An Open Software Development-based Ecosystem of R Packages for Metabolomics Data Analysis

Johannes Rainer[1], Philippine Louail[1], Andrea Vicini[2], Roger Gine[3], Josep M Badia[3,4] , Michele Stravs[5,6], Mar Garcia-Aloy[7], Carolin Huber[8], Liesa Salzer[9], Jan Stanstrup[10], Nir Shahaf[11], Christian Panse[12,13], Thomas Naake[14], William Kumler[15], Pablo Vangeenderhuysen[16], Carl Brunius[17], Helge Hecht[18], Steffen Neumann[19], Michael Witting[20], Sebastian Gibb[21], Laurent Gatto[2]

[1]Institute for Biomedicine, Eurac Research, Italy. [2]Computational Biology and Bioinformatics, de Duve Institute, UCLouvain, Belgium. [3]Department of Electronic Engineering & IISPV, Universitat Rovira i Virgili, Spain. [4]CIBERDERM, Instituto de Salud Carlos III, Spain. [5]Department of Environmental Chemistry, Switzerland. [6]Institute of Molecular Systems Biology, ETH Zurich, Switzerland. [7]Metabolomics Unit, Research and Innovation Centre, Fondazione Edmund Mach, Italy. [8]Department of Effect Directed Analysis, Helmholtz Center for Environmental Research, Germany. [9]Research Unit Analytical BioGeoChemistry, Helmholtz Munich, Germany. [10]Department of Nutrition, Exercise and Sports, University of Copenhagen, Denmark. [11]Department of Plant and Environmental Sciences, Weizmann Institute of Science, Israel. [12]Functional Genomics Center Zurich, University of Zurich/ETH Zurich, Switzerland. [13]Swiss Institute of Bioinformatics, Switzerland. [14]Genome Biology Unit, EMBL, Germany. [15]Ingals Lab, School of Oceanography, University of Washington, USA. [16]Laboratory of Integrative Metabolomics, Department of Translational Physiology, Infectiology and Public Health, Faculty of Veterinary Medicine, Ghent University, Belgium. [17]Department of Life Sciences, Chalmers University of Technology, Sweden. [18]RECETOX, Faculty of Science, Masaryk University, Czech Republic. [19]Computational Plant Biochemistry, MetaCom, Leibniz Institute of Plant Biochemistry, Germany. [20]Metabolomics and Proteomics Core, Helmholtz Munich, Germany. [21]Anesthesiology and Intensive Care Medicine, University Hospital Greifswald, Germany.

In a nutshell

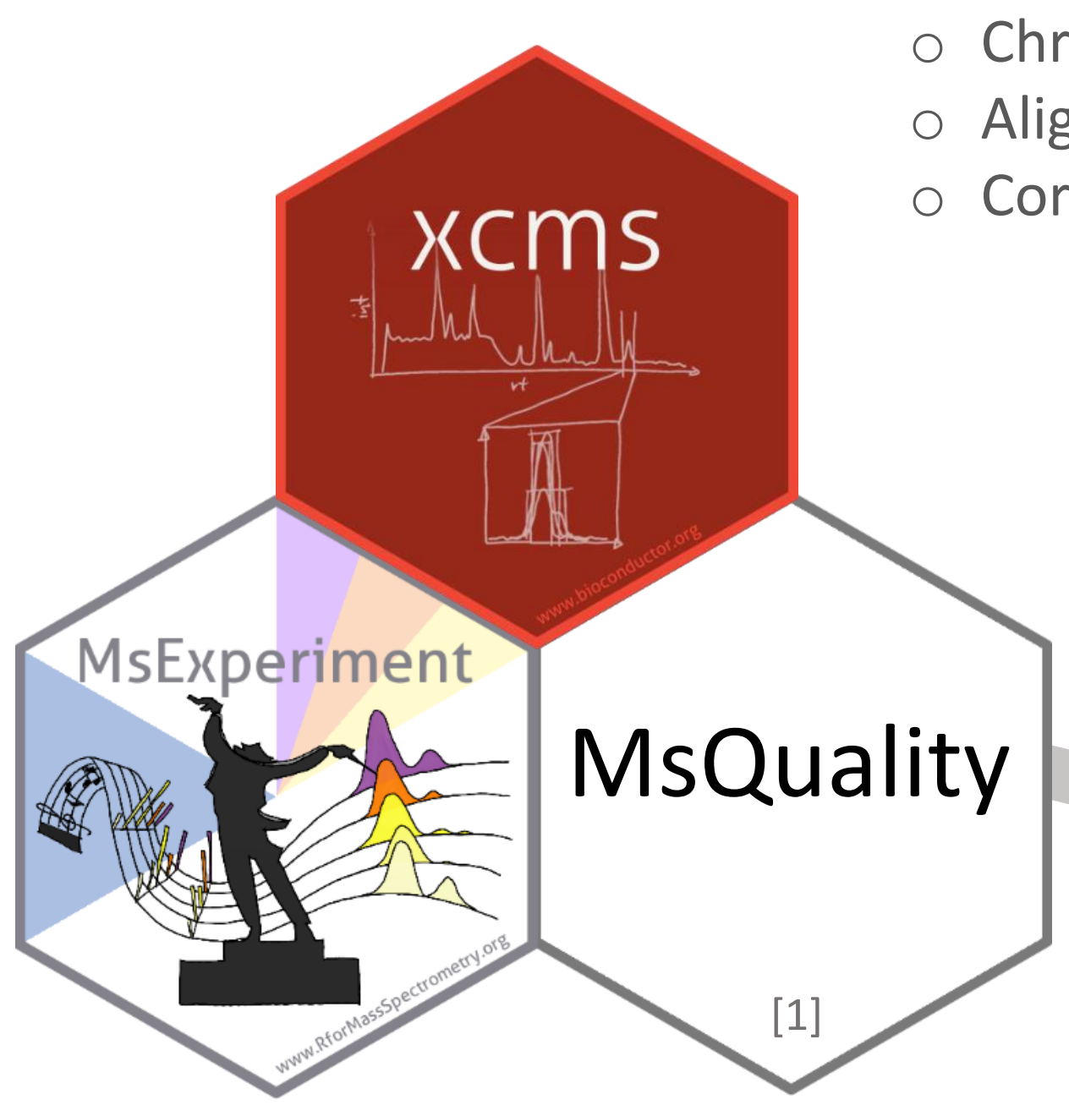
Get in touch/contribute

Poster on Zenodo

… to see it in action

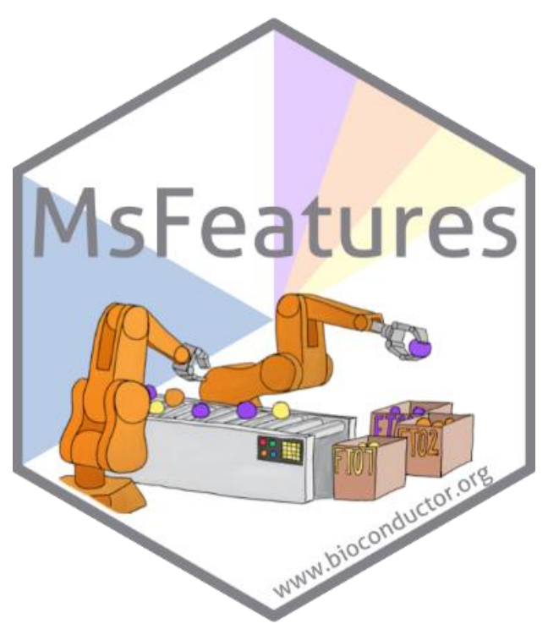
- Powerful and flexible infrastructure for efficient analysis of large-scale MS data
- Open collaborative development of software for MS data analysis:
  - Well documented and thoroughly tested
  - Long-term support and maintenance
- Modular package ecosystem to allow creation of custom analysis workflows
- Tightly integrated into Bioconductor

LC-MS(/MS) preprocessing



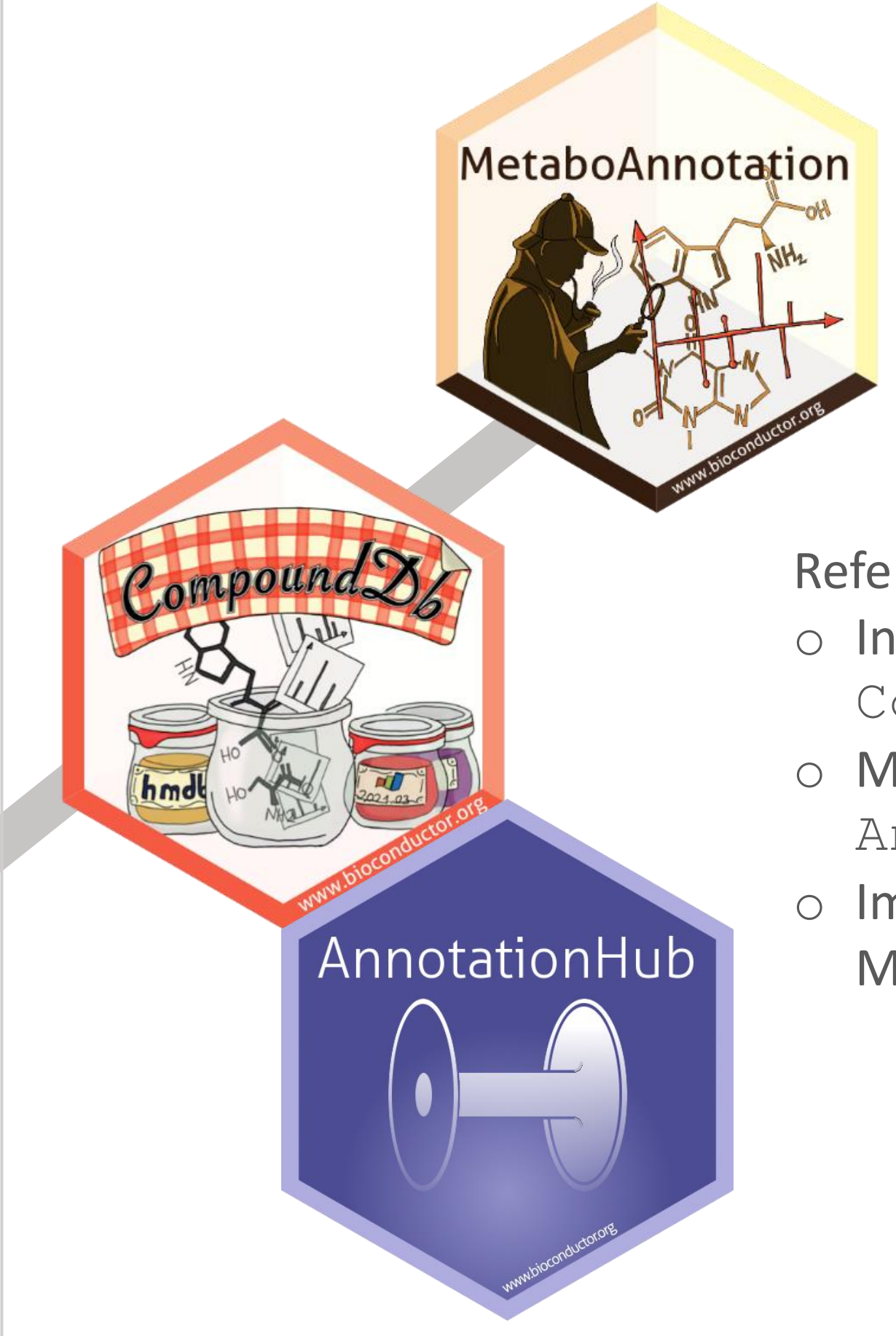
- Quality assessment
- Chromatographic peak detection
- Alignment
- Correspondence

Compounding



- Group adducts
- Group isotopes

Annotation [2]

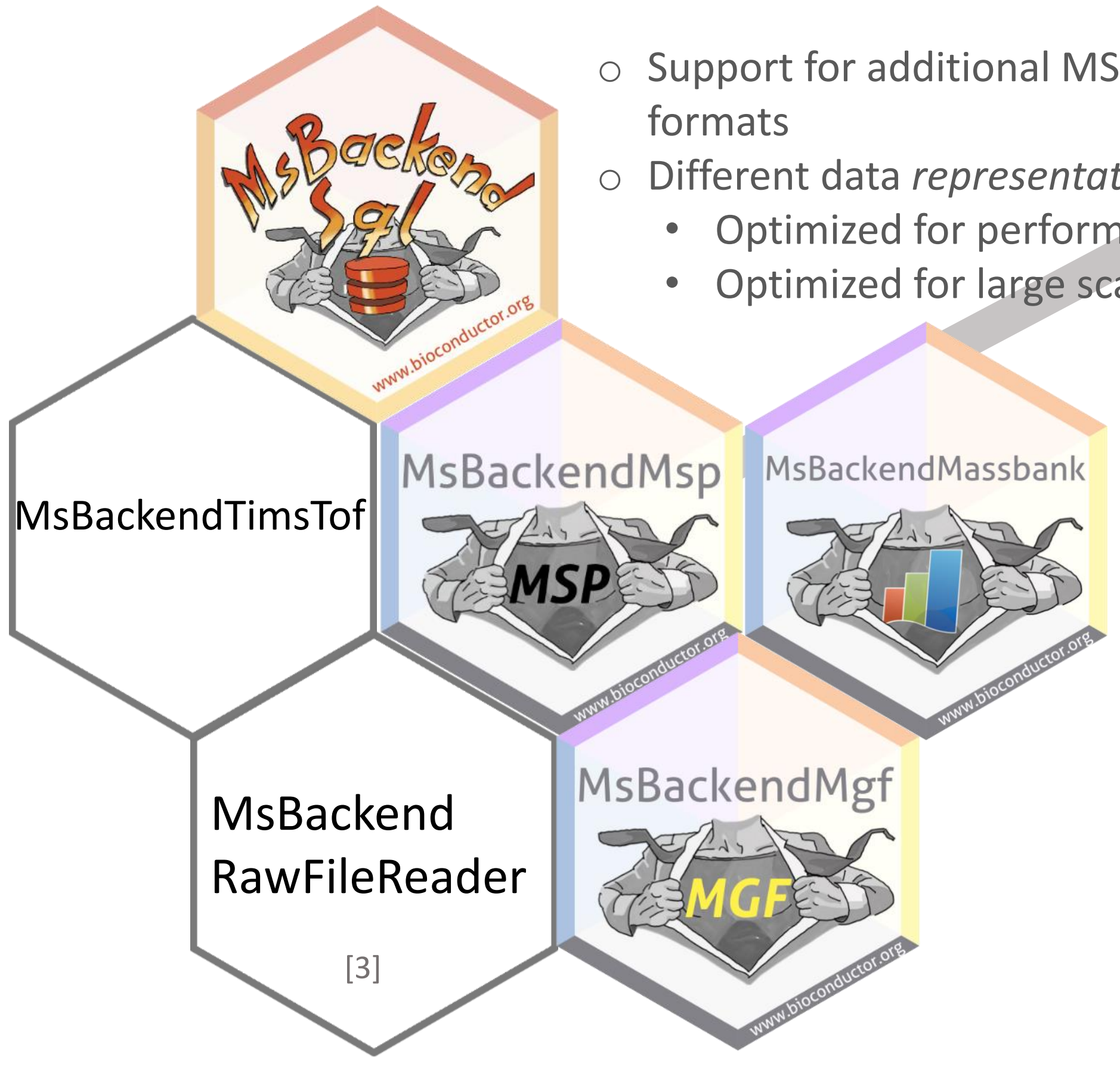


- MS1 and MS2-based annotation
- Fragment spectra matching

Reference libraries

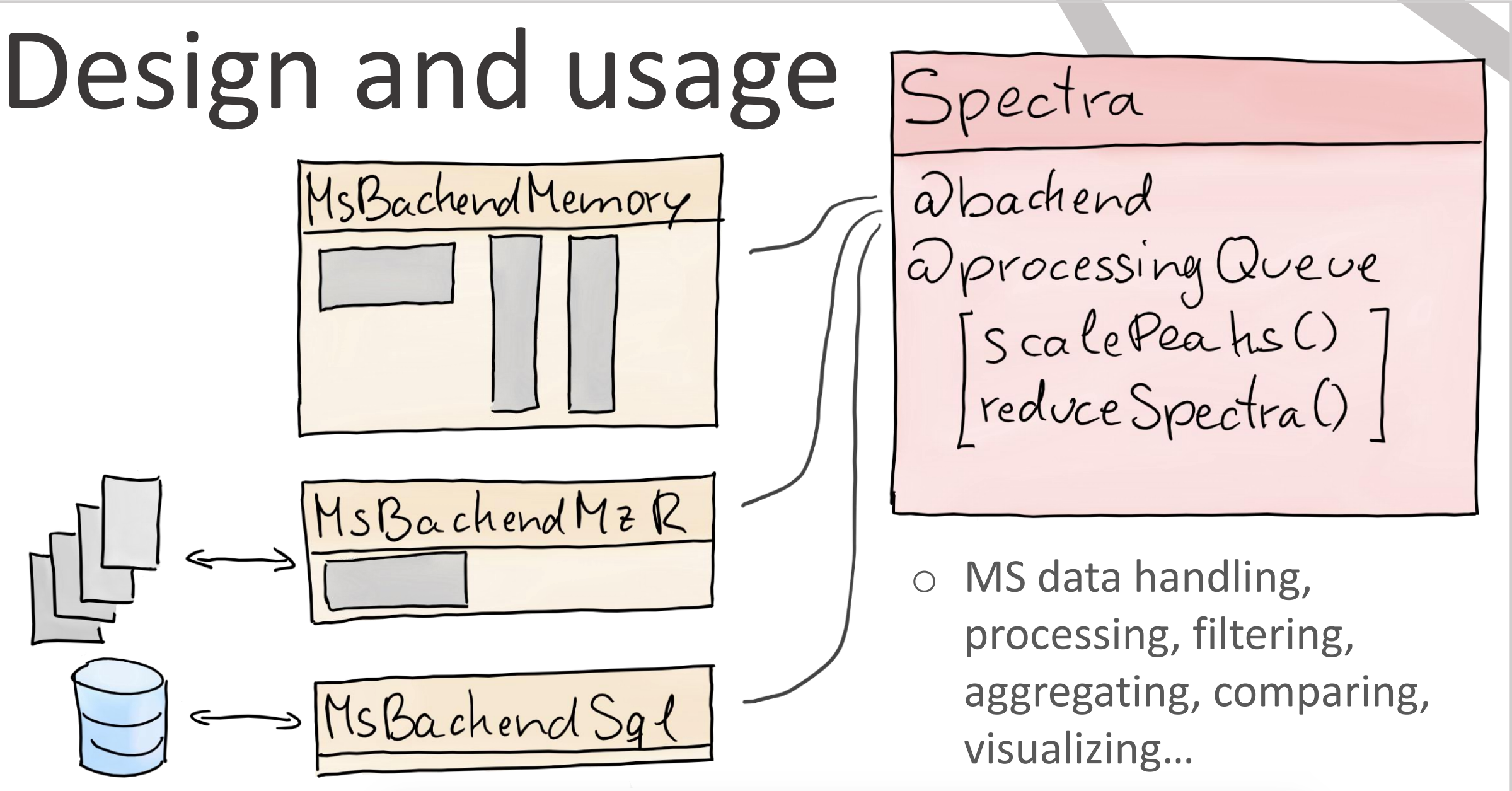
- In-house databases through `CompoundDb`
- MassBank through Bioconductor's `AnnotationHub`
- Import **GNPS** or others from MGF, MSP files.

Data I/O and representation



- Support for additional MS data file formats
- Different data *representations* :
  - Optimized for performance
  - Optimized for large scale data

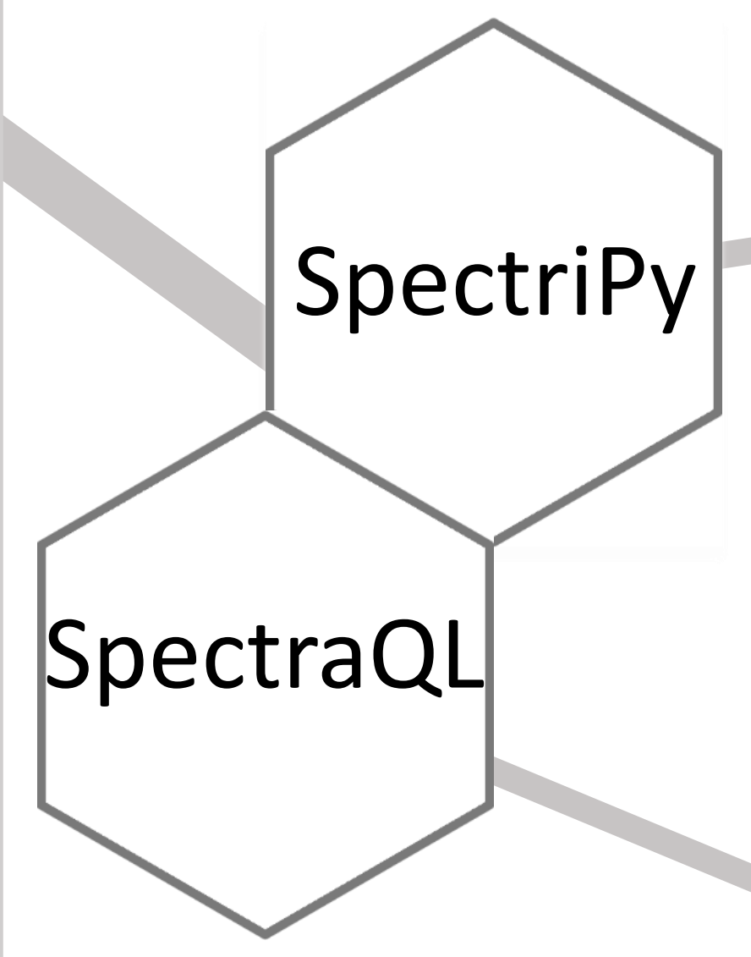
Design and usage



- MS data handling, processing, filtering, aggregating, comparing, visualizing…

```
> filterMsLevel(sps, msLevel = 1L) |>
+     scalePeaks() |>
+     reduceSpectra()
```

Integration with external tools



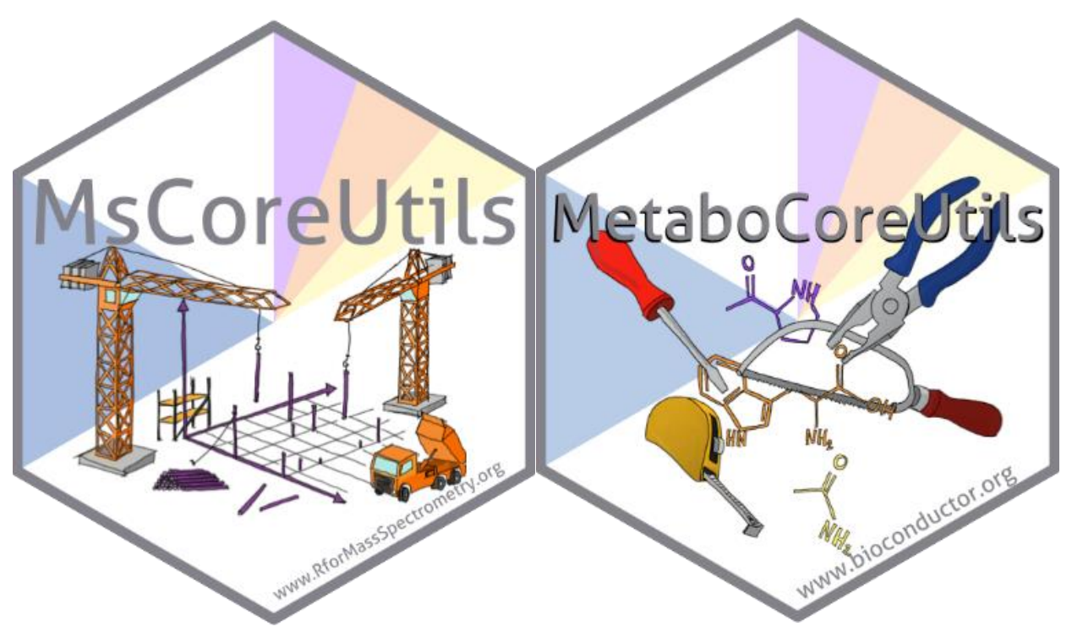
python

MATCHMS [4]

- Integrate/use mstchms-based similarity scores for fragment spectra
- Query MS data using the MassQL query language in R.

MassQL [5]

Core *low-level* functions



- Spectra similarity scores
- Peak alignment
- Handling chemical formulas
- Isotope peak prediction
- Adduct/exact mass calculations
- etc

Outlook

- Continue integration of R/Bioconductor packages and external software
- Emphasis on seamless integration to improve user experience
- Open for contributions from the community -> reach out via github issue or email

References

[1] MsQuality – an interoperable open-source package for the calculation of standardized quality metrics of mass spectrometry data. Naake T et al. Bioinformatics 2023. https://doi.org/10.1093/bioinformatics/btad618
[2] A Modular and Expandable Ecosystem for Metabolomics Data Annotation in R. Rainer J et al. *Metabolites* 2022. https://doi.org/10.3390/metabo12020173
[3] The rawrr R Package: Direct Access to Orbitrap Data and Beyond. Kockmann T et al. Journal of Proteome Research 2021. https://doi.org/10.1021/acs.jproteome.0c00866
[4] matchms – processing and similarity evaluation of mass spectrometry data. Huber F et al. JOSS 2020. https://doi.org/10.21105/joss.02411
[5] A Universal Language for Finding Mass Spectrometry Data Patterns. Jarmusch AK et al. bioRxiv 2022. https://doi.org/10.1101/2022.08.06.503000

@jorainer@fosstodon.org

https://doi.org/10.5281/zenodo.11370345

eurac research