The Journal of Open Source Software

patRoon 2.0: Improved non-target analysis workflows including automated transformation product screening

Rick Helmus*[1], Bas van de Velde[1, 2, 3], Andrea M. Brunner[2, 4], Thomas L. ter Laak[1, 2], Annemarie P. van Wezel[1], and Emma L. Schymanski[5]

1 Institute for Biodiversity and Ecosystem Dynamics (IBED), University of Amsterdam, P.O. Box 94240, 1090 GE Amsterdam, The Netherlands. **2** KWR Water Research Institute, Chemical Water Quality and Health, P.O. Box 1072, 3430 BB Nieuwegein, The Netherlands. **3** Division of BioAnalytical Chemistry, Amsterdam Institute of Molecular and Life Sciences (AIMMS), Vrije Universiteit Amsterdam, 1081 HV Amsterdam, The Netherlands. **4** TNO, Environmental Modelling, Sensing and Analysis (EMSA), Princetonlaan 8, 3584 CB, Utrecht, The Netherlands. **5** Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, L-4367 Belvaux, Luxembourg.

DOI: 10.21105/joss.03978

Software

- Review
- Repository
- Archive

Editor: Pending Editor

Submitted: 03 December 2021
Published: 03 December 2021

Summary

Non-target analysis (NTA) via chromatography coupled to high resolution mass spectrometry (HRMS) is used to monitor and identify organic chemicals in the environment. Biotic and abiotic processes can transform original chemicals (parents) into *transformation products* (TPs). These TPs can be of equal or more concern than their parent compounds and are therefore critical to monitor and identify in the environment (Escher & Fenner, 2011; Farré et al., 2008), often with NTA. Given the amount of data generated by NTA, advanced automated data processing workflows are essential. The open-source, `R`-based (R Core Team, 2021) platform `patRoon` (Helmus, ter Laak, et al., 2021) offers automated, straightforward, flexible and comprehensive NTA workflows.

This article describes improvements introduced in `patRoon 2.0`, including extensive TP screening and simultaneous processing of positive and negative HRMS data. The updated documentation and code are available via https://rickhelmus.github.io/patRoon and archived in Helmus, Velde, et al. (2021).

Statement of need

The identification of chemicals in NTA still remains a grand challenge (Vermeulen et al., 2020); only a small percentage of detected masses can be confidently annotated with spectral libraries (Silva et al., 2015). The unidentified "dark matter" is partly due to TPs, motivating the need for TP screening workflows. Reported approaches include screening of known/predicted TPs, parent/TP classification techniques, isotope labeling experiments and identifying expected (dis)similarities in MS data (shown in Table 1; also reviewed in Li et al. (2021)). However, these approaches are typically designed for a single study or available only as a stand-alone and/or commercial tool. `patRoon 2.0` implements a consistent interface to the complementary approaches from Table 1 and other novel functionality to provide comprehensive TP screening workflows.

*corresponding author, r.helmus@uva.nl

The Journal of Open Source Software

Table 1: Implemented TP screening approaches.

Approach	Principle	Requirements	Examples
TP suspect screening	Screen known/predicted TPs	Knowledge base/study applicable prediction tools, parent structures	Djoumbou-Feunang et al. (2019), Wicker et al. (2015), Schymanski, Kondić, et al. (2021)
Metabolic logic	Screen molecular mass differences from known transformations	Known (elemental) transformations	Schollée et al. (2015)
Mass spectral	Cluster similar MS data	Correlation structural and spectral similarity. Availability of fragmentation spectra for respective parents/TPs	Schollée et al. (2017), Treutler et al. (2016), Naake & Gaquerel (2017), Depke et al. (2017)
Classification	Parent/TP classification	In-house statistical/computational expertise	Schollée et al. (2015), Brunner et al. (2019), Schollée et al. (2021)

Since wide chemical coverage is desired with NTA and since TPs can ionize differently to their parent, HRMS analyses are often performed using positive and negative ionization mode. `patRoon 2.0` is now capable of simultaneously processing, integrating and interpreting mixed mode data - a functionality not available in most workflows due to complexity and long processing times.

Further improvements to `patRoon` include interactive data curation and new prioritization and identification strategies, described further below.

New functionality

Transformation product screening workflow

The `patRoon 2.0` TP screening workflow starts with features (data points with unique chromatographic/MS information) obtained from a 'classical' `patRoon` workflow (Figure 1A). Then, data from one or more of TP screening (B,C), MS similarity (D) and parent/TP feature classification (E) is combined to link parent and TP features into components (F). The resulting data is then prioritized (G), corresponding features are annotated (H), and finally all data is reported (I). All algorithm parameters are configurable, yet simplified via defaults. This enables flexible and customizable workflows for a wide variety of applications.

TP screening uses known/predicted TP structures from parents (Figure 1B) or mass differences of transformations using metabolic logic (Schollée et al., 2015) (C). Parents for (B) are specified from (1) a target list, (2) results of a *suspect screening* to find parents by mass, or (3) candidates of feature compound annotation (see Helmus, ter Laak, et al. (2021)). Corresponding TP structures are then either obtained *in silico* with BioTransformer (Djoumbou-Feunang et al., 2019), or through a library search from PubChem data (Kim, 2021; Krier et al., 2022; Schymanski, Kondić, et al., 2021; Schymanski, Bolton, et al., 2021) or a custom-made library. Metabolic logic (C), which does not depend on parent structural data, uses transformation reactions from Schollée et al. (2015) or a custom-made list. TP suspect screening then matches candidate TPs with detected features by mass.

MS similarity (Figure 1D) is calculated, without a predefined parent list, from spectral match and/or equivalence of spectral annotations. Spectral match compares MS fragment spectra

(MS/MS) with a cosine or Jaccard index similarity score (Stein & Scott, 1994). This was largely implemented in C++ to allow efficient comparison of large numbers of spectra (typically thousands). The calculation can be adjusted by (1) pre-treatment of spectra, e.g. with peak count and intensity thresholds, (2) weight assignment to intensity and *m/z* data, and (3) shifting TP spectra to highlight equal neutral losses (Schollée et al., 2017; Watrous et al., 2012). Furthermore, combining matched mass peaks from shifted and non-shifted spectra was implemented for similarity calculation of equivalent fragments *and* neutral losses. MS similarity from annotation equivalence compares formulas of annotated MS/MS fragments and neutral losses, based on additional data such as isotopic fit and spectral libraries. This potentially increases accuracy, but requires presence of annotations for parent/TP features.

Parent/TP feature classification (Figure 1E) is typically performed by statistical analyses with R, facilitated by the `patRoon` data export functionality. Fold-change calculation and visualization with volcano plots (Cui & Churchill, 2003) was implemented in `patRoon 2.0` to simplify the usage of this common classification technique.

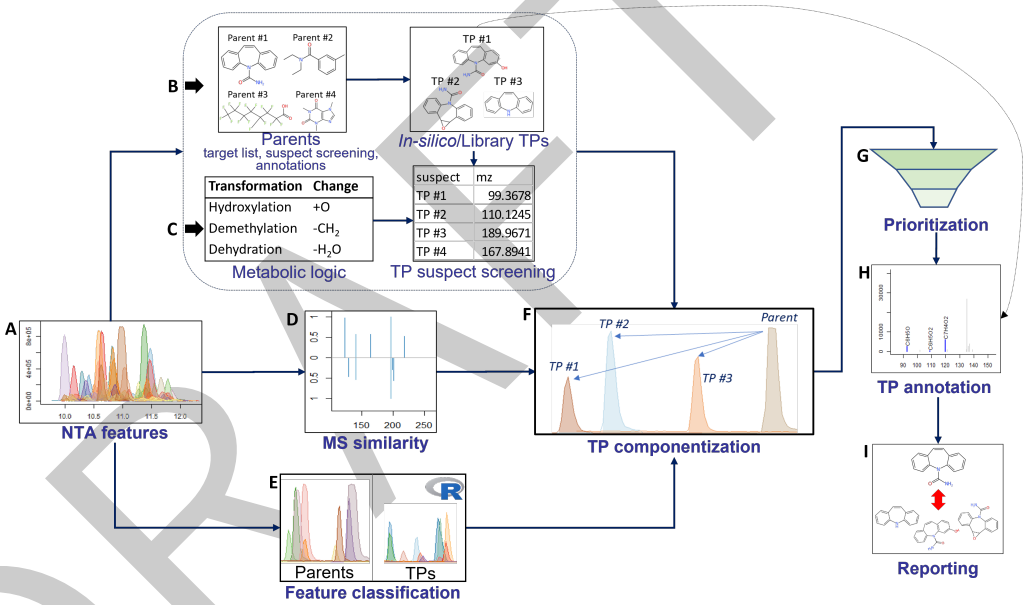

Figure 1: TP screening workflow in `patRoon 2.0`. One or more of steps B/C, D and E are used to generate TP components by linking and grouping parent/TP features (F). The TP annotation (H) can be enriched with data from (B).

During TP componentization (Figure 1F), each parent feature is linked with corresponding TP features and grouped in a TP component. Data prioritization (G) can then be performed with the subsetting functionality of `patRoon` and several newly implemented filters (Table 2). Existing MetFrag (Ruttkies et al., 2016) annotation functionality was extended to include predicted TP structures (B) to allow *in silico* MS/MS annotation (H) of TPs absent from commonly used databases. The interactive reporting (I) functionality was extended to simplify inspection of TP screening results (see Figure 2).

Table 2: Filters to prioritize TP components.

Filter	Remarks
RT direction	Verifies expected relative retention time direction of each parent/TP pair based on chemical polarities.
Structure similarity	Removes TPs with high MS similarity but low structural similarity.
Explained transformation	Verifies the proposed metabolic logic transformation with feature formula annotations.
Remove isomers	Removes isomers, which can be difficult or impossible to distinguish with MS.
Duplicates	Removes duplicate TPs formed from the same parent through different pathways.

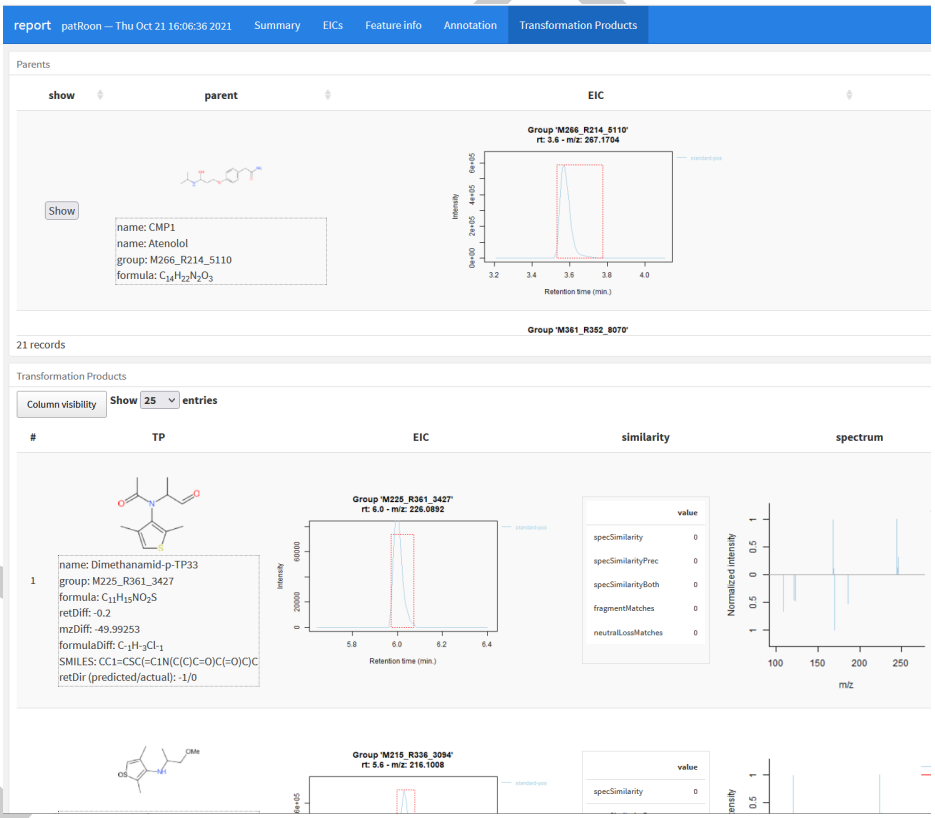

Figure 2: Example report with TP screening results (bottom) for a selected parent (top).

Sets workflows: combining positive and negative MS ionization data

In a *sets workflow*, positive and negative data is automatically processed and combined (Figure 3A). Features are obtained for each polarity, and optionally prioritized with polarity specific conditions (e.g. minimum intensity). Then, the feature *m/z* values are replaced with neutral masses calculated from adduct information (defined manually or via feature adduct annotations), and subsequently aligned and grouped across polarities (with configurable tolerances). Subsequent steps largely follow the `patRoon 1.0` workflow (Helmus, ter Laak, et al., 2021). Algorithms incapable of processing polarity mixed data are automatically executed with polarity specific data, and outputs are subsequently combined. Moreover, a consensus

for formula/compound annotations can be reached, for instance, to eliminate candidates not found for both polarities.

Sets workflows follow a generic design, where each *set* is a group of analyses that demand independent processing of MS related data (features, mass spectra etc). Therefore, sets can also be differentiated by other MS parameters such as MS/MS fragmentation technique or ionization source. Furthermore, the design allows future implementation of workflows with different chromatographic methods, for instance, to simultaneously process data from different instruments.

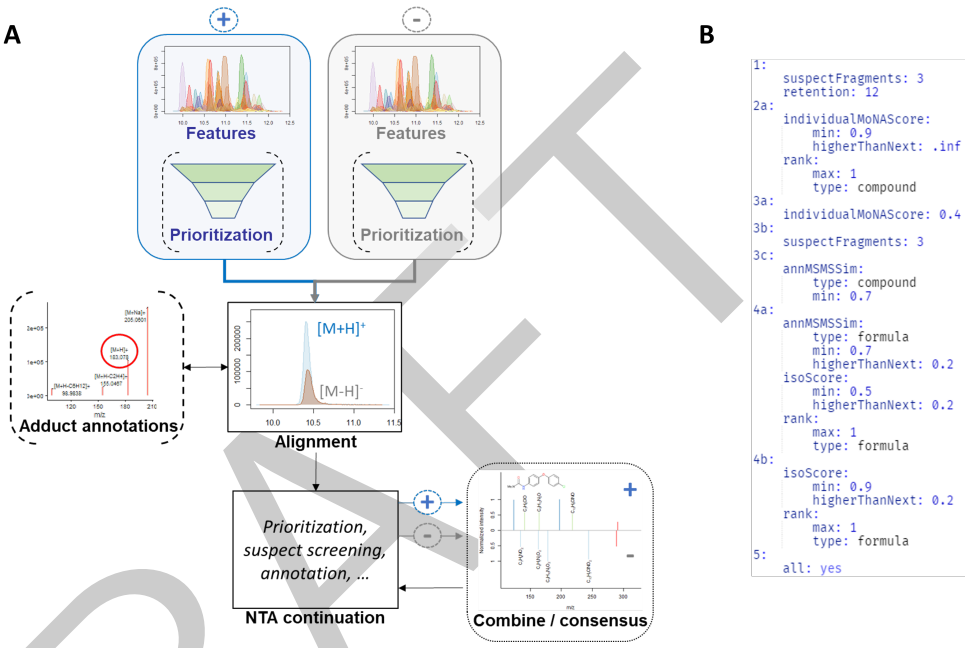

Figure 3: A Sets workflow with simultaneous processing of positive and negative data. Alignment of positive/negative features can be improved with adduct annotations. The workflow continues identically to the `patRoon 1.0` workflow, and positive/negative data is automatically processed separately for algorithms without mixed polarity support. **B** Default `YAML` configuration file used for estimation of suspect identification levels from annotation scores, candidate rankings and other data.

Other new functionality

Other new functionality of `patRoon 2.0` includes:

- Improved suspect screening
  - Automatic estimation of identification levels (Schymanski et al., 2014) using a configurable and extensible rule based approach (see Figure 3B).
  - Combining suspect and non-target screening workflows.
  - Merging results from different screenings.
- Improved adduct annotation
  - Automatic prioritization of features with preferred adducts.
  - Use of adduct annotations with formula/compound annotation.
  - Support for the algorithms of cliqueMS (Senan et al., 2019) and OpenMS (`MetaboliteAdductDecharger`) (Röst et al., 2016).
- Improved feature data

  - Inclusion of SIRIUS (Dührkop et al., 2019), SAFD (Samanipour et al., 2019) and KPIC2 (Ji et al., 2017) algorithms.
  - Integration of MetaClean (Chetnik et al., 2020) for chromatographic peak quality calculation and validation.
  - Calculation and prioritization with peak scores derived from aforementioned peak qualities.
- Interactive graphical tools to inspect and curate workflow data and to train and inspect feature classifications with `MetaClean`.
- Refactoring and updates of `newProject()` to generate code for the new functionality.
- A `delete` function to remove unwanted workflow data, e.g. to implement custom filters.
- More approaches to parallelize `R` code and support high performance computing using the future package (Bengtsson, 2020).
- Bug fixes and improvements from user feedback.

A complete listing of all changes is outlined the project news file.

Example workflows

Simultaneous processing of positive/negative data

Performing a sets workflow is straightforward, and requires only few additions to a `patRoon 1.0` workflow.

```
# Load patRoon and patRoonData libraries.
library(patRoon)
library(patRoonData)

# Obtain features for positive/negative data
fListPos <- findFeatures(exampleAnalysisInfo("positive"), "openms")
fListNeg <- findFeatures(exampleAnalysisInfo("negative"), "openms")

# Initiate sets workflow with default (de-)protonated adducts
fListSet <- makeSet(fListPos, fListNeg, adducts = c("[M+H]+", "[M-H]-"))

# Neutralize features and group across analyses and sets
fGroups <- groupFeatures(fListSet, "openms")

# Perform prioritization, annotation, reporting etc as a 'classical' workflow
...
```

TP screening

The code below demonstrates a simple TP screening workflow where (1) parents are screened, (2) corresponding TPs are predicted with `BioTransformer`, (3) the TPs are screened, (4) TP components are generated and (5) all results are reported.

```
# (1) Screen parents
parentSuspects <- data.frame(name = c("Carbamazepine", "Benzotriazole"),
                             SMILES = c("C1=CC=C2C(=C1)C=CC3=CC=CC=C3N2C(=O)N",
                                        "C1=CC2=NNN=C2C=C1"))
fGroupsScr <- screenSuspects(fGroups, parentSuspects, adduct = "[M+H]+")
```

```
# (2) Predict TPs with BioTransformer
TPs <- generateTPs("biotransformer", parents = fGroupsScr)

# (3) Screen for the TPs and amend previous results
TPSuspects <- convertToSuspects(TPs)
fGroupsScr <- screenSuspects(fGroupsScr, TPSuspects, adduct = "[M+H]+",
                             amend = TRUE)
fGroupsScr <- filter(fGroupsScr, onlyHits = TRUE)

# (4) Generate the TP components
componentsTP <- generateComponents(fGroupsScr, "tp", TPs = TPs)

# (5) Generate interactive HTML report
reportHTML(fGroupsScr, components = componentsTP)
```

The code for these and more advanced workflows are easily generated with the `newProject()` function. The Handbook outlines more examples of typical TP screening workflows.

Acknowledgements

The authors thank Carl Emil Eskildsen for his suggestions on implementing and evaluating different MS similarity approaches, Lisa Kooy for her feedback and testing of the TP screening functionality, Pim de Voogt for general supervision and constructive feedback on the manuscript, as well as María Domínguez Román for her feedback on the manuscript.

Funding

R. Helmus was funded by the Institute of Biodiversity and Ecosystem Dynamics (University of Amsterdam) and the Netherlands Organisation for Scientific Research (NWO) as part of the TooCOLD project carried out within the framework of TTW Open Technology Programme (project number 15506). A.M. Brunner was funded by the Joint Research Program of the Dutch and Belgian drinking water companies. E.L. Schymanski was funded by the Luxembourg National Research Fund (FNR) for project A18/BM/12341006.

References

Bengtsson, H. (2020). *A unifying framework for parallel and distributed processing in R using futures*. https://arxiv.org/abs/2008.00553

Brunner, A. M., Vughs, D., Siegers, W., Bertelkamp, C., Hofman-Caris, R., Kolkman, A., & Laak, T. ter. (2019). Monitoring transformation product formation in the drinking water treatments rapid sand filtration and ozonation. *Chemosphere*, *214*, 801–811. https://doi.org/10.1016/j.chemosphere.2018.09.140

Chetnik, K., Petrick, L., & Pandey, G. (2020). MetaClean: A machine learning-based classifier for reduced false positive peak detection in untargeted LCMS metabolomics data. *Metabolomics*, *16*(11). https://doi.org/10.1007/s11306-020-01738-3

Cui, X., & Churchill, G. A. (2003). Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology*, *4*(4). https://doi.org/10.1186/gb-2003-4-4-210

Depke, T., Franke, R., & Brönstrup, M. (2017). Clustering of MS2 spectra using unsupervised methods to aid the identification of secondary metabolites from pseudomonas aeruginosa. *Journal of Chromatography B*, *1071*, 19–28. https://doi.org/10.1016/j.jchromb.2017.06.002

Djoumbou-Feunang, Y., Fiamoncini, J., Gil-de-la-Fuente, A., Greiner, R., Manach, C., & Wishart, D. S. (2019). BioTransformer: A comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *Journal of Cheminformatics*, *11*(1). https://doi.org/10.1186/s13321-018-0324-5

Dührkop, K., Fleischauer, M., Ludwig, M., Aksenov, A. A., Melnik, A. V., Meusel, M., Dorrestein, P. C., Rousu, J., & Böcker, S. (2019). SIRIUS 4: A rapid tool for turning tandem mass spectra into metabolite structure information. *Nature Methods*, *16*(4), 299–302. https://doi.org/10.1038/s41592-019-0344-8

Escher, B. I., & Fenner, K. (2011). Recent advances in environmental risk assessment of transformation products. *Environmental Science & Technology*, *45*(9), 3835–3847. https://doi.org/10.1021/es1030799

Farré, M. la, Pérez, S., Kantiani, L., & Barceló, D. (2008). Fate and toxicity of emerging pollutants, their metabolites and transformation products in the aquatic environment. *TrAC Trends in Analytical Chemistry*, *27*(11), 991–1007. https://doi.org/10.1016/j.trac.2008.09.010

Helmus, R., ter Laak, T. L., van Wezel, A. P., de Voogt, P., & Schymanski, E. L. (2021). patRoon: Open source software platform for environmental mass spectrometry based non-target screening. *Journal of Cheminformatics*, *13*(1). https://doi.org/10.1186/s13321-020-00477-w

Helmus, R., Velde, B. van de, Brunner, A. M., Laak, T. L. ter, Wezel, A. P. van, & Schymanski, E. L. (2021). *patRoon 2.0: Improved non-target analysis workflows including automated transformation product screening* (Version 2.0.0) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.5752071

Ji, H., Zeng, F., Xu, Y., Lu, H., & Zhang, Z. (2017). KPIC2: An effective framework for mass spectrometry-based metabolomics using pure ion chromatograms. *Analytical Chemistry*, *89*(14), 7631–7640. https://doi.org/10.1021/acs.analchem.7b01547

Kim, S. (2021). Exploring chemical information in PubChem. *Current Protocols*, *1*(8), e217. https://doi.org/https://doi.org/10.1002/cpz1.217

Krier, J., Singh, R. R., Kondić, T., Lai, A., Diderich, P., Zhang, J., Thiessen, P. A., Bolton, E. E., & Schymanski, E. L. (2022). Discovering pesticides and their TPs in luxembourg waters using open cheminformatics approaches. *Environment International*, *158*, 106885. https://doi.org/https://doi.org/10.1016/j.envint.2021.106885

Li, D., Liang, W., Feng, X., Ruan, T., & Jiang, G. (2021). Recent advances in data-mining techniques for measuring transformation products by high-resolution mass spectrometry. *TrAC Trends in Analytical Chemistry*, *143*, 116409. https://doi.org/https://doi.org/10.1016/j.trac.2021.116409

Naake, T., & Gaquerel, E. (2017). MetCirc: Navigating mass spectral similarity in high-resolution MS/MS metabolomics data. *Bioinformatics*, *33*(15), 2419–2420. https://doi.org/10.1093/bioinformatics/btx159

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Röst, H. L., Sachsenberg, T., Aiche, S., Bielow, C., Weisser, H., Aicheler, F., Andreotti, S., Ehrlich, H.-C., Gutenbrunner, P., Kenar, E., Liang, X., Nahnsen, S., Nilse, L., Pfeuffer, J., Rosenberger, G., Rurik, M., Schmitt, U., Veit, J., Walzer, M., … Kohlbacher, O. (2016).

OpenMS: A flexible open-source software platform for mass spectrometry data analysis. *Nature Methods*, *13*(9), 741–748. https://doi.org/10.1038/nmeth.3959

Ruttkies, C., Schymanski, E. L., Wolf, S., Hollender, J., & Neumann, S. (2016). MetFrag relaunched: Incorporating strategies beyond in silico fragmentation. *Journal of Cheminformatics*, *8*(1). https://doi.org/10.1186/s13321-016-0115-9

Samanipour, S., O'Brien, J. W., Reid, M. J., & Thomas, K. V. (2019). Self adjusting algorithm for the nontargeted feature detection of high resolution mass spectrometry coupled with liquid chromatography profile data. *Analytical Chemistry*, *91*(16), 10800–10807. https://doi.org/10.1021/acs.analchem.9b02422

Schollée, J. E., Hollender, J., & McArdell, C. S. (2021). Characterization of advanced wastewater treatment with ozone and activated carbon using LC-HRMS based non-target screening with automated trend assignment. *Water Research*, *200*, 117209. https://doi.org/10.1016/j.watres.2021.117209

Schollée, J. E., Schymanski, E. L., Avak, S. E., Loos, M., & Hollender, J. (2015). Prioritizing unknown transformation products from biologically-treated wastewater using high-resolution mass spectrometry, multivariate statistics, and metabolic logic. *Analytical Chemistry*, *87*(24), 12121–12129. https://doi.org/10.1021/acs.analchem.5b02905

Schollée, J. E., Schymanski, E. L., Stravs, M. A., Gulde, R., Thomaidis, N. S., & Hollender, J. (2017). Similarity of high-resolution tandem mass spectrometry spectra of structurally related micropollutants and transformation products. *Journal of the American Society for Mass Spectrometry*, *28*(12), 2692–2704. https://doi.org/10.1007/s13361-017-1797-6

Schymanski, E. L., Bolton, E., Cheng, T., Thiessen, P., Zhang, J. (Jeff), & Helmus, R. (2021). *Transformations in PubChem - full dataset*. Zenodo. https://doi.org/10.5281/zenodo.5644561

Schymanski, E. L., Jeon, J., Gulde, R., Fenner, K., Ruff, M., Singer, H. P., & Hollender, J. (2014). Identifying small molecules via high resolution mass spectrometry: Communicating confidence. *Environmental Science & Technology*, *48*(4), 2097–2098. https://doi.org/10.1021/es5002105

Schymanski, E. L., Kondić, T., Neumann, S., Thiessen, P. A., Zhang, J., & Bolton, E. E. (2021). Empowering large chemical knowledge bases for exposomics: PubChemLite meets MetFrag. *Journal of Cheminformatics*, *13*(1). https://doi.org/10.1186/s13321-021-00489-0

Senan, O., Aguilar-Mogas, A., Navarro, M., Capellades, J., Noon, L., Burks, D., Yanes, O., Guimerà, R., & Sales-Pardo, M. (2019). CliqueMS: A computational tool for annotating in-source metabolite ions from LC-MS untargeted metabolomics data based on a coelution similarity network. *Bioinformatics*, *35*(20), 4089–4097. https://doi.org/10.1093/bioinformatics/btz207

Silva, R. R. da, Dorrestein, P. C., & Quinn, R. A. (2015). Illuminating the dark matter in metabolomics. *Proceedings of the National Academy of Sciences*, *112*(41), 12549–12550. https://doi.org/10.1073/pnas.1516878112

Stein, S. E., & Scott, D. R. (1994). Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry*, *5*(9), 859–866. https://doi.org/10.1016/1044-0305(94)87009-8

Treutler, H., Tsugawa, H., Porzel, A., Gorzolka, K., Tissier, A., Neumann, S., & Balcke, G. U. (2016). Discovering regulated metabolite families in untargeted metabolomics studies. *Analytical Chemistry*, *88*(16), 8082–8090. https://doi.org/10.1021/acs.analchem.6b01569

Vermeulen, R., Schymanski, E. L., Barabási, A.-L., & Miller, G. W. (2020). The exposome and health: Where chemistry meets biology. *Science*, *367*(6476), 392–396. https://doi.org/10.1126/science.aay3164

Watrous, J., Roach, P., Alexandrov, T., Heath, B. S., Yang, J. Y., Kersten, R. D., Voort, M. van der, Pogliano, K., Gross, H., Raaijmakers, J. M., Moore, B. S., Laskin, J., Bandeira, N., & Dorrestein, P. C. (2012). Mass spectral molecular networking of living microbial colonies. *Proceedings of the National Academy of Sciences*, *109*(26), E1743–E1752. https://doi.org/10.1073/pnas.1203689109

Wicker, J., Lorsbach, T., Gütlein, M., Schmid, E., Latino, D., Kramer, S., & Fenner, K. (2015). enviPath the environmental contaminant biotransformation pathway resource. *Nucleic Acids Research*, *44*(D1), D502–D508. https://doi.org/10.1093/nar/gkv1229