Environmental Sciences Europe

RESEARCH

Open Access

Development of a retrospective processing and data aggregation platform for non-target screening of surface waters

Kevin S. Jewell[1*], Nina Hermes[1], Franziska Prodöhl[1], Ole Lessmann[1], Georg Dierkes[1], Björn Ehlig[1], Tobias Schulze[2], Eric Winter[2], Ronya Mona Wallis[2], Anna Lena Kronsbein[2], Jan Koschorreck[2] and Arne Wick[1]

Abstract

Background Data accessibility and exchange are current key topics in the exploitation of the enormous potential of non-target screening (NTS) for surface water monitoring, especially when monitoring water bodies crossing jurisdictional boundaries. So far, major challenges stem from (1) the size of NTS raw measurement files, (2) the diverse array of instrumental and computational methods, and (3) a need for continuous reprocessing of measurement files as spectral libraries grow. Pre-processing workflows connected to databases and online dashboards are one crucial way to address these challenges and provide direct access to monitoring data.

Results The NTSPortal is a data assessment dashboard and repository for pre-processed and annotated NTS data for online access by users in the fields of analytical chemistry and environmental management. It provides both a graphical interface for search and visualization and a programming interface for integration into tailored data science workflows. We present the functionality using data from over 12,000 measurements from over 300 sampling sites within Germany analyzed with a uniform method over the course of 7 years. Using a collaborative spectral library, over 1000 compounds were identified via library screening, including pharmaceuticals, synthesis intermediates, and transformation products. We report performance figures and several use-cases demonstrating (a) searching for multiple substances and retrieving an overview of their detections in different matrices, (b) selecting a substance based on long-term intensity trends and visualizing time series, and (c) searching for substances with an isolated spatial occurrence pattern. The programming interface is demonstrated by implementing a time-series analysis script to search for discontinuous emission patterns.

Conclusions The highly automated pre-processing workflow for library screening allows much of the "heavy lifting" in NTS data processing to be off-loaded, enabling the environmental manager to focus on data analysis. The proposed system is open source and institutional access to the database can be made available upon request. The database has a flexible schema geared towards surface water monitoring but can be adapted to different needs. Further development of the system is underway, including the integration of unknown non-target features (without a library match), different instrumental analytical methods and concentration estimates.

Keywords Non-target analysis, High-resolution mass spectrometry, Databases, Open data, Open source, Water quality monitoring, Emerging contaminants, Micropollutants, Organic trace contaminants, Repository

*Correspondence:
Kevin S. Jewell
jewell@bafg.de
Full list of author information is available at the end of the article

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Background

Accessibility of non-target-screening (NTS) data for chemical water quality management poses challenges due to the large size of data, lack of standardized formats and the need for detailed metadata to perform statistical analyses. The large size of raw measurement files requires that pre-processing must be performed asynchronously from aggregated data analysis, i.e. the data pipeline must be decoupled. Inter-laboratory collaboration is therefore often limited to small data subsets, for instance in the form of non-standardized text files such as CSV. Metadata, such as sample preparation details or instrumental performance, becomes increasingly challenging to obtain and harmonize in this context. Moreover and in our view most critically, additions of new substances to spectral libraries means that raw measurement files must be continuously reprocessed (retrospective library screening) for annotations to remain up-to-date, a very repetitive task. For large datasets, these challenges can best be met by automated data processing and a repository to store measurement files, metadata, and pre-processed results.

With an automated retrospective library screening workflow in place, it is possible to re-scan an archive of thousands of NTS measurements for newly identified, "emerging" or suspected environmental contaminants simply by adding these to the spectral library–assuming the compounds are amenable to the analysis method. We can thereby obtain timely occurrence data without the need for costly and repetitive sampling and chemical analysis. The resulting datasets are harmonized and can be directly fed into interregional comparisons and time-series or other statistical analyses. Building NTS repositories for pre-processed data, providing interactive graphical user interfaces (dashboards) and machine-readable access (APIs) for data analysis can address the technical challenges of working with NTS measurement data, allowing better data utilization while also assisting laboratories to conform to FAIR data principles [1, 2].

For the environmental management of surface waters and for chemical risk assessment at the environment agency level, the need for aggregating NTS data stems from the cross-border nature of environmental impacts and the often long monitoring time frames needed. Firstly, the sources of chemicals and their environmental impacts may be in different jurisdictional areas, requiring data to be exchanged between agencies. This is particularly important in federated governments such as Germany and in national unions such as the European Union. Secondly, chemicals' legislation often needs many years to take full effect (e.g., because of gradual production changes), necessitating secure data and metadata archiving to prevent data loss and to enable a full (retrospective) assessment of the impacts. Thirdly, the aggregated NTS data can be used to provide comprehensive information on the interregional distribution and time trends of new emerging compounds, i.e. chemicals that just recently received increased political or scientific attention as environmental contaminants. This data has an enormous potential for a more evidence-based prioritization of compound monitoring lists and tailoring of monitoring programs.

To make NTS data searchable with a web-dashboard, it must be combined with metadata and indexed in a search engine. A graphical user interface is essential to broaden the scope of potential users. Several strategies and platforms have been implemented for NTS (pre-) processing workflows and databases. A recent study gives an overview of six current software projects with various online data accessibility (search/export) tools [3], among them the Digital Sample Freezing Platform (DSFP), maintained by the NORMAN Network, one of the earliest platforms for environmental NTS data from research projects and surveys [4]. The platform allows users to upload raw measurement files and perform pre-processing with an NT-feature detection and library screening workflow, then view results in the form of NT-feature lists, spectra, time series and map overlays. In addition to the aforementioned projects, several projects have been started to aggregate NTS data at the national, river-basin, or agency/sector level. A non-exhaustive list includes: NTSuisse, a database for river monitoring for the Swiss Cantons [5], the multinational Rhine Project Non-Target Screening for the Rhine basin [6] and K2I-Xplore, a database for drinking water suppliers headed by the German Association of Water and Gas Providers [7].

The Federal Institute of Hydrology (BfG) and the German Environment Agency (UBA) embarked in 2019 on a project [8] to build a database and search engine for pre-processed NTS data, to allow the aggregation of NTS data from federal and state environment agencies in Germany, supporting inter-agency collaboration and cross-border environmental management. At the start of the NTSPortal project, there was no open source community platform that could be integrated into the existing NTS infrastructure at the BfG and could be implemented with the pre-existing data processing workflow [9, 10]. Furthermore, to allow data exchange between government agencies, the service had to be hosted locally, within the government network, so an external service could not be used.

For the purposes of scientific publication, adhering to E-government guidelines and for systems integration, an open source platform leveraging pre-existing, community-maintained front-end frameworks was considered a requirement of the project. The promotion of FAIR data, which includes code transparency, is often required

for projects funded by the public sector. The NTS and metabolomics fields are rich with open source projects, on which the NTSPortal was built.

Regarding system interoperability, the project needed to integrate into existing laboratory information systems, spectral libraries and server infrastructure. The goal hereby being to share resources and build on previous investments. The necessary flexibility was gained through open source and a modular architecture. Furthermore, by using community-maintained, front-end libraries for services that are "open to the internet", the maintenance burden, especially regarding IT-security requirements, could be shared among a much larger user and developer base. The modular approach also allows the project to more easily adapt to continuing developments in the NTS field.

Several goals were outlined at the start of the project: (1) allow aggregation of pre-processed NTS data (lists of detections with metadata) and make this searchable online, (2) provide online (spatial, temporal) analysis tools and (3) provide API access for integrating the database into data analysis scripts and reporting software. Providing measurement file processing workflows directly to end-users was not planned since this would have required giving users full access to computing resources and would have increased the IT-security requirements, thereby exceeding the available budget.

A further goal of the project was to propose ideas for data aggregation and exchange pertaining to the development of interoperable software tools and standards, an important topic of discussion in the NTS field [11]. The German federal system entrusts the individual states with water quality monitoring, making harmonized data exchange formats and aggregation of data particularly important for water bodies crossing state boundaries.

The following article introduces NTSPortal from a technical perspective. We describe the architecture and current system requirements, the current size of the database, and the performance of the system, for example the time needed to process raw measurement files. Application of both the online GUI and the API are then described in the form of example use-cases in order to demonstrate how the system can be used in environmental management and to provide inspiration for further development. The use-cases are based on three main datasets: (1) daily monitoring of the middle Rhine River at the Koblenz station (2017–2025), (2) annual composite samples of suspended particulate matter (SPM) from river monitoring stations across Germany, provided by the German Environmental Specimen Bank (2005–2023), and (3) regional sampling of rivers and streams in the states of Saxony and Hesse (2021–2023), in cooperation with state environment agencies. These datasets were generated with the same instrumental analysis method and processed with a library screening method. The related study by Kronsbein et al. [12] describes a detailed application of NTSPortal datasets to assist environmental chemicals management and risk assessment specifically in the German and EU regulatory context.

Methods

Dataflow concept

Measurement files were processed using a previously published library screening algorithm in R [9], currently available in the R-ntsworkflow package v0.2.10 [10] and using the Collective Spectral Library (CSL) v26.1 [13] as the reference spectral library. This algorithm searches the measurement files for compounds found in the library and builds a list of detections which are stored in the Annotated Feature Database. See Sect. "Data Processing Method" for more details. Figure 1 shows the general concept for data import and export using NTSPortal.

Software architecture

The highest-level software components are the NTSPortal backend R-package (R-ntsportal v25.3) and the NTSPortal web dashboards (v25.3), which use the Kibana visualization platform (Elastic NV). R-ntsportal handles both measurement file processing and also

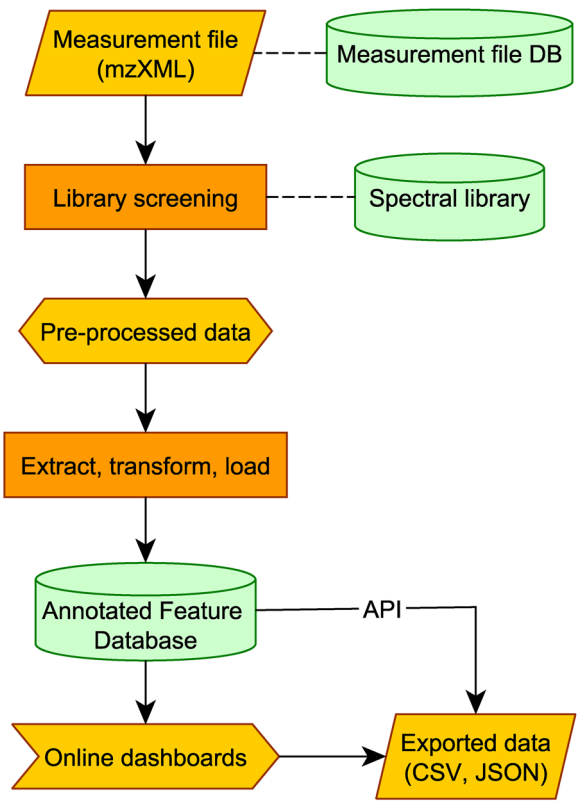


Fig. 1 Overall dataflow concept of the NTSPortal system

communication with the Elasticsearch API, for example to fetch data for analysis scripts. The Elasticsearch search engine (Elastic NV) is used for storing and searching the pre-processed NTS data. More information on software dependencies is available in Additional File 1, Sect. S1.

Hardware architecture

The hardware components used to run NTSPortal can be categorized under 1) measurement file processing, 2) data management and 3) visualization/export (Fig. 2). The core component of NTSPortal is a high-performance computing cluster (HPC) used for measurement file processing. This system consists of 7 compute nodes, each with 48 CPUs and 503 GB of memory. Measurement file processing is submitted to the HPC via a job scheduler (Slurm Workload Manager). The data management component consists of a separate server ("R server") and a 3-node cluster running Elasticsearch. The visualization component is covered by a server running the GUI platform (Kibana) and hosting the dashboards. All traffic to and from the Internet runs through a reverse proxy for IT security purposes. All servers are virtualized and run Rocky Linux 8.10. The Elasticsearch cluster, R and Kibana servers run in an HCI (hyper-converged infrastructure). All servers are part of a shared IT infrastructure and are administered by the IT department of the institute.

Database schema

NTSPortal uses Elasticsearch as a search and analytics engine, it stores data in documents and does not have a relational schema as found in conventional databases. There are several document types included in NTSPortal as listed in Table 1. In "feature" documents, each entry is a NT-feature, i.e. a detection representing a chromatographic peak. Each document includes fields such as m/z and retention time (t_R), library screening annotations such as compound name and InChI, sample information such as sampling time and location as well as method information such as instrument and chromatographic method. The full list of fields is available in the documentation [14]. Elasticsearch is designed to process large volumes of computer-generated, time-series data (e.g., log files) and was chosen for this project because NTS is seen as a stream of sensor data and in the future could become fully automated (online analysis).

Analytical methods

Sampling of water and SPM follows the methods described in [15] and [16], respectively. The analytical method for all measurements is described in [17]. A brief method description for both matrices follows. Water samples were filtered (0.45 µm regenerated cellulose, Whatman PLC) and 100 µL were injected without extraction. SPM samples (0.5 g dry weight) were extracted with methanol/water (Milli-Q) (1:1, v/v). Both were spiked with a mix of three deuterated internal standards (ISTD). From 2025 onwards, these were replaced by *Pharma-DX Mix 16* (Neochema GmbH, Bodenheim), containing 16 ISTD (see Additional File 1, Sect. S2 for details). Analysis was performed on an LC-ESI-QqToF system

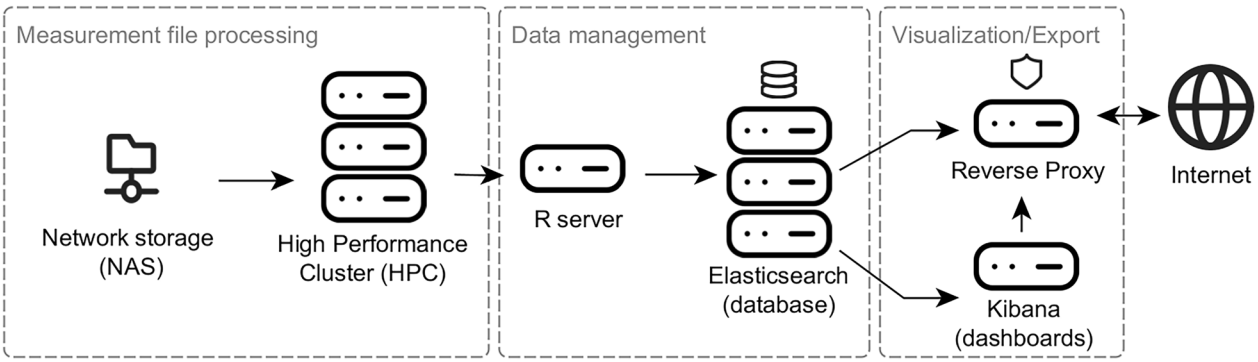


Fig. 2 Hardware components of NTSPortal. Arrows indicate general data flow. Adapted from the NTSPortal Documentation [14]

Table 1 List of document types used to store data in NTSPortal

Document type	Description
Feature	Results of measurement file processing (e.g. library-screening)
Msrawfiles	Registry of measurement files, metadata and processing parameters
Analysis	Results of summary statistics on specific subsets in the feature table

(Agilent 1260 Infinity, Sciex TripleTOF 6600) equipped with a reversed phase, C18 column (Agilent Zorbax Eclipse Plus, 2.1×150 mm, 3.5 µm). Chromatographic gradient elution (30 min run time) with water and acetonitrile (0.1vol% formic acid) was used. The main acquisition parameters were: MS^1 in full scan (m/z 100–1200, resolution > 25,000 at m/z 400) followed by MS^2 in data-dependent acquisition mode (m/z 30–1200, 8 scans, high sensitivity mode) set at 40 ± 15 V (average of 3 scans) and a cycle time of 0.5 s. An MS^2 fragmentation spectrum is made for an isolated m/z precursor ion and only triggered if a threshold intensity for the precursor in MS^1 is reached. Separate runs were used for ESI+ and ESI−. Data independent acquisition mode data cannot be processed with the current workflow and until now, only liquid chromatography has been tested.

Data processing method

The data processing method is written in R, C++ and Python and proceeds batch-wise, meaning one group of measurement files (usually representing a measurement batch) are processed in isolation. The library screening module is presented in this study and was previously reported in [9]. Measurement files in proprietary vendor formats were converted to the open format mzXML using MSConvert 3.0.10800 [18]. The mzXML format was used to maintain consistency within longer time-series.

Non-target feature annotation is carried out at two confidence levels, the first is based on comparing molecular ion m/z, MS^2 fragmentation spectra, and method-specific t_R with data in the spectral library (CSL). This is labelled an *MS2-RT* level match. The method is described in detail in [9] and parameters are provided in Table 2. Currently, there is no t_R correction performed. For low-intensity peaks with missing MS^2, a lower-confidence (*Batch-RT* level) match, based only on m/z and t_R, is made. It is still required that at least two *MS2-RT* level matches are made in a batch for the compound to be annotated at all. At the *Batch-RT* level, the t_R is taken as the average from all *MS2-RT* level matches in the current batch, not from the CSL. This is to reduce false positives by allowing a smaller t_R tolerance, see "reprocess compounds" in [9] for further details. All processing settings are stored in the msrawfiles table individually for each file and can be modified for different analytical methods.

Table 2 Processing parameters used for library screening of the datasets used in this study

Parameter	Tolerance
MS^1 m/z match	< 5 mDa
MS^2 m/z match	< 15 mDa
MS^2 dot prod. score[a]	> 300 (of 1000)
t_R match (*MS2-RT* level)	< 90 s
t_R match (*Batch-RT* level)	< 20 s
Peak area[b]	> 1

[a] Using the "Purity" algorithm described in [20], also known as "cosine similarity"

[b] Peak areas are in arbitrary units. Due to the Sciex centroiding algorithm implemented in MSConvert, peak areas are in the range of $1–10^5$

The results of library screening (initially stored as tables in R) are then reformatted and imported into Elasticsearch. This ETL (extract, transform, load) step is accomplished using R-ntsportal [19]. In short, each feature is extracted as a separate entity and annotations and metadata are added sequentially.

ISTDs are processed as normal features but additionally, the peak intensity and area of a pre-selected ISTD is added to each feature during ETL. This is used for computing relative instrumental response values to account for analytical performance changes.

Quality assurance

At least two field blanks are measured in each batch to remove background features (unless a factor 5 intensity difference is exceeded). A partial intensity correction for features with low-intensity blank values is not performed since it was found to lead to additional noise in time-series without significantly changing the results.

In some datasets, replicate measurements were performed (Sect. "Datasets Used in this Study"). The results of all measurements are included in the data in NTSPortal, i.e. there is no replicate summarizing during preprocessing. This allows more flexibility in subsequent data analysis workflows. To reduce false positives, at least two *Batch-RT* level matches of each compound must be present in each replicate group. If no replicates are available, there is still the requirement of at least two *MS2-RT* level matches in the batch as a whole, as described in the previous section.

Quality control

Currently, the system uses two levels of custom built, interactive control charts to monitor system performance and verify adherence to tolerances. The first level is an automated data processing and visualization system installed on the instrument computer itself, covering short-term (batch-to-batch) QC. For long term, aggregated analysis, the Annotated Feature Database of NTSPortal includes ISTD and reference data for QC, and the GUI includes several diagnostics dashboards to view the data.

Short term quality control at the instrument

Quality parameters (intensity, area, MS^1 and MS^2 m/z error, t_R, chrom. peak width and spectral peak resolution)

are monitored directly on the MS instrument control computer using two R-based QC software tools (1) R-shewhart4cds, for visualizing data from automated direct infusion experiments every 2 h and (2) R-shewhart4hrms, for processing and viewing LCMS measurements of the ISTD mix in purified water at the start and end of every batch [21–23]. When tolerance levels (see Additional File 1, Sect. S3) are exceeded, the analyst can react quickly, stop further measurement, perform necessary fixes such as instrument maintenance, and reacquire data.

Long term quality control using the Annotated Feature Database

The NTSPortal includes internal standards as regular features (labelled as surrogate standard). The ISTD Diagnostics dashboard uses this data to show time series and distributions of peak intensity and area, m/z and t_R for the user-selected data subset, as well as warnings when an ISTD is not detected. In addition, one reference sample using a standard mix of 137 compounds [24] at 1 µg/L is also measured once per batch and undergoes the standard processing workflow. The Reference Diagnostics dashboard shows the data for these measurements (the same parameters as shown for ISTD). This dashboard additionally has time series for the number of detected compounds. Both diagnostics dashboards are used to answer a QC checklist after every new upload (See Additional File 1, Sect. S4). If tolerance levels are exceeded, the data is removed.

Despite these quality checks, the data in NTSPortal is nevertheless computer-generated, it is critical that users are aware of uncertainties and verify the results of their research. A disclaimer with this information is prominently displayed on the landing page. The repository [19] also provides an issue board where results can be discussed.

Datasets used in this study

The datasets used in this study are an aggregation of ongoing monitoring programs, discrete sampling campaigns and research projects in collaboration with federal and state environmental agencies in Germany. Although these measurements were performed at the BfG, the NTSPortal was designed to aggregate data from different laboratories and trial studies have been completed to include data analyzed at other agencies [8].

A full overview of all data in NTSPortal is given in Sects. "Performance" and "Compound Search Dashboard". The three main datasets used for demonstrating the GUI and API are listed in Table 3. In the case of annual SPM sampling, these samples were prepared and measured in triplicate.

Results and discussion

Performance

The full dataset in NTSPortal contained, at the time of writing, more than 14,000 LC-HRMS/MS measurement files (including ~ 11% blank and QC samples), split into 456 batches (Fig. 3b). The spectral library (CSL 26.1) contained 1821 compounds (ca. 37,000 spectra for different instrumental parameters and for some compounds, multiple adducts) [13]. The library is a collaborative effort of federal and state environment agencies in Germany. The compounds were chosen based on research in the field and results from projects and monitoring programs performed at the contributing environmental institutes. It therefore has both a multi-regional and surface-water-monitoring focus and has been continually growing since 2017. Figure 3a shows the distribution of the main compound classes in the library. A compound may belong to multiple groups and around 21% of compounds are not yet categorized. This empirical classification scheme has evolved organically with the CSL. It focuses on the main usage categories of compounds based on literature reports and online databases (e.g., PubChem [25]). It is currently being revised to include new schemes based on market usage and regulatory classifications [12]. Currently, the largest group is human and veterinary medicinal products, followed by industrial chemicals and plant protection products (when grouped together). Human metabolites are grouped together with environmental transformation products. The CSL, rather than larger online libraries, was used to ensure more uniform quality, allow faster (in-house) corrections and additions as well as provide method-specific t_R to increase the annotation confidence [9].

Table 3 Overview of the main datasets implemented in the use-cases

Dataset or project	Time frame	Matrix[a]	Composite sampling	Implemented in use-case (section)
Monitoring station Koblenz/Rhine, daily samples	2018–2025	water	1 day	Example Use of the API
Collaboration Environmental Specimen Bank	2005–2023	SPM	1 year	Time Series Dashboards
Collaboration with state environment agencies of Saxony and Hesse	2021–2023	water	grab samples	Spatial Occurrence Analysis Dashboard

[a] For sample and method details see [15] for water and [16] for SPM

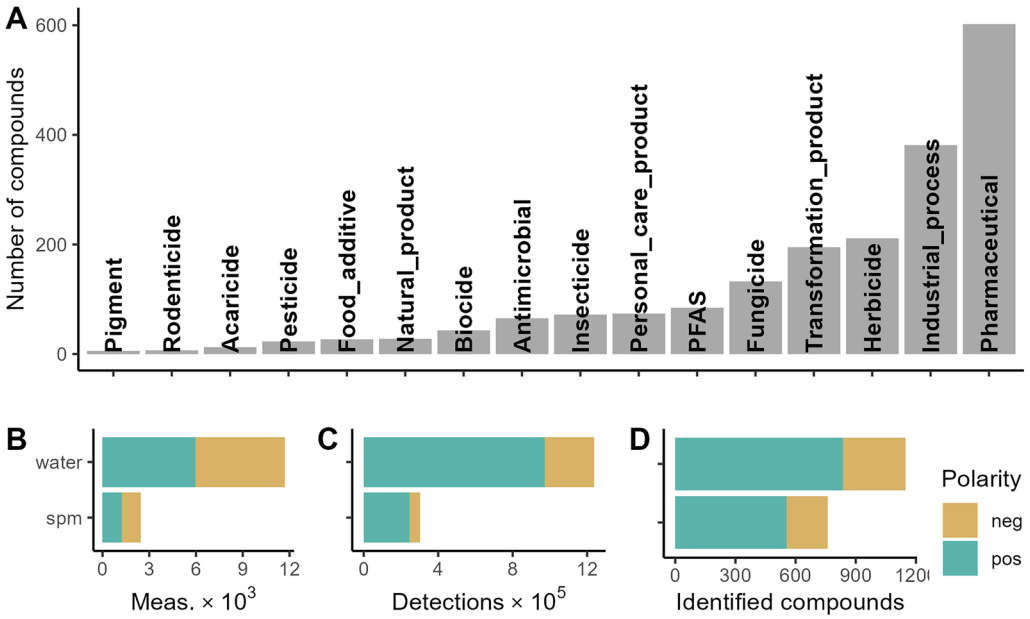


Fig. 3 Overview of data in the CSL and NTSPortal. **A** Distribution of compound classes in the spectral library (CSL), where "transformation product" also includes metabolites. **B** Number of raw LC-HRMS measurement files in NTSPortal. **C** Total number of detections (annotated NT-features) in environmental samples. **D** Total number of different compounds detected in environmental samples (identified by library screening)

The processing of the measurement files with the library screening algorithm on the HPC took approximately 14 h. The parallelization is batch-wise, with a maximum of 10 parallel batch processes, meaning on average one file is processed per minute per thread but this can vary depending on the size of the file, the size of the batch and the current load from other users on the cluster. The number of parallel threads can be increased (up to 336) if more processing power is needed. The resulting, pre-processed data contained around 1.4×10^6 annotated features, of which 42% included an MS^2 spectrum. Importing this data into Elasticsearch took approximately 0.5 h. Once imported, searching and viewing the data in the dashboards is interactive, i.e. a delay of under one minute is expected, even when searching the data for MS^2 fragment masses (in total ca. 2×10^7).

High latency can occur when retrieving large amounts of data in one request (e.g., time series of daily measurements over 6 years for 100 compounds). The dashboards automatically convert requests to summary statistics or truncate the results if the requested data is too large. For example, in the case of the peak area-time series dashboards, daily values may be converted to weekly median values for the top 100 most frequently detected compounds. In the case of replicate injections, the median peak areas per replicate group are shown on the time series dashboards. The API offers more flexibility, but requests may take longer to complete.

Use-cases

The following use-cases demonstrate how different dashboards of the online GUI can be utilized to perform basic search and prioritization tasks. The last example demonstrates how the database can be incorporated into a more complex statistical analysis workflow using the API to fetch data for processing in R. This article was itself built using the NTSPortal API and the *Quarto* publishing tool (Posit PBC) and as such is also a demonstrator for using the NTSPortal API for generating reports. The code for the manuscript is in Additional File 2. The figures shown below are made using the API and third-party packages in R. Screenshots of the GUI for each example and an overview table of all dashboards are available in Additional File 1, Sect. S5.

Compound Search dashboard

Currently, the NTSPortal contains data for substances identified during processing by means of library screening using the CSL for two sample preparation methods and one analytical method. Data from 351 sites for surface water and 16 sites for SPM were analyzed and processed.

For water, data from 10,286 measurements are included, containing detections of 926 identified compounds (Fig. 3b-d). For SPM, 2398 measurements are included (either yearly or monthly composite samples, measured in triplicate). The number of substances

detected in SPM (591) was lower in comparison to water, which could be due to several reasons, including a different sampling frequency, the different analytical window due to the SPM extraction method, a lower diversity of sampling locations, or the current distribution of compounds in the CSL in terms of their expected partitioning to SPM (only 140 compounds have, e. g., an estimated log K_{ow} above 5 [26, 27]).

An overview of all identified substances is available in the Compound Search dashboard. A table containing substance-specific parameters (e.g., compound name, classification, m/z) as well as detection statistics is provided (number of detections, number of sites at which the compound was detected, etc.).

The table can be sorted or filtered according to the parameter of interest (Table 4). For example, one can view only detections in water samples and then sort these according to the number of sites at which a substance was detected, giving a first overview of ubiquitously detected substances or those with a more isolated occurrence pattern (see Sect. "Spatial Occurrence Analysis Dashboard" for more spatial analyses). Each table entry includes links to further dashboards to look at a specific compound in more detail, including MS^2 spectra and CSL entries.

Time series dashboards

The NTSPortal dataset currently contains continuously updated time series from 13 sampling stations for the SPM matrix and from one monitoring station for the water matrix (Koblenz/Rhine). The time series dashboards (one for each matrix) allow the user to view peak area–time series of searched compounds. The interface shows two time series using either absolute peak areas or peak areas relative to an ISTD.

These dashboards can be used, for example, (a) for assessing the effectiveness of contaminant reduction measures (e.g., new regulations, investments into wastewater treatment) by observing decreasing trends after reduction measures are put in place or (b) to assess increasing risk potential from contaminants with increasing trends.

On the Time Series SPM dashboard, to assist in the discovery of compounds with increasing intensity trends, a table lists the pre-computed slope of the linear regression of peak area over time for each compound at each station. To compute this slope, the peak area is normalized by (1) the pre-selected ISTD peak area and (2) the maximum peak area in the time series, as shown in the simple linear model 1, where A_{2N} is the double-normalized area. The normalization accounts for shifts in instrumental response over time and differing absolute response between compounds. The slope of the regression (β), which expresses the average relative change in response per year, is displayed in the table.

$$A_{2N} = \alpha + \beta t$$

where

$$A_{2N} = \frac{A/A_{ISTD}}{(A/A_{ISTD})_{max}} \quad (1)$$

The user can then sort the table by this relative change and view the time series. This simple prioritization tool can assist users to focus their search, but more detailed statistical (trend) analyses can be realized using external scripts and the API (see Sect. "Example Use of the API").

Example use of the Time Series SPM dashboard One of the top-ranked compounds in the list of increasing response trends in SPM was rivaroxaban (RIV), a blood thinning agent. Selecting this compound showed a map of the detections and the time series of the peak area of RIV split into multiple traces for each station and ESI molecular ion specification (combination of compound name, polarity, adduct or in-source fragment, and isotopologue). In total, five different adducts or in-source fragments for RIV were annotated: $[M+H]^+$, $[M-H]^-$, $[M+HCOO-]^-$, $[M+C2H7N+H]^-$ and the in-source fragment with a nominal mass of 390 (referred to in

Table 4 Summarized table of entries in the compound search dashboard. Example for five selected substances

Comp. name	Adduct	Isotopologue	*m/z*	Classification	Det. SPM	Det. water	Stations
Diclofenac	[M – H]–	monoisotopic	294.0082	Pharmaceutical	0	3040	297
Diclofenac	[M + H]+	monoisotopic	296.0230	Pharmaceutical	3	4172	326
Diphenylphosphine oxide	[M + H]+	monoisotopic	203.0612	Industrial_process	0	2366	18
Fluoxetine	[M + H]+	monoisotopic	310.1408	Pharmaceutical	832	13	11
Sitagliptin	[M + H]+	monoisotopic	408.1244	Pharmaceutical	1088	4883	320
Sitagliptin	[M – H]–	monoisotopic	406.1112	Pharmaceutical	0	4	1
Sitagliptin	[M + H2CO2 – H]–	monoisotopic	452.1148	Pharmaceutical	1023	3221	195
Tetrabutylammonium	[M]+	monoisotopic	242.2834	Industrial_process	1198	4808	333

the CSL as $[QF390]^-$), while isotopologue was always "monoisotopic". $[M+H]^+$ had the most detections, so this was selected. From the time series it is possible to select a time range and view the chromatograms, spectra, and data quality statistics of the selected features on the Spectra and Chromatograms dashboard, thereby allowing an estimate of the detection confidence. In the case of RIV ($[M+H]^+$), 750 annotations were made in total, 262 at the *MS2-RT* level. The average MS^2 similarity score was optimal at 900 (/1000). The m/z distribution had a standard deviation of 1.5 mDa and t_R was within a 0.2 min window. This information gave a high confidence in the annotation.

The time series showed that four (out of 16) stations had continuous detections of RIV and increasing peak areas. These stations were on the Mulde, Saale and in the lower Elbe. In contrast, RIV was not detected at the stations Zehren and Prossen on the upper Elbe. The sharpest increase and highest peak areas were found at the Mulde-Dessau station (Fig. 4). Consistent with these findings, it was found that RIV is produced by a pharmaceutical manufacturer with sites in the Mulde catchment [28] but further investigation would be needed to confirm any connection.

Based only on the information from the dashboard it can already be hypothesized that a localized source dominates the emissions (detections in SPM mostly in the Mulde and in the Elbe below the confluence of the Mulde) (Fig. 4). This data mining example shows the strengths of long-term, multi-catchment sampling, as provided by the Environmental Specimen Bank, combined with the accessibility of pre-processed NTS data in the NTSPortal. The information gained can be used as a starting point for more detailed investigations into the compound (e.g., concentration estimates, analysis of water samples) or into possible impurities or transformation products [29, 30].

Furthermore, the identification of RIV was possible because high-quality reference spectra were submitted to the CSL by the Bavarian Environment Agency in 2024 [13]. By reprocessing the NTSPortal's archive of measurement files against the updated CSL, we were able to generate the time trend data shown here. This highlights the critical role of inter-agency spectral libraries in

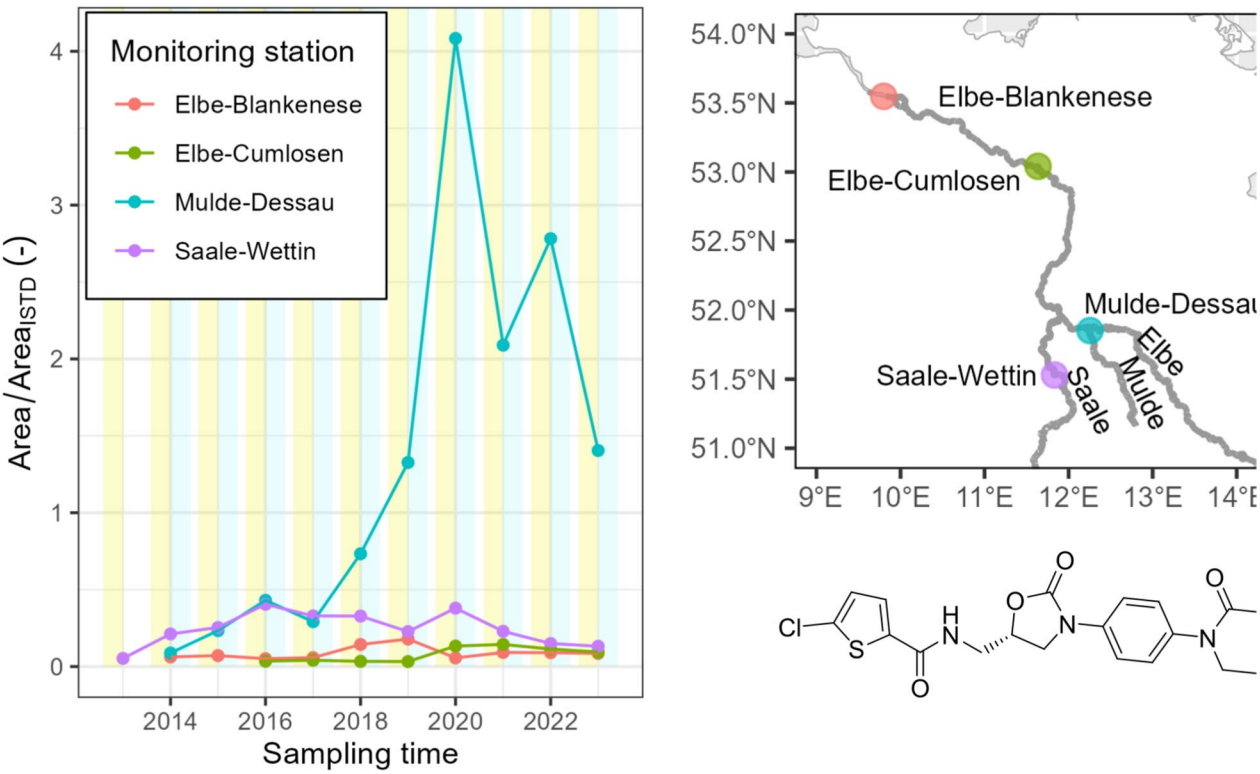


Fig. 4 Left: Peak area-time series of rivaroxaban (RIV) in annual, composite SPM samples. Peak area is relative to the internal standard (ISTD) peak area, bezafibrate-*d*4 at 1 µg/L. Data is reproduced as it is displayed on the Time Series SPM dashboard. Yellow bars indicate a measurement file is available for at least one station for that year, blue bars indicate at least one *MS2-RT* level annotation for that year. The points indicate a detection, showing the median peak area of triplicate measurements. Lines are the linear interpolation between points. Right-top: Map of the Elbe catchment in Germany showing the locations where RIV was detected (map does not show non-detects). Right-bottom: Structural formula of RIV

maximizing the value of NTS data archives through automated retrospective library screening.

Spatial occurrence analysis dashboard

Compounds with a regionally confined (localized) occurrence pattern but a high detection frequency may indicate a unitary source, e.g., due to industrial production or specialized uses, whereas wide-spread occurrence is an indication of diffuse sources or common consumption. With the nation-wide data available in NTSPortal, the Rarity dashboard was developed to enable the sorting of detected compounds for such spatial patterns utilizing an approach similar to the rarity score proposed by Krauss et al. [31]. A scatterplot sorts the detected compounds by their rarity/ubiquity (Fig. 5a). The user can interactively select a compound, for example, the quaternary ammonium compound tetrapropylammonium, and the adjoining map shows its occurrence pattern (Fig. 5b), indicating that it has multiple sources in both the Rhine and Elbe catchments. However, as evidenced by the large number of detection locations in tributaries of the Elbe (Mulde, Pleiße) in Saxony, there appear to be a few significant sources in that region.

If the user is interested in localized emissions in a specific region, the inverse search approach is also implemented on the same dashboard. The user draws a bounding box on the map and compounds that may have an unusually high detection frequency in that region (compared to the full dataset) are listed in a table. This approach utilizes the "significant terms" data aggregation algorithm provided by Elasticsearch [32]. In an example use-case, the area of the Hessian Ried in southern Hesse was selected and the top ranked compound with an unusually high frequency in this area was diprogulic acid (dikegulac), which is a waste product of past vitamin C production in the area. The compound can still be detected at elevated concentrations at some local groundwater wells and current detections in surface water can be explained by (a) infiltration of contaminated groundwater and (b) prior to the commissioning of an advanced treatment step in 2024, also by the emissions of an industrial wastewater treatment plant due to (non-potable) groundwater reuse by the connected industrial facility [33]. This basic use of the "significant terms" algorithm has limited parameterization options and results must be verified. Compounds may be prioritized even though they are not unusual to a region due to statistical bias. Using the Elasticsearch API in a data analysis script, fine-tuning the algorithm's parameters is possible, for example, to define the background dataset for comparison.

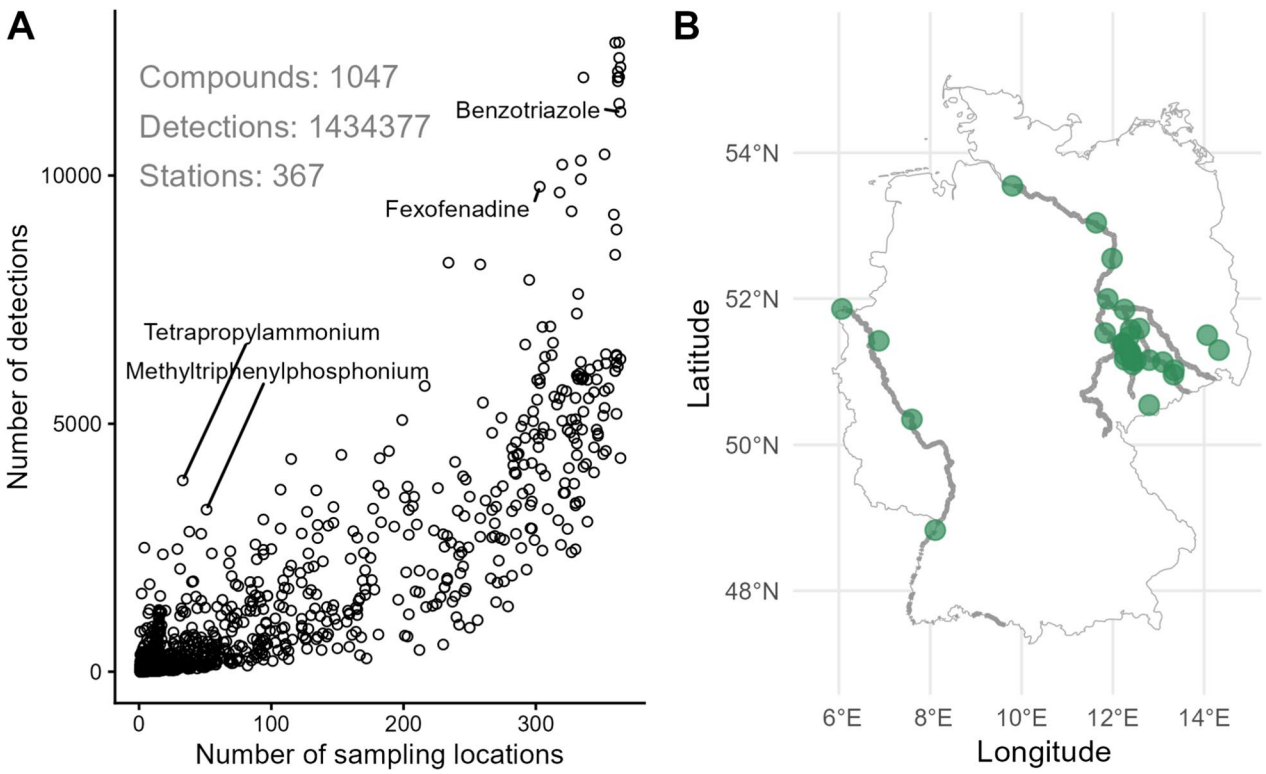


Fig. 5 **A** Rarity scatterplot to sort compounds into ubiquitous (upper-right) and localized (high number of detections at a limited number of locations) (upper-left). **B** Map of Germany showing the detection locations of tetrapropylammonium

Example use of the API: fetching and analyzing time-series data

The previous examples demonstrate tools for visualizing and exploring the data using online dashboards. However, to keep the interface intuitive and to work within the GUI framework used, there are limitations to the flexibility of the GUI elements. More detailed exploration of the datasets is intended to be performed by fetching appropriate data via the API and developing data analysis scripts.

One example use-case is the identification of discontinuously discharged organic compounds, which are often associated with industrial emissions [15]. Using the API through wrapper functions in R-ntsportal [19], an external R-script was prepared to analyze time-series data using a set of four parameters to search for discontinuous emissions:

A) *Detection proportion* proportion of measurements in which the compound is detected.
B) *Intensity spread* span of intensities, calculated by the ratio of the 95%-quantile to the 5%-quantile [34]. Intensities rather than normalized peak areas were chosen in accordance with the published method.
C) *Break period* discontinuous emissions are characterized by a short peak of detections followed by a time period with low intensities or no detection. This pause of emission is the break period. For each timepoint t it is calculated if intensity falls under a cut-off, i.e., if the intensity at t subtracted by the overall minimum intensity is less than 1/30 of the total intensity range [35].
D) *Anomaly points* in the case of unusual events, data points in time series may deviate extremely from the majority of detections. These outliers or anomalies can be of particular interest. For the detection of anomalies, a function from the R-timetk package was used [36]. The function removes the trend and seasonality components from the time series and for the remainder, time points are marked as an anomaly when the value is outside 1.5 times the interquartile range (IQR).

The script was applied to the time-series data from the Koblenz/Rhine station, for which daily, 24 h composite samples were analyzed from January 2018 to May 2025. The script source is in Additional File 3.

For the detection proportion, Chonova et al. [35] set 70% as a threshold. In this study a larger time period was analyzed and it was expected that discontinuous emissions may show a higher number of non-detects. Thus, the threshold was lowered to 25%. This threshold reduced the number of candidates from 492 to 113 compounds. Only 9 of these did not show a break period and these were also filtered out by the intensity spread filter when set to a threshold of 10 [35], making the break period filter unnecessary in this particular dataset. Using both filter stages (detection proportion and intensity spread), 62 candidates for discontinuous emission were left. The visual inspection of these time series confirmed some substances being relevant candidates, showing indeed well-defined peaks in the time series (e.g., Metalaxyl metabolite NOA409045 and Chlortoluron) but still a high number of false positives (continuous emissions without any substantial peaks). Therefore, the fourth criterion, anomaly points (at least one), was added. The intensity spread threshold was tuned to further refine the results list. In this demonstrative example, the threshold was increased to 25. The final settings for the parameters are given in Table 5. All thresholds had to be fulfilled for a time series to be prioritized.

The script produces a scatterplot of the computed parameters for all compounds (Fig. 6) and a table of the prioritized compounds (Table 6). The script output (Additional File 4) also includes time series for the 11 prioritized compounds, annotated with the detected anomaly points. For 10 of the 11 compounds, the visual inspection of the time series showed a clearly discontinuous time series and many of these compounds are known to be produced in the Rhine catchment in the industrial regions south and west of Frankfurt [15, 37]. Benzyltriethylammonium, for example, was detected with several time-series peaks in the years 2020 and 2021 (Fig. 6) [38, 39], (methoxymethyl)triphenylphosphonium had time-series peaks in 2017 and 2018 [15]. The time series of the natural compound 4-Methoxycinnamaldehyde, however, did not show a clear discontinuous pattern and must be treated as a false positive.

Although adenosine has reported therapeutic uses, it was surprising to see it linked to a pronounced discontinuous time series (see Additional File 4). The MS^2 spectra for this feature (available through the Spectra and Chromatograms dashboard) indicate a poor match compared to the spectral library. Thus, the annotation itself may be

Table 5 Metrics and thresholds used in the prioritization analysis. Intensity spread in arbitrary units. The full algorithm can be viewed in Additional File 3

Metric	Threshold
Detection proportion (%)	> 25%
Intensity spread (–) [34]	> 25
Break period (Boolean) [35]	true
Anomaly points [36]	> 0

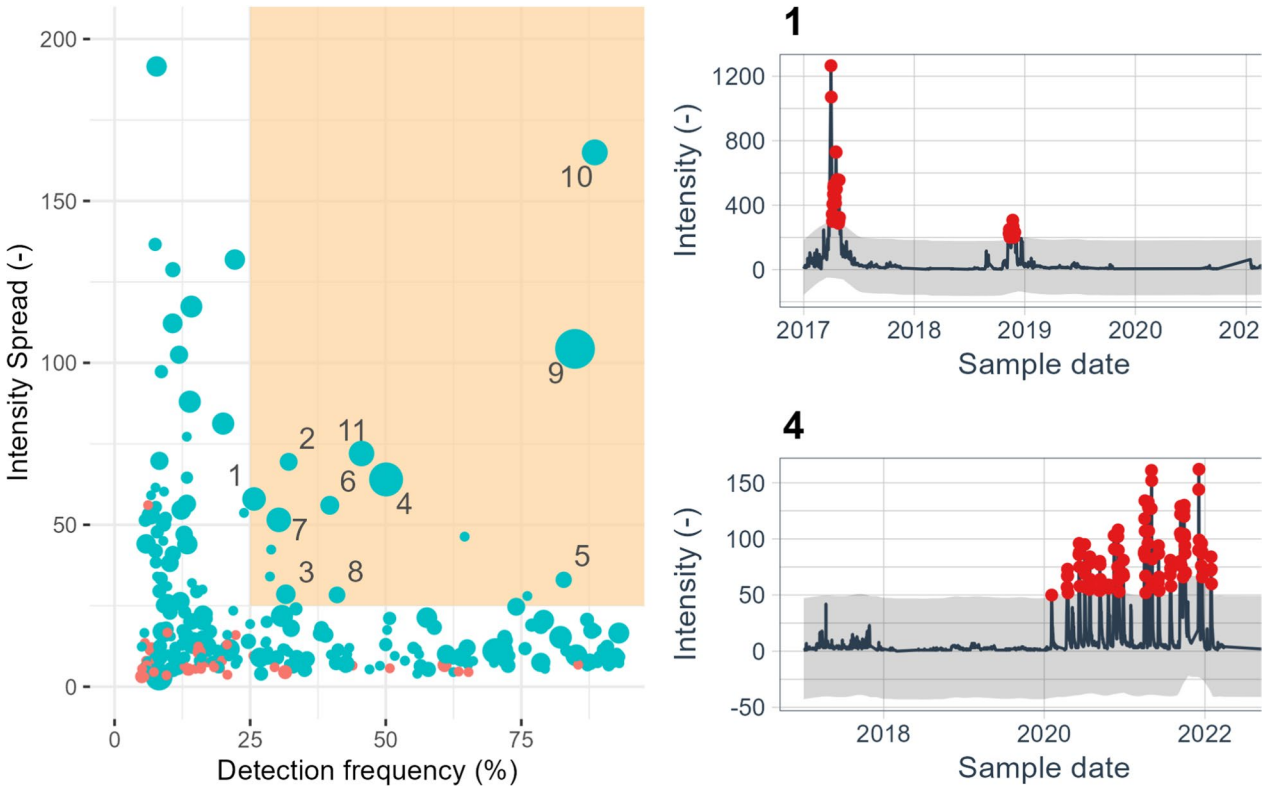


Fig. 6 Left: Scatterplot of all considered compounds (492) showing the four computed prioritization metrics. Compounds with a break period are teal (otherwise red). Point size corresponds to the number of anomalies detected (smallest point corresponds to no anomalies). Prioritized compounds (for which all thresholds were fulfilled) are (1) teal, (2) within the shaded area, and (3) have at least one anomaly. These compounds are numbered in alphabetical order (Table 6) and labelled on the plot (1–11). Right: Time series of two prioritized compounds (1: (Methoxymethyl) triphenylphosphonium, 4: Benzyltriethylammonium) with detected anomalies in red

Table 6 Details of compounds selected by the external prioritization script

Number	Name	Detection Frequency (%)	Intensity spread (–)	Anomaly points
1	(Methoxymethyl)triphenylphosphonium	26	58	32
2	4-Methoxycinnamaldehyde	32	69	11
3	Adenosine	32	29	16
4	Benzyltriethylammonium	50	64	100
5	Edetol	83	33	7
6	Ethyltriphenylphosphonium	40	56	14
7	Lumichrome	30	52	37
8	Repaglinide	41	28	8
9	Tetrabutylammonium	85	104	154
10	Tetrapropylammonium	89	165	44
11	Tributylmethylammonium	46	72	42

a false positive and therefore needs to be verified before further assessments are made.

While for some compounds (e.g. triphenylphosphoniums) more detailed data on concentrations and ecotoxicological relevance are already available [40], further investigation is needed for the detected quaternary ammonium compounds, to assess the environmental impact of these concentration peaks found in the Rhine. Data on concentrations in SPM is available [41], while toxicity data for these relatively short chain-length

QACs is limited [42, 43]. For the other compounds, e.g., repaglinide, there is limited data for (aquatic) ecotoxicity [44]. While edetol, a chelating agent similar to EDTA, has a relatively high reported PNEC of 15 mg/L [45]. Lumichrome is a bacterial metabolite of riboflavin so the detected time-series peaks may have a natural origin [46].

In this study, parameter thresholds were set to reduce the length of the results list, in order produce a manageable number of prioritized compounds. However, the approach needs to be further optimized and systematically validated based on both false negative and false positive rates. This was out of the scope of the current study as here the main focus was to demonstrate the general value of the NTSPortal API when performing bespoke analyses of the data using a statistical programming language and open source libraries.

Conclusions and outlook

In the context of water quality monitoring, we see that the data "pre-processing" workflow for this form of annotated-NTS (NT-feature detection and library screening) can be defined as a fixed set of steps and automated to run on a server cluster. The product of this processing, the Annotated Feature Database can be used as the basis for a wide variety of different spatial and temporal analyses, either through a graphical web-application or by developing more advanced data analysis scripts. This suggests that (a) the logistics of managing many terabytes of measurement data can be off-loaded to servers and automated workflows, freeing time for developing more detailed statistical analyses and (b) the large data-requirements of statistical analyses, at the very least for annotated-NTS, can be met through aggregating pre-processed data in a NoSQL database.

The preceding use-cases demonstrate a wide variety of applications of the NTSPortal to support surface water monitoring. The design demonstrates how long-term and interregional analysis of annotated NTS data can be implemented, for example, for (1) characterizing occurrence patterns and emission sources, (2) assessing the impacts of new chemical regulation or investments in wastewater treatment, and (3) giving advanced warning of new or increasing potential chemical hazards.

The platform provides substantially more efficient and integrative data analysis tools than were previously available and can build a basis for data exchange of NTS monitoring data in both regulatory and research contexts. However, the development of NTSPortal is ongoing and several additions, extensions and improvements are under way or planned:

- Development of data import tools, such as a user interface for uploading data, will lower the technical hurdles for external laboratories to submit data, encouraging data exchange. It has previously been shown how data from other labs, using different chromatographic methods, can be included into both library screening and non-target screening workflows in NTSPortal [8]. This is accomplished through the msrawfiles table, in which processing parameters can be set individually and the CSL, which contains spectra from a wide range of instrumental parameters and modeled and experimental t_R for multiple liquid chromatographic methods.
- In addition to t_R modeling in the CSL, t_R correction to account for instrumental drifts within one method is also foreseen to reduce the t_R tolerance for the *MS2-RT* level match.
- Inclusion of multiple ISTDs for normalization of intensities/peak areas and enabling the user to choose the number of used ISTDs interactively is foreseen.
- Further development of external analysis scripts for the NTSPortal is planned, including advanced spatial analyses to compare different regions and catchments.
- The processing workflow shown in this study currently only considers library screening but the inclusion of unknown features is currently ongoing and a test dataset with over 20 million unknown features has already been imported into NTSPortal. The demonstrator project [8] produced a pilot system for NT processing and the alignment (m/z and t_R grouping) of these features using a clustering algorithm. Follow-on studies should also consider expanding the annotation of features at different confidence levels using larger online libraries, t_R modeling and emerging AI techniques [47].
- Concentration estimates are vital for assessing the environmental relevance of detections of contaminants. They also provide a standardized measure for comparing results between laboratories. Quantification estimation in NTS is currently a subject of research in the field [48, 49] and an implementation for NTSPortal is in development.

Abbreviations

API	Application programming interface
BfG	Federal Institute of Hydrology
CPU	Central processing unit
CSL	Collective Spectral Library
CSV	Comma separated values
ESI	Electrospray ionization
ETL	Extract, transform, load
GUI	Graphical user interface
HPC	High-performance cluster
HRMS	High-resolution mass spectrometry
InChI	International Chemical Identifier

IQR	Inter-quartile range
ISTD	Internal standard
IT	Information technology
JSON	Javascript object notation (text-based data exchange format)
MS^2	Tandem mass spectrometry (fragmentation)
m/z	Mass-to-charge ratio
noSQL	Not only SQL
NTS	Non-target screening/non-target analysis
NT-feature	Non-target feature, a record representing a chromatographic peak
PNEC	Predicted no-effect concentration
QAC	Quaternary ammonium compounds
QqToF	Quadrupole time-of-flight mass spectrometer (with collision cell)
R	Programming language R
RIV	Rivaroxaban
SPM	Suspended particulate matter
SQL	Structured Query Language
t_R	Retention time
UBA	German Federal Environment Agency

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12302-026-01477-7.

Additional file 1. Supporting Information including screenshots and additional figures, in PDF format.

Additional file 2. Manuscript code showing use of API for reporting. QMD file.

Additional file 3. Example source code for the API and external script use-case. QMD file.

Additional file 4. Output of the data analysis script in Additional File 3, in PDF format.

Acknowledgements
The authors acknowledge the pivotal contributions of Jonas Skottnik, Lukas Müller, Arnd Weber and Thomas Scharrenbach of the BfG IT department for building and maintaining the computational, database and web-hosting infrastructure. The cooperation with the state environment agencies of Hesse and Saxony, including the provision of water samples, is gratefully acknowledged. The authors also thank colleagues Susanne Brüggen, Alexander Bammert (LANUK), André Macherius, Uwe Kunkel (LfU Bayern), Jakob Hauns and Marco Scheurer (LUBW) for contributing reference spectra as well as Martin Krauss (UFZ), Saer Samanipour (Univ. Amsterdam) and Johannes Boog (EAWAG) for constructive discussions. The authors also thank the anonymous reviewers for their many constructive suggestions.

Author contributions
KSJ, NH: conception, creation of new software, analysis and interpretation of data, design of work, drafted work; FP, BE, OL: data acquisition and analysis, creation of new software; GD, RW, ALK, TS, EW: Analysis and interpretation of data, substantive revision; AW, JK: conception, design of work, analysis and interpretation of data, substantive revision.

Funding
Financial support was provided by the Federal Environment Agency (REFOPLAN) projects 3720222010 and 3723222020 as well as the Federal Institute of Hydrology.

Data availability
The datasets supporting the conclusions of this article are available on the Zenodo repository [50]. The CSL is available on Zenodo [13] and the BfG spectra are also available on MassBank Europe. All code is available in the GitHub repositories [10, 19, 21, 23]. Access to the NTSPortal website [51] and API [52] can be provided to academic, government and non-profit institutions upon request to ntsportal@bafg.de. Access to specific datasets may be restricted depending on the open data license used by the original data provider.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Federal Institute of Hydrology, Am Mainzer Tor 1, 56068 Koblenz, Germany. [2]German Federal Environment Agency (Umweltbundesamt), Wörlitzer Platz 1, 06813 Dessau-Roßlau, Germany.

Received: 24 March 2026 Accepted: 12 July 2026
Published online: 31 July 2026

References

1. Hollender J, Schymanski EL, Ahrens L, Alygizakis N, Béen F, Bijlsma L, Brunner AM, Celma A, Fildier A, Fu Q et al (2023) NORMAN guidance on suspect and non-target screening in environmental monitoring. Environ Sci Eur 35(1):75. https://doi.org/10.1186/s12302-023-00779-4
2. Bayer JM, Scully RA, Dlabola EK, Courtwright JL, Hirsch CL, Hockman-Wert D, Miller SW, Roper BB, Saunders WC, Snyder MN (2023) Sharing FAIR monitoring program data improves discoverability and reuse. Environ Monit Assess 195(10):1141. https://doi.org/10.1007/s10661-023-11788-4
3. Feraud M, O'Brien JW, Samanipour S, Dewapriya P, van Herwerden D, Kaserzon S, Wood I, Rauert C, Thomas KV (2023) InSpectra – a platform for identifying emerging chemical threats. J Hazard Mater 455:131486. https://doi.org/10.1016/j.jhazmat.2023.131486
4. Alygizakis NA, Oswald P, Thomaidis NS, Schymanski EL, Aalizadeh R, Schulze T, Oswaldova M, Slobodnik J (2019) NORMAN digital sample freezing platform: a European virtual platform to exchange liquid chromatography high resolution-mass spectrometry data and screen suspects in "digitally frozen" environmental samples. Trends Analyt Chem 115:129–137. https://doi.org/10.1016/j.trac.2019.04.008
5. Singer H, Boog J, Kammer K, Schiappacasse FC (2025) NTSuisse - The Swiss LC-HRMS surface water data platform. https://www.ntsuisse.ch/. Accessed 12 Jun 2026
6. ICPR (2024) Pilot project on non-target screening. https://www.iksr.org/en/iksr/rhein-2040/rhine-project-non-target-screening. Accessed 12 Jun 2026
7. DVGW (2024) DVGW-Forschungsprojekt K2I-Xplore. https://www.dvgw.de/themen/forschung-und-innovation/forschungsprojekte/dvgw-forschungsprojekt-k2i-xplore. Accessed 5 Sep 2025
8. Jewell KS, Thron F, Ehlig B, Skottnik J, Scharrenbach T, Ternes TA, Wick A, Kronsbein AL, Koschorreck J (2025) Online-portal non-target screening für die Umweltüberwachung der Zukunft". https://www.umweltbundesamt.de/publikationen/online-portal-non-target-screening-fuer-die. Accessed 12 Jun 2026
9. Jewell KS, Kunkel U, Ehlig B, Thron F, Schlüsener M, Dietrich C, Wick A, Ternes TA (2020) Comparing mass, retention time and MS2 spectra as criteria for the automated screening of small molecules in aqueous environmental samples analyzed by LC-QToF-MS/MS. Rapid Commun Mass Spectrom 34:e8541. https://doi.org/10.1002/rcm.8541
10. Jewell KS, Dietrich C, Lessmann O, Köppe T, Prodöhl F, Wick A (2026) ntsworkflow - A non-target screening data processing tool. https://github.com/bafg-bund/ntsworkflow. Accessed 12 Jun 2026
11. Hollender J, van Bavel B, Dulio V, Farmen E, Furtmann K, Koschorreck J, Kunkel U, Krauss M, Munthe J, Schlabach M (2019) High resolution mass spectrometry-based non-target screening can support regulatory environmental monitoring and chemicals management. Environ Sci Eur 31(1):42. https://doi.org/10.1186/s12302-019-0225-x
12. Kronsbein AL, Wallis RM, Winter E, Bandow N, Jewell KS, Dierkes G, Wick A, Koschorreck J (2026) A regulatory perspective on the systematic use of high-resolution mass spectrometry non-target screening data

in environmental monitoring and chemicals management – use cases from the German NTSPortal. Environ Sci Eur. https://doi.org/10.1186/s12302-026-01368-x
13. Lessmann O, Jewell KS, Ehlig B, Kunkel U, Macherius A, Winter E, Schlüsener MP, Scheurer M, Hauns J, Bammert A, et al. (2026) Collective spectral library (26.1). https://doi.org/10.5281/zenodo.18745439
14. Jewell KS, Prödohl F, Ehlig B, Skottnik J, Lessmann O, Hermes N, Wick A (2025) ntsportal - Functions for building and managing the NTSPortal database (documentation). http://docs.ntsportal.bafg.de. Accessed 12 Jun 2026
15. Schlüsener MP, Kunkel U, Ternes TA (2015) Quaternary triphenylphosphonium compounds: a new class of environmental pollutants. Environ Sci Technol 49(24):14282–14297. https://doi.org/10.1021/acs.est.5b03926
16. Boulard L, Dierkes G, Schlüsener MP, Wick A, Koschorreck J, Ternes TA (2020) Spatial distribution and temporal trends of pharmaceuticals sorbed to suspended particulate matter of German rivers. Water Res 171:115366. https://doi.org/10.1016/j.watres.2019.115366
17. Nürenberg G, Schulz M, Kunkel U, Ternes TA (2015) Development and validation of a generic nontarget method based on liquid chromatography – high resolution mass spectrometry analysis for the evaluation of different wastewater treatment options. J Chromatogr A 1426:77–90. https://doi.org/10.1016/j.chroma.2015.11.014
18. Kessner D, Chambers M, Burke R, Agus D, Mallick P (2008) ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics 24(21):2534–2536. https://doi.org/10.1093/bioinformatics/btn323
19. Jewell KS, Prödohl F, Ehlig B, Skottnik J, Lessmann O, Hermes N, Wick A (2025) ntsportal - Functions for building and managing the NTSPortal database (repository). https://github.com/bafg-bund/ntsportal. Accessed 12 Jun 2026
20. Sokolow S, Karnofsky J, Gustafson P (1978) The Finnigan library search program
21. Jewell KS, Köppe T, Ehlig B, Prodöhl F, Wick A (2025) shewhart4hrms - A package for automated processing of LC-HRMS data for analytical quality control purposes. https://github.com/bafg-bund/shewhart4hrms. Accessed 12 Jun 2026
22. Schulz W (2025) Application of non-target screening by means of LC-ESI-HRMS in water analysis. https://www.gdch.de/fileadmin/downloads/Netzwerk_und_Strukturen/Fachgruppen/Wasserchemische_Gesellschaft/Publikationen/Leitfaden_NTS_EN_V2_final.pdf. Accessed 12 Jun 2026
23. Jewell KS, Ehlig B, Wick A (2021) shewhart4cds - Interactive Shewhart quality control charts for HRMS instruments with automated calibration. https://github.com/bafg-bund/shewhart4cds. Accessed 12 Jun 2026
24. Hermes N, Jewell KS, Wick A, Ternes TA (2018) Quantification of more than 150 micropollutants including transformation products in aqueous samples by liquid chromatography-tandem mass spectrometry using scheduled multiple reaction monitoring. J Chromatogr A 1531:64–73. https://doi.org/10.1016/j.chroma.2017.11.020
25. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA (2016) PubChem substance and compound databases. Nucleic Acids Res 44(D1):D1202–D1213. https://doi.org/10.1093/nar/gkv951
26. Ulrich N, Goss K-U, Ebert A (2021) Exploring the octanol–water partition coefficient dataset using deep learning techniques and data augmentation. Commun Chem 4(1):90. https://doi.org/10.1038/s42004-021-00528-9
27. Loos R, Daouk S, Marinov D, Gómez L, Porcel-Rodríguez E, Sanseverino I, Amalric L, Potalivo M, Calabretta E, Ferenčík M et al (2024) Summary recommendations on "Analytical methods for substances in the Watch List under the Water Framework Directive." Sci Total Environ 912:168707. https://doi.org/10.1016/j.scitotenv.2023.168707
28. Daoud F, Zuehlke S, Spiteller M, Kayser O (2020) Ozonation of rivaroxaban production waste water and comparison of generated transformation products with known in vivo and in vitro metabolites. Sci Total Environ 714:136825. https://doi.org/10.1016/j.scitotenv.2020.136825
29. Wingert NR, dos Santos NO, Nunes MA, Gomes P, Müller EI, Flores ÉM, Steppe M (2016) Characterization of three main degradation products from novel oral anticoagulant rivaroxaban under stress conditions by UPLC-Q-TOF-MS/MS. J Pharm Biomed Anal 123:10–15. https://doi.org/10.1016/j.jpba.2016.01.053
30. Wingert NR, Arbo MD, Göethel G, da Costa B, Altknecht LF, Garcia SC, Steppe M (2019) In vitro toxicity assessment of rivaroxaban degradation products and kinetic evaluation to decay process. Drug Chem Toxicol 42(5):509–518. https://doi.org/10.1080/01480545.2018.1452931
31. Krauss M, Hug C, Bloch R, Schulze T, Brack W (2019) Prioritising site-specific micropollutants in surface water from LC-HRMS non-target screening data using a rarity score. Environ Sci Eur 31(1):45. https://doi.org/10.1186/s12302-019-0231-z
32. Elastic BV (2025) Significant terms aggregation. https://www.elastic.co/docs/reference/aggregations/search-aggregations-bucket-significantterms-aggregation. Accessed 15 Jun 2026
33. Hermes N, Wick A, Ternes TA, Mayer J, Martin H, Leßmann B (2025) Studie zur chemischen Charakterisierung und Quantifizierung bisher nicht bekannter Stoffe in Wässern des Landgrabensystems (Hessisches Ried). Bundesanstalt für Gewässerkunde, Koblenz
34. Anliker S, Loos M, Comte R, Ruff M, Fenner K, Singer H (2020) Assessing emissions from pharmaceutical manufacturing based on temporal high-resolution mass spectrometry data. Environ Sci Technol 54(7):4110–4120. https://doi.org/10.1021/acs.est.9b07085
35. Chonova T, Ruppe S, Langlois I, Griesshaber DS, Loos M, Honti M, Fenner K, Singer H (2025) Unveiling industrial emissions in a large European river: insights from data mining of high-frequency measurements. Water Res 268(Pt B):122745. https://doi.org/10.1016/j.watres.2024.122745
36. Dancho M, Vaughan D (2025) timetk: A tool kit for working with time series. R package version 2.9.1.9000. https://github.com/business-science/timetk. Accessed 12 Jun 2026
37. Grell W, Hurnaus R, Griss G, Sauter R, Rupprecht E, Mark M, Luger P, Nar H, Wittneben H, Müller P (1998) Repaglinide and related hypoglycemic benzoic acid derivatives. J Med Chem 41(26):5219–5246. https://doi.org/10.1021/jm9810349
38. Dierkes G, Boulard L, Wick A (2024) Non-target screening in Schwebstoff- und Biotaproben. https://www.umweltbundesamt.de/publikationen/non-target-screening-in-schwebstoff-biotaproben.
39. Dierkes G, Schmidt S, Meier C, Ziegler K, Koschorreck J, Wick A (2025) Retrospective trend analysis of biocides in suspended particulate matter of major German rivers. Environ Sci Eur 37(1):16. https://doi.org/10.1186/s12302-025-01053-5
40. Brand S, Schlüsener MP, Albrecht D, Kunkel U, Strobel C, Grummt T, Ternes TA (2018) Quaternary (triphenyl-) phosphonium compounds: environmental behavior and toxicity. Water Res 136:207–219. https://doi.org/10.1016/j.watres.2018.02.032
41. Köppe T, Jewell KS, Ehlig B, Wick A, Koschorreck J, Ternes TA (2023) Identification and trend analysis of organic cationic contaminants via non-target screening in suspended particulate matter of the German rivers Rhine and Saar. Water Res 229:119304. https://doi.org/10.1016/j.watres.2022.119304
42. Mohapatra S, Yutao L, Goh SG, Ng C, Luhua Y, Tran NH, Gin KY-H (2023) Quaternary ammonium compounds of emerging concern: classification, occurrence, fate, toxicity and antimicrobial resistance. J Hazard Mater 445:130393. https://doi.org/10.1016/j.jhazmat.2022.130393
43. Mustafa NM, Jumaah FN, Ludin NA, Akhtaruzzaman M, Hassan NH, Ahmad A, Chan KM, Su'ait MS (2024) Tetraalkylammonium salts (TAS) in solar energy applications – a review on in vitro and in vivo toxicity. Heliyon 10(7):e27381. https://doi.org/10.1016/j.heliyon.2024.e27381
44. Markiewicz M, Jungnickel C, Stolte S, Białk-Bielińska A, Kumirska J, Mrozik W (2017) Ultimate biodegradability and ecotoxicity of orally administered antidiabetic drugs. J Hazard Mater 333:154–161. https://doi.org/10.1016/j.jhazmat.2017.03.030
45. ECHA Chem (2026) REACH registration dossier for {N,N,N′,N′-tetrakis(2-hydroxypropyl)ethylenediamine} (CAS-RN 102-60-3), Environmental hazard assessment. https://chem.echa.europa.eu. Accessed 20 Jan 2026
46. Yamamoto K, Asano Y (2015) Efficient production of lumichrome by *Microbacterium* sp. strain TPU 3598. Appl Environ Microbiol 81(21):7360–7367. https://doi.org/10.1128/AEM.02166-15
47. Bushuiev R, Bushuiev A, Samusevich R, Brungs C, Sivic J, Pluskal T (2026) Self-supervised learning of molecular representations from millions of tandem mass spectra using DreaMS. Nat Biotechnol 44(4):630–640. https://doi.org/10.1038/s41587-025-02663-3
48. Tisler S, Kilpinen K, Pattison DI, Tomasi G, Christensen JH (2023) Quantitative nontarget analysis of CECs in environmental samples can be improved by considering all mass adducts. Anal Chem 96(1):229–237. https://doi.org/10.1021/acs.analchem.3c03791

49. Malm L, Palm E, Souihi A, Plassmann M, Liigand J, Kruve A (2021) Guide to semi-quantitative non-targeted screening using LC/ESI/HRMS. Molecules. https://doi.org/10.3390/molecules26123524
50. Jewell KS, Ehlig B, Prodöhl F, Hermes N, Dierkes G, Wick A (2025) Aggregated non-target screening results from surface water monitoring programs in {Germany} . https://doi.org/10.5281/zenodo.17830933 Accessed 12 Jun 2026
51. Jewell KS, Prödohl F, Ehlig B, Skottnik J, Lessmann O, Hermes N, Wick A (2025) NTSPortal (dashboards). https://ntsportal.bafg.de. Accessed 12 Jun 2026
52. Jewell KS, Prödohl F, Ehlig B, Skottnik J, Lessmann O, Hermes N, Wick A (2025) NTSPortal (API). https://elastic.bafg.de. Accessed 12 Jun 2026

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.