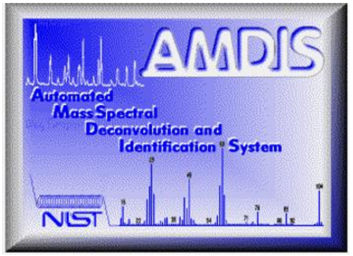
**Bad**

- not scalable → no distributed computation
- tight coupling of GUI and backend
- bad library support → programming overhead
- mostly focus on LC-MS

MZmine 2

MZmine2 (10.1186/1471-2105-11-395)

SIRIUS (10.1093/bioinformatics/btn603)


AMDIS
Automated Mass Spectral Deconvolution and Identification System
NIST

AMDIS (10.1016/S1044-0305(99)00047-1)

MS-FINDER (10.1021/acs.analchem.6b00770)

MS-DIAL (10.1038/nmeth.3393)

# State of the Art – Web-based tools

## Good

- easy to use
- partially scalable

## Bad

- data storage → sensitive data?
- difficult to modify individual steps
- little resources for GC-MS

GNPS (10.1038/nbt.3597)

MetaboAnalyst (10.1093/nar/gkab382)

XCMS Online (10.1021/ac300698c)

# State of the Art – Coded Workflows

## Good

- fairly easy to extend & modify
- scalable
- good library support

## Bad

- varying (often poor) quality
- hard to use
- low reproducibility
- poorly integrated → group specific

Bioconductor (10.1093/nar/gkab382)

RforMassSpectrometry

Bioconda (10.1038/s41592-018-0046-7)

# State of the Art – Galaxy

**Good**

- easy to use
- scalable & modular
- data management

**Bad**

- focus on LC-MS
- varying tool quality
- different application domain

PhenoMeNal (10.1093/gigascience/giy149)

W4M (10.1093/bioinformatics/btu813; 10.1016/j.biocel.2017.07.002)

# Problem Statement

We need data processing pipelines that are

1. easy to use, understand & access,
2. built for large-scale analysis,
3. including various modules & steps,
4. creating reproducible results,

consisting of tools that are

1. well tested & documented,
2. easy to extend & modify based on requirements,
3. specific to the research domain.

# Methods

We implement Galaxy pipelines using new tools that are

1. tailored for user needs & domain problems,
2. developed open-source,
3. according to professional software standards

**Good**

- best of all worlds
- long-term solution
- links to other infrastructures

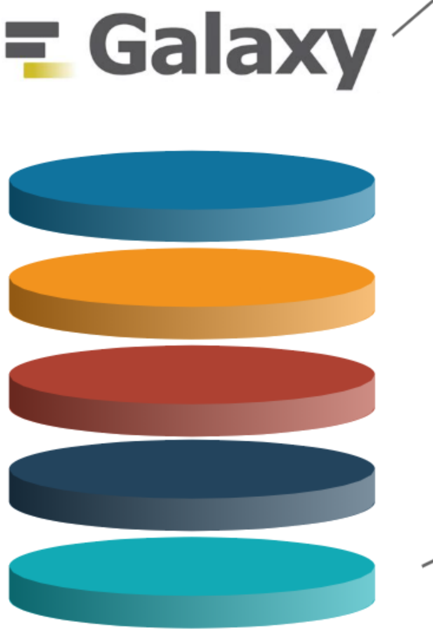
& State-of-the-Art packages with modifications to

1. test their behaviour,
2. make them easier to use & understand,
3. make them scalable.

**Bad**

- hard to achieve
- high complexity
- expensive

# Methods



**Access**

User-friendly web interface

**Workflows**

Modular tools to build multi-step processing pipelines

**Visualizations**

In-browser and interactive tools for data inspection

**Storage**

Secure data storage with federated identity & access management for individual users or groups

**Reproducibility**

Tracking of actions and tool metadata in a history that can be transformed into a workflow, published, shared or persisted on disk.

**Distributed Computing**

Implicit parallelization via job distribution for shorter runtimes

# Methods – Galaxy Tool Development

- virtualization → docker & Biocontainers (10.1093/bioinformatics/btx192)
- open-source hosting via GitHub
- testing (testthat; pytest) with widely supported frameworks & code coverage
- static code analysis (sonarcloud)
- tools according to IUC guidelines

# Results

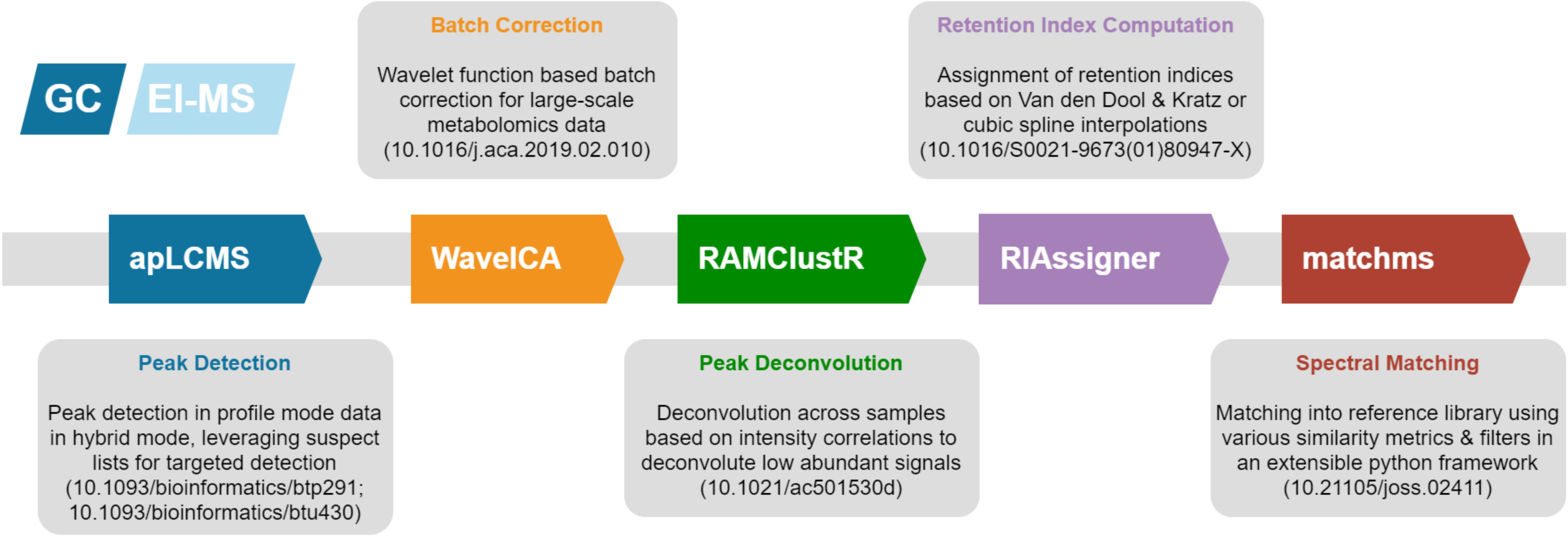
galaxy workflow for GC-MS processing

tools that are complementary to existing resources

two standalone tools extracted as extensible modules

contribution to existing open-source software

# Results – Galaxy Workflow



**Batch Correction**

Wavelet function based batch correction for large-scale metabolomics data (10.1016/j.aca.2019.02.010)

**Retention Index Computation**

Assignment of retention indices based on Van den Dool & Kratz or cubic spline interpolations (10.1016/S0021-9673(01)80947-X)

**Peak Detection**

Peak detection in profile mode data in hybrid mode, leveraging suspect lists for targeted detection (10.1093/bioinformatics/btp291; 10.1093/bioinformatics/btu430)

**Peak Deconvolution**

Deconvolution across samples based on intensity correlations to deconvolute low abundant signals (10.1021/ac501530d)

**Spectral Matching**

Matching into reference library using various similarity metrics & filters in an extensible python framework (10.21105/joss.02411)

Galaxy (cerit-sc.cz)

# Results - Tools

RIAssigner

- read & write data in various formats (csv & msp) using matchms & pandas
- extensible data & computation modules
- published via Bioconda
- makes data comparable by aligning based on high- confidence annotations

pyMSPannotator

- add various metadata fields to mass spectral libraries
- extends functionality of *webchem* to python
- leverages IDSM (10.1186/s13321-021-00515-1) service for PubChem → query via API
- first step of improved high-resolution filtering workflow

# Results – Capacity Building

- participation in Galaxy Metabolomics Community calls & member in ELIXIR Metabolomics Community
- participation in de.nbi network events (metaRbolomics)
- Netherlands metabolomics infrastructure (eScienceCenter → matchms)
- contributions to Galaxy Training Network
- member in US Thermo GC Orbitrap working group, BP4NTA, mQACC

# Summary

- strong need for automation & harmonization in data processing
- state-of-the-art tools are scattered
- lack of high-quality resources
- Galaxy as platform for harmonization of large scale analysis
- rapid progress (compared to others!)
- high quality developments take time
- potential for publication of workflow & tools

# Future Work

- integration of complementary tools → ADAP-GC4.0 (10.1021/acs.analchem.9b01424), NormAE (10.1021/acs.analchem.9b05460) etc.
- additional steps → reporting (10.1021/acs.analchem.8b04310, biotransformation, prediction of RI (10.1016/j.aca.2020.12.043)
- applying machine learning techniques
- experimenting with similarity scores (10.1371/journal.pcbi.1008724; 10.1101/2021.04.18.440324) & molecular networking
- workflow for improved high-resolution filtering

# Acknowledgements

Martin Čech
Elliott James Price
Aleš Křenek
Jiří Novotný
Maksym Skoryk
Matej Troják
Ondřej Melichar
Gabriela Karásková

Vojtěch Bartoň
Jana Klánová
Muhammad Usman
Zdenka Dudová
Karolína Trachtová

RECETOX research infrastructure supported by the Ministry of Education, Youth and Sports of the Czech Republic (LM2018121) and funding from the CETOCOEN EXCELLENCE Teaming 2 project supported by Horizon 2020 (857560) and the Ministry of Education, Youth and Sports of the Czech Republic (02.1.01/0.0/0.0/18_046/0015975).

Computational resources provided by the e-INFRA CZ (LM2018140).

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

MUNI
ICS

MUNI | RECETOX

# Thank you for your attention!

Questions?