EnDiTrap: Pull-down-based pipeline for detection of endocrine-disrupting chemicals

Sehnal, L.[1], Codling, G.[3], Smutná, M.[1], Grabic, R.[2], Hilscherová, K.[1*]

Affiliations:

[1]*Masaryk University, Faculty of Science, RECETOX, Kamenice 753/5, 625 00 Brno, Czech Republic*

[2]*University of South Bohemia in České Budějovice, Faculty of Fisheries and Protection of Waters, Research Institute of Fish Culture and Hydrocenoses, Zátiší 728/II, 389 25 Vodňany, Czech Republic*

[3]*Centre for Resilience in Environment, Water and Waste (CREWW), University of Exeter, N. Park Road Exeter, Devon EX4 4QE, UK*

**corresponding author: Klara Hilscherova (klara.hilscherova@recetox.muni.cz)*

Abstract

Most environmental matrices contain a diverse range of synthetic and natural compounds, some of which exhibit toxic effects. Nevertheless, current methods often do not provide sufficient resolution and power to specifically identify compounds responsible for the biological effects of the complex environmental mixtures. Here, we present the development of the EnDiTrap pipeline that facilitates the identification of compounds that specifically interact with targets involved in the regulation of the endocrine system. We used heterologously expressed ligand binding domain (LBD) of retinoic acid receptor alpha (RARα) as a model system for pull-down-based pipeline development, consequent optimization, and standardization. The applicability of the EnDiTrap pipeline was validated using standard ligand and tested through a case study with environmental samples of freshwater blooms. Results showed that the EnDiTrap pipeline significantly helps to reduce the number of putative features and facilitates the identification of suspect compounds responsible for observed biological effects. We also compared the performance of the software tools Compound Discoverer and MSDial commonly used for processing and analysis of mass spectrometry data. This comparison provided insight into the impact of different software processing on the outcome that brought interestingly contrasting results. This study enhances our ability to specifically identify effect drivers in environmental mixtures of chemicals. Moreover, the developed EnDiTrap pipeline can be applied to various protein targets thus presenting broad applicability.

Introduction

Endocrine-disrupting chemicals (EDCs) have become one of the greatest concerns globally, regarding their adverse effect on the environment and human health. In animal and human health, hormonal regulation by nuclear receptors (NRs) plays an essential role in maintaining an organism's homeostasis, however, this system can be strongly affected by even low doses of EDCs. There are numerous examples of NRs dysregulation by ECDs of environmental origin; for instance, retinoic acid receptor (RAR) activation by cyanobacterial water blooms extracts ((1), thyroid hormone receptor (TR) activation by indoor dust (2) or estrogen receptor (ER) and aryl hydrocarbon receptor (AhR) activation by wastewater samples (3). Moreover, La Merrill et al. (4) summarized key characteristics of EDCs showing that NRs can be affected in several ways. Although many EDCs have been identified, these compounds can often explain only small fractions of the bioactivity detected in environmental pollutant mixtures (5) Thus, the development of novel approaches for the effective identification of EDCs responsible for/contributing to the effects of environmental mixtures is necessary.

Utilizing *in vitro* bioassays to guide chemical analyses provides a robust approach for identifying bioactive compounds. The emergence of high-resolution mass spectrometers (HRMSs) capable of operating in full scan has further increased this capability allowing an extensive range of compounds to be measured in a single instrumental run compared to traditional target methods. However, simply running a complex sample with a known biological effect on an HRMS system does not provide the solution. A single sample chromatogram may contain thousands or hundreds of thousands of chemical features, most of which do not account for/contribute to the biological effect. Thus, either a list of candidate compounds or, in some cases, specific chemical markers such as halogenated mass spectra patterns are used. Both approaches lead to an incomplete picture of the total effect, as they capture only known or suspect compounds. Libraries and databases of chemicals are expanding exponentially and, when properly curated, can identify numerous untargeted compounds in complex mixtures. With thorough investigation, some of these may be linked as potential candidates for the effects detected in the bioassay. But features not in libraries or not previously associated with the bioassay effect will still be omitted. The pull-down assay offers a solution by exposing a sample to a specific protein, where only compounds that will bind to that protein are captured. This *in vitro* technique has been largely used for the analysis of physical protein-protein interactions(6,7) or protein-ligand interaction. Thus, adapting a pull-down assay to capture the EDCs interacting with specific proteins (mainly receptors), offers a powerful approach to search for unknown EDCs. This approach, coupled with quality controls, can reduce the number of features investigated in a sample from thousands to hundreds or fewer. Successful application of pull-down assay concept to investigate EDCs in environmental samples have been already presented. Study by Peng et al. (8) used complete PPARγ to study oil sands process-affected water (OSPW), Sun et al. (9) used complete transthyretin (TTR) and ligand binding domain (LBD) of TR to study commercial chlorinated paraffins (CDs) mixtures while Jia et al. (10) used RARα ligand binding domain (LBD) to investigate EDCs in an organic extract of dust particles. In all these studies, pull-down assay served as a very effective tool to facilitate bioactive ligand detection. Based on these successful applications, framework *the environmental Chemical-Protein Interaction Network* (eCPIN) has been proposed (11). However, standardization and optimization of pull-down-based experiments, that would allow the wider and faster application of this effective technology to explore unknown effect-driving EDCs, is required.

Herein, we present a pipeline EnDiTrap that should serve for standardization of pull-down-based procedure for the investigation of effect-driving EDCs. The NR RARα was used as a model receptor for optimization and consequent case study with freshwater blooms. RAR plays a major role in the regulation of retinoids-mediated processes in animals. The most potent and natural ligand of RAR is all-*trans* retinoic acid (ATRA). The proper regulation of this NR is particularly important for normal embryogenesis. Sehnal et al. (1) identified broad spectra of retinoids produced by freshwater blooms dominated by cyanobacteria that have an affinity to RAR and thus can affect retinoid signalling, e.g., 5,6epoxy-ATRA or 4keto-ATRA.

Interestingly, retinoids are not the only group of natural products able to interact with RAR. Eroglu et al. (12) showed that apocarotenoids, specifically β-apo-14´-carotenal, β-apo-14´-carotenoic acid, and β-apo-13-carotenone, antagonized ATRA-induced transactivation of RAR or study Jia et al. (10) identified antagonists of RAR receptor in an organic extract of dust particles using the pull-down assay. The identified RAR antagonists constitute industrial contaminants (e.g., di(2-ethylhexyl) phthalate (DEHP), tris(2-ethylhexyl) phosphate (TEHP), or bis(2-ethylhexyl)phenyl phosphate (BEHPP)). Therefore, the identification of broader spectra of molecules interacting with RAR constitutes essential knowledge to fully identify the hazardous potential of environmental mixtures of chemicals regarding disruption of RAR signalling, not only in water blooms or dust particles but elsewhere.

Frequently, compounds interacting with NR can cause a significant response at low concentrations. The experimental setup must minimize the matrix effect as much as possible. Here, we aim to (i) develop a standardized and optimized pull-down-based pipeline for the identification of compounds interacting with the nuclear receptor, (ii) develop an analytical approach combining targeted and non-target analysis to identify the broadest spectra of compounds separated by pull-down assay, and (iii) apply developed methodologies to samples from freshwater blooms dominated by cyanobacteria that showed retinoid-like activities to identify compounds that can interact with RAR. Our developed EnDiTrap pipeline can be directly used for the study of other targets and also implemented into broader initiatives such as eCPIN framework.

2. Materials & Methods

2.1 Chemicals and reagents

Solvents methanol (MeOH), acetonitrile (ACN), dichloromethane (DCM) and acetone used in extraction or preparation of equipment were from J.T. Baker (Pennyslvania USA) and were pesticide reagent grade. For LC-Orbitrap MS, solvents (MeOH absolute) and ultra-pure water were from Biosolve Chimie (Dieuze, France) grade U LC/MS-CC/SFC. 2ml vials, spring inserts, and 9mm screw caps (PFTE/S) were from Agilent (CA, USA). Formic Acid and Ammonium formate (>99.9%) were supplied by Honeywell-Fluka (ThermoScientific, MA, USA) and Sigma Aldrich (MI, USA), respectively. Separation of compounds on HPLC used an Aquity UPLC BEH C18 1.7uM, 2.1 x 100mm column with a Vanguard pre-column 3.2 x 5mm (Waters, MA, USA). Where possible, all glassware was washed in a laboratory dishwasher, baked at 450 °C for 12 hours, or sterilized. Six standards of retinoic acids (RA), along with one mass-labelled standard, were selected for method development: ATRA, 4 keto-ATRA, 4 OH-ATRA, 9 *cis*-RA, 13 *cis*-RA, 5,6 epoxy-ATRA, and ATRA-D_5 (supplied at >99% purity). The six RA standards were dissolved in MeOH and combined into a stock solution at 0.1 µg/mL. The nine-point calibration curve was diluted from this stock within the range of 0.1 to 1000 ng/mL. As the standard was likely to degrade, a new calibration solution was prepared for each instrument batch to ensure peak intensities comparable with the previous sample series.

2.2 Sample collection and sites description

Cyanobacterial blooms' biomass was collected from water bodies where freshwater blooms had been observed. The freshwater bloom biomass was sampled from multiple independent stagnant water bodies across South Bohemia and South Moravia in the Czech Republic. From several sampling campaigns, three samples, one collected in the summer of 2013, one from 2019, and one from 2020, that exhibited high RAR-mediated bioactivity detected by *in vitro* assay were selected for use in the assessment of pull-down assay method (for a real-world test of the RAR pull-down process). Information on the sampling sites and sampled biomass is included in Supplementary materials (SI, Table S1). Sampling was conducted between July and September using a plankton net (20 mm mesh). For taxonomic classification, 40 mL of biomass was collected into a 50 mL polypropylene Falcon tube (Fisher Scientific. MA, USA), with 10 mL of 10% aqueous formaldehyde solution (final concentration 2%) to fixate the samples immediately after sampling, and placed in a cool box with ice packs before transport to the laboratory, and stored at -20°C.

2.3 Processing of freshwater blooms biomass

Collected biomass samples were lyophilized (Freeze dryer Gamma 1e16 LSCplus, Martin Christ Freeze Dryers, Germany), and dry biomass was extracted according to Sehnal et al. (1). Briefly, 100 mg of sample biomass was extracted by sonication for 2 x 2 min (100% power, cycle 0.9) in 5 mL of MeOH on ice. The sample was centrifuged at 3050 x g for 5 min, and the supernatant was transferred to a 15 mL pre-baked glass vial. The biomass pellet was resuspended in 1 mL of MeOH and sonicated for 0.5 min, followed by centrifugation. The supernatant was combined with the first extract. A final addition of 5mL of MeOH to the pellet with resuspension, sonication (0.5 min), and centrifugation was repeated. All extracts from the repeated extractions were combined into one sample. The extract was evaporated to ~50 µL under a gentle stream of nitrogen at room temperature and diluted in 250 µL MeOH to reach a biomass concentration of 400 g dry mass (dm)/L. The extracts prepared from the three selected cyanobacterial blooms collected in 2013, 2019, and 2020 were labelled as BM13, BM19, and BM20, respectively, where the BM stands for biomass and the number for the last digits of the collection year.

2.4 Protein expression and purification

2.4.1 Bioinformatic analysis

Basic bioinformatic analysis of the human RARα ligand binding domain (LBD) was carried out to facilitate expression and purification steps. The protein molecular weight (Mr) and isoelectrical point (pI) were calculated using ProteinPredict (13) and SnapGene Viewer (www.snapgene.com), the secondary structure of the protein was calculated using JPred4 (14), protein solubility in Escherichia coli was predicted using SoluProt (15), and transmembrane topology using Phobius (16).

2.3.1 Expression

The nucleotide sequence of the human RARα LBD region, including Ser175 - Asp421 (NCBI protein accession No. NP_000955.1) fused in frame with T7 tag at the N-terminal of the protein-coding sequence was commercially synthesized by GeneArt Service (ThermoFisher Scientific, USA) and consequently subcloned into expression plasmid pET28a+ (as shown in Figure S1).

Expression plasmid was transformed into chemically competent Escherichia coli BL21(DE3) cells, and 10 mL 2 x LB starter culture with kanamycin (50 µg/mL) was inoculated with a single colony of 2 x LB agar culture with kanamycin. The starter culture was grown at 37°C, 150 rpm, for 6 hours, and then 1 mL of starter culture was inoculated into 100 mL expression media (2xLB medium). Expression culture was grown at 37°C, 150 rpm, into OD_{600} = 1, and then expression was started by the addition of isopropyl thio-β-D-galactoside (IPTG) to a final concentration of 0.5 mM. Protein was expressed at 18°C, shaking at 150 rpm, for 16 hours. After that, cells were collected by centrifugation at 4000 x g, 4°C, for 30 min. Cells were resuspended in 8 mL purification buffer (20 mM Tris-Cl, 200 mM NaCl, 5% glycerol, 0.05% Tween20, pH 7). The cell suspension containing expressed protein was stored at -80°C, or the protein was immediately purified.

2.3.2 Purification

The human RARα LBD (247aa) protein expressed in frame with polyhistidine (His) tag, thrombin cleavage site, and T7 tag was purified using HisPur™ Ni-NTA Resin (ThermoFisher Scientific) and Poly-Prep® chromatography columns (Bio-Rad). To extract expressed protein, cells were resuspended in the purification buffer and sonicated for 3 x 20 sec followed by 3 x10 sec. After that, the extract was centrifugated at 17000 x g, 4 °C, for 30 min, and the supernatant was incubated with 2 mL HisPur™ Ni-NTA Resin at 4 °C, for 90 min in Poly-Prep® chromatography columns.

The resin with bound protein was washed twice with 6 mL of purification buffer containing 10 mM imidazole. After the wash steps, the purified protein was eluted with a purification buffer containing 250 mM imidazole. Purified protein was then desalted by ZEBA spin desalting columns (ThermoFisher Scientific) in a purification buffer with 5mM β-mercaptoethanol (β-ME) and filtered by sterile filter (22µm), also to remove imidazole. The protein concentration was quantified using RC-DC protein assay (Bio-Rad) and nanodrop.

2.4 Pull-down assay procedure and optimization

Critical conditions of the pull-down assay procedure were identified and optimized. These conditions included protein amount, incubation temperature, MeOH concentration in the pull-down mixture, sequential elution by imidazole and MeOH, and the addition of BSA. The impact of individual conditions was analyzed using *in vitro* bioassay (Figure S2).

For the final optimized procedure, 100 µg of RARα LBD was mixed with 990 µL of purification buffer. This solution was spiked with 10 µL of ATRA (50 µM) or MeOH extracts of freshwater bloom biomass (BM) samples (400 g/L dm) to the final concentration of 1% MeOH and incubated at 4 °C for 60 min. Next, 10 µL of HIS-select nickel magnetic agarose beads were added to each sample, and the mixture was incubated at 4 °C on a horizontal mixer for 60 min. Then, tubes were placed in a magnetic separator for 10 seconds, and the supernatant was removed. Samples were washed three times with 500 µL of purification buffer to remove chemicals that were non-specifically bound to RARα and tubes. The complex of ligands and RARα LBD was eluted with 50µL of elution buffer (20 mM Tris-Cl, 200 mM NaCl, 5% glycerol, 0.05% Tween20, and 250 mM imidazole, pH 7), and this step was repeated two times. In the optimization experiment, three initial elutions by elution buffer with imidazole were followed by three elutions by 50uL MeOH. However, only one MeOH elution followed in experiments with extracts of freshwater bloom biomass. Individual elutions prepared with elution buffer (containing imidazole) were mixed with 450 µL of MeOH and incubated for 10 min at room temperature to dissociate the protein-ligands complex. These elutions were consequently concentrated to a volume of 50 µL.

Two types of negative controls were used. The first, control for non-specific binding (NSB) of other components present in the pull-down system. This control variant contained all reaction components except of the His-RARα LBD. This control is further referred to as NSB control. Importantly, for the NSB samples, each BM had a corresponding triplicate of NSB sample with one imidazole and one MeOH. The second, protein control (PC) contained all reaction components except of the ATRA or sample of freshwater bloom biomass. The purpose of this negative control is detection of background features that originate from the His-RARα LBD. This control is further referred to as PC control. The PC control samples were processed with the same elution procedure as BM samples. For chemical analysis, instrumental blanks (n=30) were also prepared.

2.5 In vitro activity testing

The retinoid-like activity was analyzed using the RARα Reporter (Luc)-HEK293 Cell Line containing a firefly luciferase gene under the control of retinoic acid response elements stably integrated into HEK293 cells along with full-length human RARα (NM_000964). Further details about bioactivity analysis by this method are provided in Sehnal et al. (17).

2.6 Instrumental Analysis

2.6.1 Target analysis of fortified samples for method development

Each sample and control were fortified with 2.5 µL of 1 µg/mL solution of ATRA-D_5 prior to injection. For method development, the samples were fortified with six retinoic acids (RAs) at differing concentrations (100 – 10,000 ng/mL). All injections were performed blind, which means that codes for each sample were given, and no sample details presented until after integration. The fortified samples that exceeded the linear range of the calibration curve (0.1 to 500 ng/mL) were diluted 10 and 100-fold and reinjected for reanalyses.

Target analysis of extracts was performed on a Xevo TQ-S (Waters, Manchester UK) connected to an Acquity UPLC (Waters, Manchester UK). The autosampler was thermostatted to 10 °C, and the column temperature was maintained at 40 °C. An elution gradient method was used with 0.1% formic acid in nanopure water (A) and ACN (B). The gradient started with 20% B, holding for 1 minute before increasing to 70% B over 1 minute and then increasing to 100 % B over the next four minutes and holding for two minutes before decreasing to 20% B and holding for 2 minutes. The sample was sent to waste during the first 1.5 minutes and the last 3 minutes of the instrument run. The flow rate was 300 µL/min, and the injection volume was 5 µL per sample. Detection by MS was in ESI positive (ESI +) using tandem mass spectrometry. The capillary voltage was 0.8 kV, and the collision gas, cone, and desolvation flow were 0.01, 150, and 600 L/h, respectively. The quantification was based on the ratio of the analyte to the internal standard (ATRA-D_5). The limit of detection (LOD, S/N>3.3) and quantification (LOQ, S/N>10) were generated from the regression of the calibration curve. All parent ions and transitions for target and internal standard, along with the LOD and LOQ, are presented in supplementary Table S2. Results were processed using Skyline software for small molecules (Adams et al., 2020), with two transition ions required for confirmation.

2.6.2. Nontargeted HRMS analysis

As the purpose of full scan acquisition in this trial was to identify the greatest number of features binding to the RAR LBD, chromatographic separation conditions varied from that used in target analysis. Changes included choices of buffer, solvent, and gradient to maximize the number of detected features, but at the expense of a more sensitive method optimized for detection of retinoids.

2.6.2.1 Instrumental conditions

Chromatographic separation was performed on a Shimadzu LC (LC-20A, Kyoto Japan). Both mobile phase A (nanopure water) and mobile phase B (MeOH) were buffered with 0.1% formic acid and 5 mM ammonium formate. For separation, Waters C18 column with an AQUITY guard thermostatted at 35 °C was used. Before injection, the column was equilibrated for 3 minutes with 90% mobile phase A. This condition was held for 1 minute after injection when the gradient of B was increased until it was 50:50 at 10 minutes. During the following 3 minutes, the proportion of B increased to 100% and was held for 5 minutes before returning to 90% A. The autosampler was thermostatted at 15 °C, the flow rate was 300 μl min, and the injection volume was 10 μl.

The LC was coupled to an Orbitrap Fusion™ Lumos™ Tribrid™ Mass Spectrometer (Breman Germany) equipped with Thermo Scientific™ OptaMax™ NG (H-ESI II) ion source. The orbitrap was tuned before instrument runs and checked between positive and negative scans and after running for mass drift. The ion source parameters for both positive and negative mode were kept identical with gas rates of 40, 5, and 1 AU (arbitrary units) for sheath, auxiliary, and sweep, respectively. The ion transfer tube and vaporizer temperatures were 325 and 350 °C, respectively. Spray voltage differed with 3500 V and 3000V for positive and negative mode, respectively.

Six MeOH injections were included before the sample injection, and an exclusion list was generated. A list of suspect RAR-agonists was supplied (Table S3), and the [M+H] or [M-H] ion masses were included for suspect screening. The instrument operated in a combination of Full Scan MS and data-independent acquisition (DIA). The mass range used was 100-1000 m/z, Orbitrap resolution was 60,000 FWHM in MS^{1}and the top 10 ions in a scan were selected for DIA. MS^2 resolution was 30,000 FWHM. HPLC eluate was sent to waste, and no spectral data was collected for the first minute and last 4 minutes of each run.

During sample injection, every six injections, a MeOH blank was added, primarily to test for carryover, and used as an instrumental blank. The all-*trans* retinoic acid calibration curve used in target analysis was injected before and after sample injections. Also, one point of the calibration curve at random was injected and checked for mass drift or signal degradation every nine injections.

2.7 Full scan data processing

Results from the raw data files were processed in multiple ways to identify target, suspect and non-target features. Target features, in this case, refer to features where the calibration curve and full spectra are present in high resolution. These features can be attributed to known-known or level 1 confidence (19). The analytes used in the method development of the RAR pull-down assay were individually injected onto Orbitrap at 1 μg/mL in ESI$^+$ to identify peaks and MS^2 ions under full scan data-independent acquisition (DIA; Table S2). Up to fifteen MS^2 fragments were identified for each retinoic acid (RA) compound injected, though priority for confirmation was given to the transitions used in the target method. Target detection was performed using Trace Finder 4.1 (Thermo Fisher, MA, USA).

2.7.1 Full Scan analysis

Raw files were submitted to Compound Discoverer (CD) 3.1 software (Thermo Fisher, MA, USA), which enabled online and library searches for unknown features. CD was developed primarily for hybrid orbital trap instrumentation, and an existing workflow, "Untargeted environmental research ID workflow with statistics' was applied. The method aligned retention times (tolerance 0.1 min), performed unknown compound detection and grouped the compounds based on specified criteria. Background compounds were identified from the instrumental blanks (n>30), and these were excluded from the results using a minimum sample area to the blank ratio of 5. Gap filling was restricted to a signal-to-noise ratio (S/N) of 5, and peak detection S/N was set at 10. Elemental composition was predicted along with online searches on mzCloud, and ChemSpider. ChemSpider was restricted to eight databases to speed up processing (ACToR: Aggregated Computational Toxicology Resource, DrugBank, EAWAG Biocatalysis/Biodegradation Database, EPA DSSTox, EPA Toxcast, FDA UNII – NLM, Marine Drugs and the Royal Society of Chemistry). Results from ChemSpider were further restricted to 20 potential compounds per feature and 3 ppm mass tolerance. A local database search was also included using libraries aggregated from online sources (Massbank, MassBank-EU, MSDial, and RIKEN-MS) supplemented with a formula from the suspect list (Table S3). The S/N for peak detection and gap filling was adjusted from 3 to 10 to observe variation in final feature detection, as was the minimal peak area from 10^3 to 10^6. The lowering of this parameter, although it increased the processing time, led to the identification of thousands more features that were almost all omitted through blank and control settings, but was included as many compounds that bind to RAR are known to be at ultra-low concentrations. It agrees with results of our target analysis that detected also low levels of targeted retinoids.

Analytical ratios between pull-down samples and protein control (PC) or non-selective binding (NSB) were implemented in the workflow for post-processing. During post-processing, all features with RT < 1.5 min, with a molecular weight > 600 m/z, and a ratio of sample peak area to instrument blank < 5 were automatically excluded. Further, workflow continued by sample group and elution filtering based on the PC and NSB controls. The filtration steps are such that a feature is retained if a single elution group from any BM passes all filtration criteria.

The PCs consist of a triplicate protein assay without biomass extract with four elutions per sample, mirroring the sample pull-down procedure. As the first three elutions (EL 1-3) in any sample used the same solvent (imidazole), the maximum peak area of EL 1-3 in the PC samples was compared to each of the first three elutions in the samples. The ratio of sample peak area to PC had to be >4 to retain that feature. For the MeOH elution (EL 4) as there was a single elution, the ratio of BM to PC >4 was used. For NSB, there was one imidazole, and one MeOH elution for each BM sample and a ratio of > 4 was set again for retaining features. Within each BM the ratio of the maximum peak area in EL 1-3 was compared to the maximum peak area in MeOH (EL 4). As EL1-3 and EL 4 were filtered independently potentially a feature may be retained where the elution that passed the filtration steps had a lesser peak area than other elutions that did not pass the filtration step. The ratio criterion was set to be greater than 1 for EL1-3 or less than 1 for EL4.

Raw data files were also submitted to MSDial (MSD) 4.7 software (RIKEN, Kanagawa, Japan) after file conversion using Reifycs Analysis Base File Converter (from Thermo's raw format to .abf format RIKEN, Kanagawa, Japan). The parameters used for MSDial were an MS^1 and MS^2 tolerance of 0.001 and 0.0025, respectively, a smoothing level of 3, and a minimum peak height of 1000. The slice width was set at 0.1 m/z. Library searches used a mass tolerance of 0.01 and 0.05 m/z for MS^1 and MS^2, respectively, and a cut-off of 80 in identification scoring. The retention time tolerance between features was 0.1, and the accurate mass cut-off was set at 0.01. Additional reprocessing was performed using differing MS^1 and MS^2 tolerances and minimum peak heights, but the post-processing outcomes were similar. For each mode (ESI^+ and ESI^-) searches used the ESI(+)-MS/MS from authentic standards (16,481 unique compounds), ESI(-)-MS/MS from authentic standards (9,033 unique compounds) libraries that are available from the RIKEN web page (http://prime.psc.riken.jp/compms/index.html). Subsequent library searches (MassBank, MassBank-EU, ReSpect, RIKEN, GNPS, Fiehn, CASMI2016, MetaboBase, PFAS, and an in-house library) were performed. The use of smaller libraries against the more extensive library was to confirm detection agreement across different records, as though the larger database houses over 60'000 records across both modes, where different records provide variation, false positives can occur.

Additional post-processing of the deconvoluted results from MSD and the unfiltered CD results were performed on JupyterLab (Python Software Foundation, https://www.python.org/), subjected to filtration following the same process as outlined for CD. With instrumental blanks, PC and NSB controls were used to omit features. Where features were non-detect in controls, a value of 1 was submitted to prevent infinite ratios. Python allowed parallel processing of outputs (i.e., both CD and MSD were treated the same) and a more rapid adjustment of filtration parameters. Additional filtration, such as the percentage of variance within a sample, excluded features where the maximum peak area in a sample was more than double the area of other peaks. In addition to comparisons of output from both methods, suspect searches were also done on JupyterLab, as were graphs, diagrams, and statistical analysis.

A final assessment of features was performed by visual observation of each feature, its occurrence/ abundance across all samples, and number of replicates in a sample. The results were grouped into three levels, with level 1 features having multiple detections within an elution and across samples along with PC as background noise or non-detect. Where the NSB was observed but with less than 10% of the greatest peak area, the feature was also considered Level 1. Level 2 features had some PC or NSB noise/background or low feature abundance, and Level 3 features were considered poor candidates for further investigation most often due to features in controls having similar elution pattern to the sample indicating the feature is likely an artefact of the extraction process. For an example of the levels, see the supplementary information (Figure S4).

Those features, from either CD or MSD, reaching level 1 or 2 were checked against the initial deconvolution and library searches for tentative identification. A peak shape score was applied from 1-5 that looked manually at the background noise, peak shape, number of scans, if peak is a suspected fragment and the alignment of peaks to score candidates. Where MS^2 features were present, the spectra were submitted to mzCloud™ (HighChem LLC, Slovakia) and MassBank EU (https://massbank.eu/MassBank). For mzCloud™, spectrum searches were performed in three steps, all requiring a score above 70: initially, an identity search (target); if no candidates were found, a similarity search; and finally, a substructure search. Only the target features are considered in detail in the discussion. Where multiple candidates were presented, a logical score was applied based upon the potential that the compound would be in the target location, (i.e. in this study an anti-cancer drug used in low doses is less likely than a compound associated with algae). For MassBank, only MS2 features were searched, a score >0.6 used and the output compared to the parent mass for confirmation, again a logic in selection was applied. In addition, libraries both in house and within the CD and MSD were applied during the deconvolution and compared to the final candidate list of compounds.

Screening of suspect RAR binding compounds from literature was performed, and a detailed explanation of the suspect screening workflow is provided in the supplementary information (Text S1.1). In brief, a list of 64 suspect compounds was selected based on literature and molecular structure as being suspected RAR agonists (Table S3). Searches on MZCloud by name or fragments of name identified 9-cis Retinal, but no other features were in the database at the time of searching. Other databases had low resolution or no available fragmentation patterns (Mass Bank and PubChem). Initial screening therefore compared the final filtered samples to the parent mass of suspects assuming [M+H] or [M-H]. Where suspected candidates were identified, *in silico*-fragmentation was used to predict the MS^2 (AB, Canada, https://cfmid.wishartlab.com/). As these compounds are detected at low concentrations in the environment, some concern that the filtration process may have excluded features prompted additional screening. In Python, the results from both CD and MSD after instrumental blank, high mass (m/z >600), and low retention time (<1.5 min) exclusion, were screened against the parent masses of the library of 64 candidates at 10 and 5 ppm mass error. As this resulted in some candidates, possible MS^2 fragments were predicted with in silico fragmentation for this selection using CFM-ID on Python for batch prediction. For final confirmation, raw data files were screened for the presence of the predicted fragments using Trace Finder and Skyline.

2.8 Quality Control and Quality Assurance

All used solvents were HPLC grade with the exception of the water and methanol used on the LC-Orbitrap that was ultra-pure. All equipment was run through a laboratory dishwasher, baked where appropriate, solvent washed, and autoclaved.

During extraction, sample and instrument blanks were prepared, and integration was performed blindly; except for the instrumental blanks, samples were coded, and the researcher was not informed of what the codes represented until after the data analysis was completed. This blind assessment prevented biased peak selection. For the target recovery during method development, no compounds were observed above the limit of detection (LOD) in blanks.

Protein controls (PC; n=3) contained all components of the pull-down assay except sample material and were extracted alongside the samples. Non-specific binding controls (NSB; n=3 per sample) consisted of the same BM for each sample and were ran through the same pull-down assay process but without the addition of the RAR-LBD. The PC was used to exclude features from the protein, and the NSB identified features retained within the assay but not bound to the LBD.

Online searches for unknown compounds were restricted only to features with MS^2. Predicted features had to have a fit >70% for consideration and logically be present in the sample environment. This approach is similar to that used in other studies where a review of fragment evidence and sample evidence is used to provide greater confidence (20). Where MSD or CD presented MS^1 candidates through their searches the candidate is included in the SI (Table S6) unless online MS^2 compounds were found. However, these are not discussed further and included only as potential compounds to be included in future targeted screening lists for either confirmation or exclusion.

3. Results & Discussion

EnDiTrap pipeline requires the combination of interdisciplinary approaches including bioinformatics, molecular and synthetic biology, bioactivity detection, and advanced methods of analytical chemistry as shown in Figure 1. This study can serve as a guideline for other researchers interested in application of the EnDiTrap pipeline for discovering target specific compounds that originate from any source.

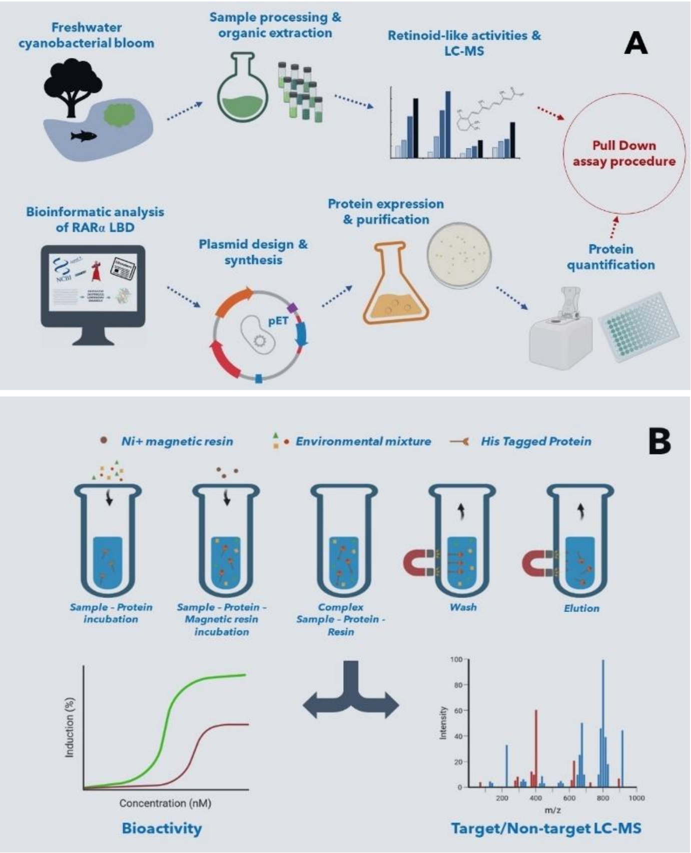
Freshwater cyanobacterial bloom
Sample processing & organic extraction
Retinoid-like activities & LC-MS
A
Pull Down assay procedure
Bioinformatic analysis of RARα LBD
Protein expression & purification
Plasmid design & synthesis
Protein quantification
pET
Ni+ magnetic resin
Environmental mixture
His Tagged Protein
B
Sample - Protein incubation
Sample - Protein - Magnetic resin incubation
Complex Sample - Protein - Resin
Wash
Elution
Induction (%)
Concentration (nM)
Bioactivity
Intensity
100
80
60
40
20
0
200
400
600
800
1000
m/z
Target/Non-target LC-MS

Figure 1. Work pipeline and scheme of the final pull-down assay procedure. **A** – Work pipeline of sample pre-processing and protein preparation. **B** – Scheme of pull-down assay procedure with consequent bioactivity detection, and target and non-target LC-MS analysis.

<u>3.1 RARα LBD protein preparation and pull-down optimization</u>

The cornerstone of the presented approach is an expression and purification of the ligand-binding domain of the target receptor, in this case, human RARα. The use of LBD instead of full-length RARα protein was inspired by the study's objective and the method itself. Since the basic condition is analysis only of compounds that interact with the ligand binding domain of RARα, other protein regions are superfluous and could be responsible for detecting compounds that do not specifically interact with LBD. Moreover, the suitability of RARα LBD for research focused on compounds interacting with this receptor was already shown in previous studies (10,21).

Our bioinformatic analysis of the amino acid sequence of RARα LBD showed that the LBD part does not contain any transmembrane region and is predicted to be soluble in E. coli. Analysis of secondary structure allows protein design that results in protein expression only with complete alpha-helixes and beta-sheets. Such design helps increase yield, stability, and proper folding of protein together with decreasing precipitation tendency. Protein characteristics used during design are summarized in Figure S3.

In order to verify the functionality of the expressed RARα LBD and maximize the informativeness of the pull-down assay, optimization experiments were carried out with standard RAR ligand ATRA. Based on the results of optimization experiments where individual factors were evaluated, final optimized procedure was tested (Figure 2). We decided to analyze broader spectra of retinoids also in optimization experiments based on a recent publication (17) that showed the susceptibility of retinoids to oxidation in water environment. Moreover, these compounds could significantly contribute to the final detected bioactivity, as shown earlier (1,22). Despite of the fact that only ATRA was added to the optimized pull-down assay, mass spectrometry analysis allowed the detection of ATRA, 9/13cis-RA and 4keto-ATRA. The ATRA was the most abundant detected compound, followed by 9/13cis-RA. In addition, the optimized procedure provided information about the elution dynamics of individual compounds (Figure 2). All three replicates provided consistent results where the level of detected compounds correlated with the level of detected bioactivity. The greatest portion of the retinoids was eluted by the first imidazole elution. A significantly lower amount was eluted in the two following imidazole elutions, and only the first MeOH elution contained detectable levels of the retinoids and retinoid-like activity. Therefore, we decided to apply only one MeOH elution in the experiments with the environmental samples.

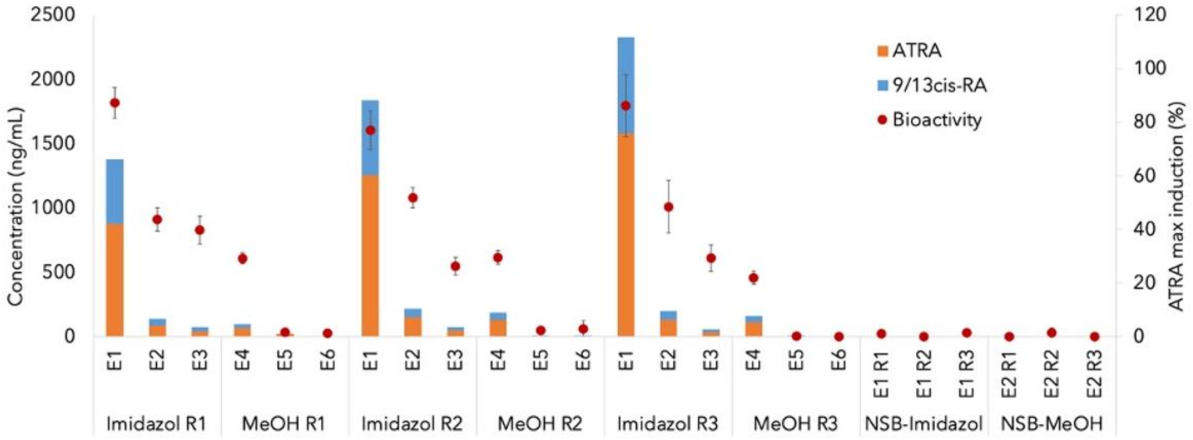

Figure 2. The pull-down assay with sequential elution for procedure optimization. The pull-down assay reaction mixture contained RARα LBD protein and its native ligand ATRA. ATRA was eluted 3x by protein buffer containing 250mM imidazole followed by 3x elution by 100% methanol (MeOH). The NSB control samples were eluted 1x by protein buffer containing 250mM imidazole followed by 1x elution by 100% methanol (MeOH). The concentration of detected retinoids after the pull-down assay is shown in ng/mL. Bioactivity is shown as % of ATRA max induction. E – elution, R – replicate, NSB – nonspecific binding control.

Furthermore, we could quantify total retinoids' recovery after the pull-down assay. Cumulative total retinoids recovery from all consecutive elutions based on LC-MS data were 57%, 77.5%, and 91% in replicates 1, 2, and 3, respectively. The relatively high results variance delineates the importance of a triplicate experiment. In contrast to variations in the concentration of individual retinoids, detected bioactivities reached comparable levels across triplicates. The highest concentration of retinoids was detected in E1 across all replicates of all samples.

3.2 Pull-down assay targeted at RARα ligands in freshwater blooms extracts

The applicability of the developed method for studies involving complex environmental samples was verified using extracts from freshwater bloom biomass dominated by cyanobacteria that are well known for their retinoid-like activities (1,23,24)

Targeted analysis of a broad spectrum of retinoid compounds in samples following the pull-down assay demonstrated successful recovery of various retinoid compounds in all three tested freshwater blooms together with retinoid-like activity. The highest concentration of retinoids and bioactivities across all three replicates were detected in pull-down samples from BM20 (Figure 3). Contrarily, the lowest ones were detected in pull-down samples from BM19. We also calculated the relative recovery of retinoids in pull-down samples compared to the original extract of BM samples (Supplementary Table S4). Percentage recovery after pull-down assay ranged from 73-76%, 144-150%, 16-37% for BM13, BM19, and BM20, respectively. Interestingly, the highest total retinoid concentration in the original extract before pull-down was detected in BM13, and the lowest total retinoid concentration in the original extract before pull-down was detected in BM19. Significant differences in percentage recovery can be explained by the limited capacity of the protein to bind all available ligands and also a different composition of available ligands and their affinities to the ligand binding domain of the protein, as was shown in earlier studies (1,17,25). Moreover, the vulnerability of retinoids to oxidation and degradation by reactive oxygen species in water is an important factor that must be considered in the assessment of retinoids recovery (17). Sample BM19 with higher recovery than 100% can serve as an example. The most probable scenario is the oxidation retinal (not analysed within this study due to very low affinity to RAR) to ATRA and other ATRA derivatives; thus artificially increasing recovery. This process was explained in the study by Sehnal et al. (17), where directed oxidation of less bioactive retinal led to the detection of more bioactive ATRA, and consequent oxidation of ATRA can lead to the production of other retinoids such as 4OH-ATRA, 5,6epoxy RA, or 4keto-ATRA.

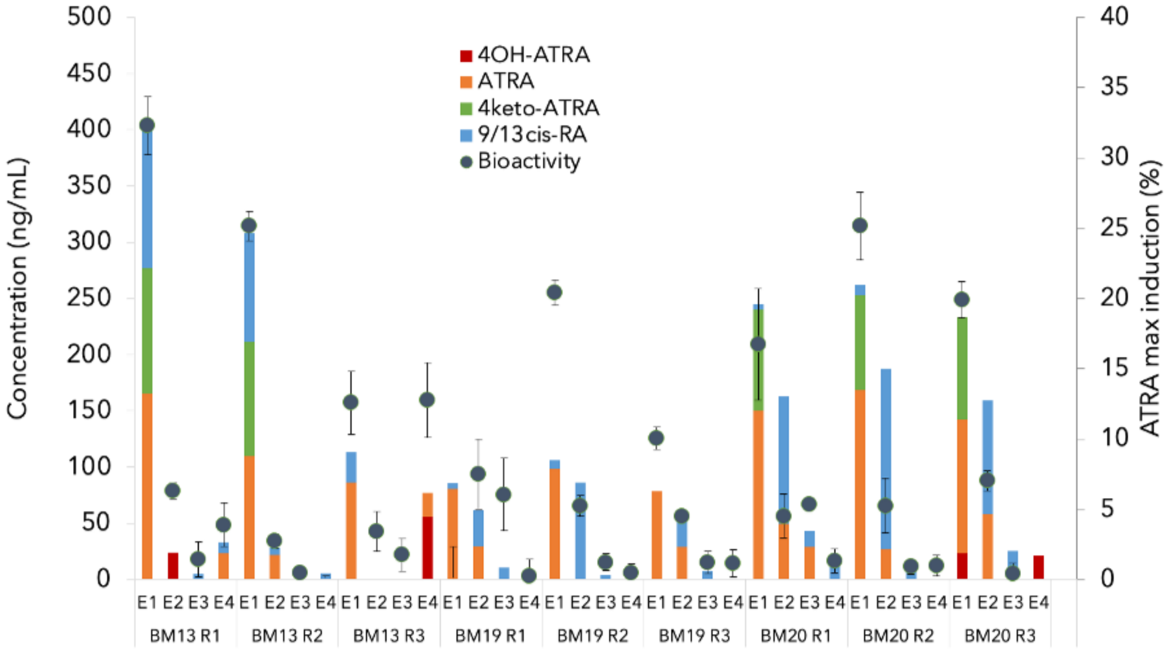

Figure 3. The concentration of retinoids and bioactivities of three different freshwater bloom biomasses after pull-down assay with RARα LBD. The first three elutions (E1-E3) in each replicate (R1-R3) represent elution by protein buffer containing 250mM imidazole. The last elution (E4) in each replicate was carried out by methanol (MeOH). The detected retinoid concentration after the pull-down assay is shown in ng/mL. Bioactivity is shown as % of ATRA max induction. E – elution, R – replicate.

Detection of different retinoids also constitute an important factor regarding total recovery. In particular, different retinoid profiles across samples directly impact detected bioactivity since individual retinoids significantly differ in their binding affinity to RAR. Retinoids ATRA and 9/13cis-RA were detected and quantified in all BM samples after pull-down assay, although these metabolites were not quantified or even detected in original samples of BM (this phenomenon is discussed below). Moreover, these two retinoids were detected in the highest concentration of all analysed retinoids. Besides ATRA and 9/13cis-RA, retinoids 4OH-ATRA, 4keto-ATRA, and 5,6epoxy-RA were detected or quantified in samples after the pull-down assay with extracts from cyanobacterial water blooms.

Our results also document changes in the profiles of detected retinoids in the original BM sample and pull-down samples. Although the most potent RAR ligand, ATRA, was not quantified (NQ) in the original samples BM19 and BM20, this highly bioactive retinoid was detected in the concentration range from 98 ng/mL to 110 ng/mL and 195 ng/mL to 228 ng/mL across pull-down replicates of BM19 and BM20, respectively (Table S4). Similarly, the concentration of 4keto-ATRA in pull-down treated sample BM20 was almost twice as high as in the original biomass. This phenomenon is associated with oxidation/degradation of retinoids by ROS as described above. In contrast, profile of detected retinoids in sample BM13 was different. The ATRA was detected in both original biomass and samples after pull-down in approximately same level. However, compound 9/13cis-RA was detected in all pull-down samples despite its absence in the original biomass which can be again explained by oxidation/degradation of other retinoids. Furthermore, although 5,6epoxy-RA was detected in all original BM samples, this retinoid was not quantified in any sample after pull-down and was detected only in sample BM20. It can be associated with lower potency of 5,6epoxy-RA to bind to the RAR (25). In general, compound affinity is an essential factor that must be considered when using the pull-down approach, e.g., during design of positive control.

3.2 HRMS full-scan data of cyanobacteria samples after RAR pull-down assay

3.2.1 Target Library Search

RAs were detected in target analysis in this study in the BM samples with < 160 ng/mL for individual compounds (Figure 3). However, many RAs were far less than this level. Given the low abundance for many RAs, and the non-specific method, the peaks identified as RAs in samples had low S/N ratios, and when comparing to controls and blanks, no compounds were identified at a high certainty level. Furthermore, 9 *cis*-RA, 13 *cis*-RA and ATRA were not separated chromatographically by the full-scan method. However, multiple compounds with the same MS^1 and some corresponding MS^2 ions as ATRA were observed eluting prior to the retention time window.

3.2.2. Suspect Library Screening

Initial suspect screening of fully processed non-target data identified six potential compounds across both ESI$^+$ and ESI$^-$ in either Compound Discoverer (CD) or MSDial (MSD) data sets. These are 4-Keto 13-*cis*-Retinoic Acid Methyl Ester and β-Apo-14'-carotenal in CD ESI$^-$, 4-Methoxy Retinoic Acid in CD ESI$^+$, β-Apo-14'-carotenal in MSD ESI$^-$ and 3-Dehydro Retinol Acetate, Caged Retinoic Acid, all-trans Retinal MSD ESI$^+$. However, in silico generated MS^2 fragments did not match MS^2 ions for suspect m/z acquired by DIA, and without the reference standards, further investigation of fully filtered data was not pursued.

However, using the data output from MSD and CD in Python (after removal of instrument blank features as described above), the *m/z* selected by either software in MS^1 was screened to ascertain if filtration of the data had omitted the compounds of interest. For CD in ESI$^+$ 48 and 10 features were detected at 10 and 5 ppm mass error, respectively, while for CD ESI$^-$ the same four features were measured at 10 and 5 ppm. In case of MSD ESI$^+$ 9 and 7 features were detected at 10 and 5 ppm error, respectively, while only one feature was observed at 10 ppm and none at 5 ppm for MSD ESI$^-$. Of these features, some were identical for both MSD and CD.

For the potential compounds identified within the 10 ppm mass range the in-silico-generated product ions were predicted. This search identified eleven potential RAR agonists among ESI+ features in the samples corresponding to level 3 identification (19). These compounds had MS^1, and at least two of the predicted MS^2 fragments. However, without the reference standards, or online data bases confirmatory identification was not possible (Table S4). The predicted fragments for compounds in ESI^- was also tested however, no compound was detected with the corresponding MS^2. None of the suspect compounds were measured in either controls or blank samples.

3.2.3 Non-target data processing

The target and suspect screening relied on searching for known-known and unknown-known compounds within the raw or processed data. The non-target utilized processed data and strict filtration criteria for comparing samples to controls and blanks. There were 180,877 and 47,512 features identified via the CD method and 56,761 and 24,995 using MSD for ESI^+ and ESI^- ionization, respectively. These feature numbers are primarily bound by the detection limit criteria set in data processing and include many noise signals. Variable deconvolution criteria, such as changing the gap filling from a signal to noise (S/N) of 1.5 to 5 and adjusting the minimum peak area threshold were tested during processing. The different settings generated more or fewer features, but similar final number of features was obtained when the filtration criteria was applied in Python. Consequently, the use of these high-density output files was not considered problematic, but also did not provide much in the way of additional compounds of interest.

To assess the effectiveness and reproducibility of the pull-down assay and to characterize the differences among samples and respective controls, an OPLS-DA was applied to the raw output from CD and MSD (under both ESI^+ and ESI^- modes; Figure S5). It is apparent in the model that the instrumental blanks stand apart from the sample groups. It is also clear that the features observed in imidazole differ from those of the MeOH elution (EL4). This finding indicates that the MeOH elution may provide additional compounds. This is also substantiated by the fact that the instrumental blank, made up of MeOH used for the extractions and elution, does not overlap with the sample MeOH elution. The feature density over the EL4 was also substantially less implying that the majority of features eluted in the imidazole fraction. Furthermore, though there is some overlap between the NSB and the biomass extracts, the PC fraction overlaps more with the BM samples indicating that the protein may be a primary source of features in the initial observations. As the models are very data-rich and contain a large proportion of noise features, the model fit was poor but did provide a first insight into the results of the pull-down assay (PDA).

Primary filtration, removing features from instrumental blanks (ratio of sample max/blank max >5), m/z > 600 amu, retention time less than 1.5 min and features where the maximum peak area in the samples did not exceed 2000 AU (as described in paragraph 2.7.1.), removed between 76 to 97 % of all the detected features (17,427 and 1889 features retained for CD and 13,346 and 699 features retained in MSD for ESI^+ and ESI,$^-$ respectively). The large discrepancy between the initial feature numbers in data sets by MSD and CD significantly reduced and indicate how many features were noise and for expedience using higher thresholds during processing does not lose a significant number of compounds. To better understand the observed differences among protein controls (PC), non-specific binding controls (NSB), and the three sample groups, the primarily filtered datasets were extracted for OPLS-DA analysis (under both ESI^+ and ESI^- modes). The results and observed distributions indicate that many pull-down specific features were eluted in the first imidazole elution (Figure 4a, b). This finding is documented by its most evident separation from the protein and non-specific binding controls. For ESI^+ (Figure 4a), clear separation between NSB and PC controls is also observed, while for ESI^- (Figure 4b), NSB and PC do not separate well, but the samples do. This separation indicates that many of the features in the controls may be contamination artefacts from sample handling and processing. For the MeOH elution, the distinction between controls and groupings is less clear (Figure 4 c-d). As the MeOH elution finishes the extraction process, many features would likely have already been eluted and the feature density for MeOH is much less than for the imidazole elutions. The separation observed for imidazole extraction EL 1 from controls is observed for BM 13's second and third imidazole elution (Figure S6, c-e) though one sample in BM 13 EL2 appears as an outlier. For BM 19 and 20, particularly for elution 2 (Figure S6c-d), quite a lot of overlap is observed between PC and NSB, potentially indicating that the majority of pull-down specific features were eluted in the first imidazole fraction.

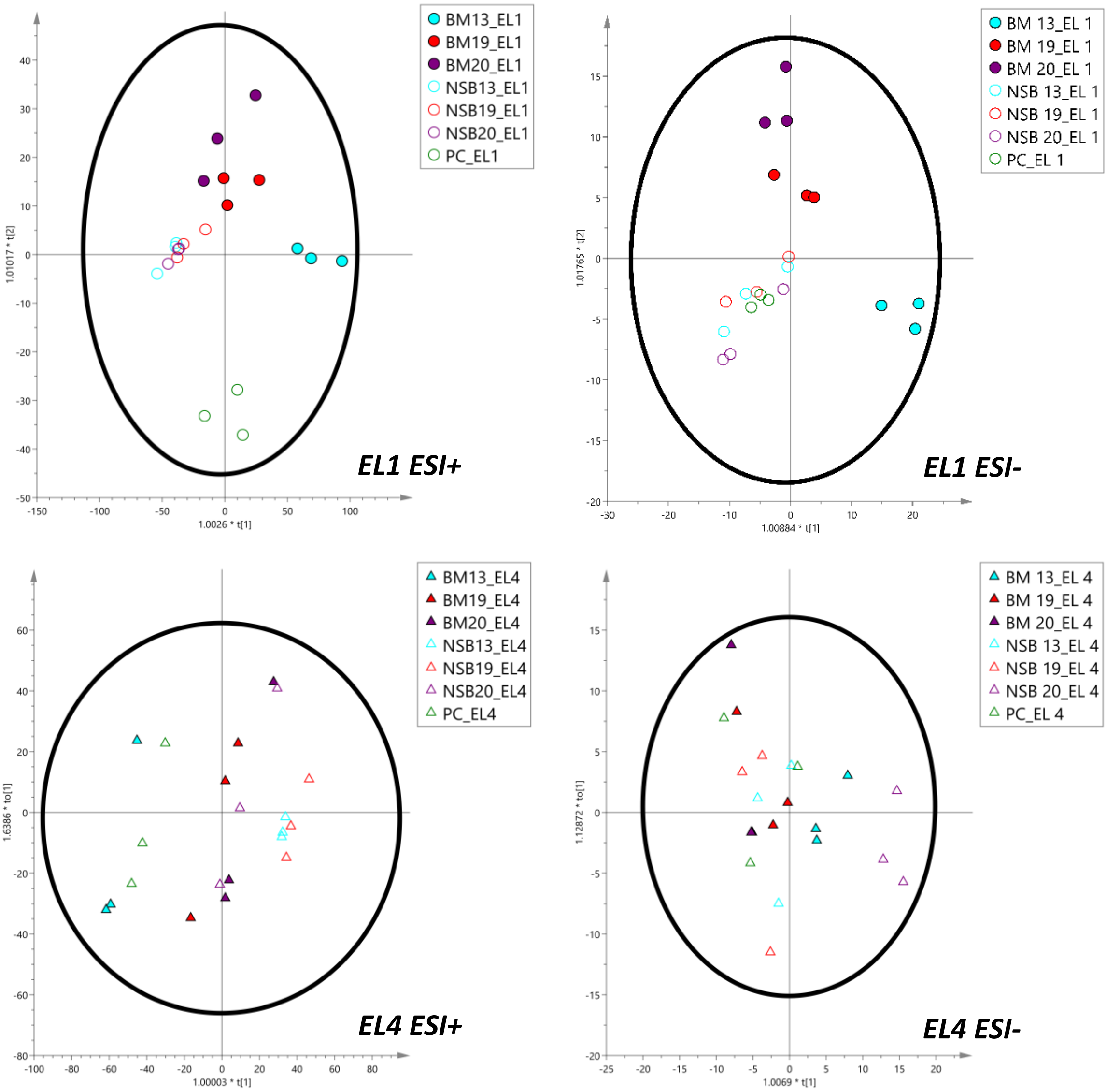

Figure 4: Orthogonal projections to latent structures discriminant analysis (OPLS-DA) score and scatter, for results after instrumental blank, retention time, mass filtration, and low maximum peak area (<2000 AU) for CD results in both ESI$^+$ and ESI$^-$ modes, for elution 1 (EL1; imidazole) and elution 4 (EL4; methanol) (SIMCA 14, MKS Umimetrics, Gottingen Germany). All filled shapes represent sample BM (Biomass) with numbers 13, 19, 20 representing the year of sample. All empty shapes are controls, colour between filled shape and corresponding nonspecific binding (NSB) control are matched for ease of identification. The protein control (PC) data are presented as the green non-filled marker points. Additional OPLS-DAs of all features and elutions 2 and 3 are presented in the SI (Figure S6 and the results from MSDial are in Figure S7).

When all elutions and features are modelled together (Figure S6a, b), distinctions between all MeOH elutions, including controls, are straightforward compared to those of imidazole. The imidazole and the corresponding controls also show separation, which is the strongest for the BM13 imidazole elutions, though with so many factors, the separation distance is not as well defined as when each elution is observed individually. To compare between outputs from the two processing softwares, the same model was also run for the MS Dial output, and similar groupings were observed, indicating that both methods give similar outputs (Figure S7). In some respects, separation is better for MSD, especially for elution 2 and 3, compared to CD. This may be due to the integration thresholds implemented in initial processing, which is also reflected in the far lower number of primary features.

When each biomass sample is investigated individually, the elution replicates in ESI$^+$ tend to cluster together and have a greater distance from PC and NSB controls (Figures S8 and S9). For BM 13 one elution (EL 2) appears as an outlier in many models in both ESI$^+$ and ESI$^-$ for both CD and MSD models, which indicates probable relation to the extraction or sample-handling process. It is also clear, that the MeOH elution (EL4) and the corresponding controls tend to occupy a close space within the OPLS-DA, indicating more similarity between the elutions across the samples than within the same biomass. This observation is, in some respects, ideal as it implies that the imidazole elutions are removing most of the LBD-bound compounds.

Secondary filtration using the ratio of each sample to the maximum PC peak area and the sample ratio to the corresponding NSB (each BM had its own NSB) reduced the number of features to ~2000 for ESI$^+$ and 400 for ESI$^-$ for either CD or MSD. Differing ratios were tried for sample/PC and Sample /NSB (from 2 to 6) with 4 selected as it reduced the number of features to a manageable data set without excluding too much information. Additional filtration criteria were then applied firstly for the triplicates produced per elution per sample: a minimum of two out of the three points for a feature had to be above the PC and NSB ratio threshold. Secondly, the coefficient of variation among replicates above the ratio threshold was less than 30%. A final number of 918 features (366 and 396 features in ESI+ and 109 and 47 in ESI$^-$ for CD and MSDial, respectively), was obtained. The initial detected features were reduced to between 0.2 to 0.7 %.

3.2.3 Features after filtration

Scatter graphs for all 918 final features were generated and assigned a level (1-3) based on the criteria discussed previously (Figure S4). Features were characterized into three levels 1-3, by primarily checking PC and NSB concentrations against the sample and first excluding features where a clear protein elution pattern was observed in the PC. The elution pattern typically observed was either an increase or decrease in peak area from elution 1-4. Also, the spectra of these features exhibited protein fragmentation spectra (26,27). Ideal features (Level 1) were detected in multiple elution replicates and had low or non-detectable control concentrations (Figure 6). There were 324 features retained, those found by both CD and MSD with similar m/z, RT and elution patterns were merged (for example Figure 6 c-d), leaving 306 unique features that were level 1 or 2. Of these 178, and 110, were only from CD, or MSD respectively and 16 were found in both. Comparison of ESI- and ESI+ also highlighted some features in both with the same molecular mass (assuming M+H or M-H) and similar RT, however these were not merged.

For each biomass there were 3 replicates of 4 elutions, and ideally features should be identified across all three replicates, however as this is a preliminary study with limited samples, those found in 2 out of 3 were also retained (Figure S10). For BM13 and 19 in EL 1, 3 and 4 approximately 50% of features were present in all three replicates. EL2 is slightly different with around 1/3 of features in all three elutions. Looking at the features in EL2 many were in EL1 or EL3 at greater peak area so it may reflect the lower detection limits. BM20 is slightly different with around 1/3 of features in all three elutions. In EL 1 the number of features in BM20 (n=68) is similar to BM 19 (n=70), but the number of features is far less in BM20 in subsequent elutions. As the three biomass samples were collected from different locations (Table S1) at different times and contained a differing assemblage of cyanobacterial species, differences in the detected compounds were expected.

Comparing the features across the three biomass samples it is anticipated that there would be some identical features as well as some unique features (Figure 5). There are far greater numbers of unique features in BM13 than in BM 19 and 20, though without a greater data set little can be inferred as to any cause for the difference in the number of features. Those found across all samples however may present ubiquitous RAR features that in a larger study should be included in target assessment and may represent a baseline for RAR binding in any cyanobacterial system. When the features overlap is subdivided between ionization modes and deconvolution software, the pattern of features is largely similar with more than twice the number of features in the ESI$^+$ method (Figure S11).

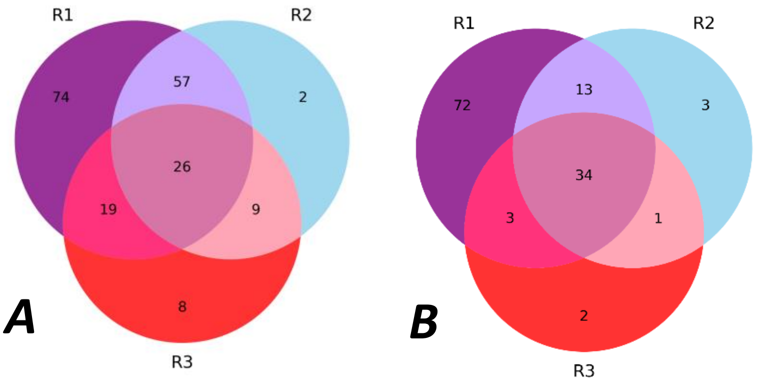

Figure 5: Venn diagram showing the combined ∑ESI$^+$ and ESI$^-$ features from CD (A) and MSDial (B) data across the three biomass samples, after filtration against lab blanks and control samples.

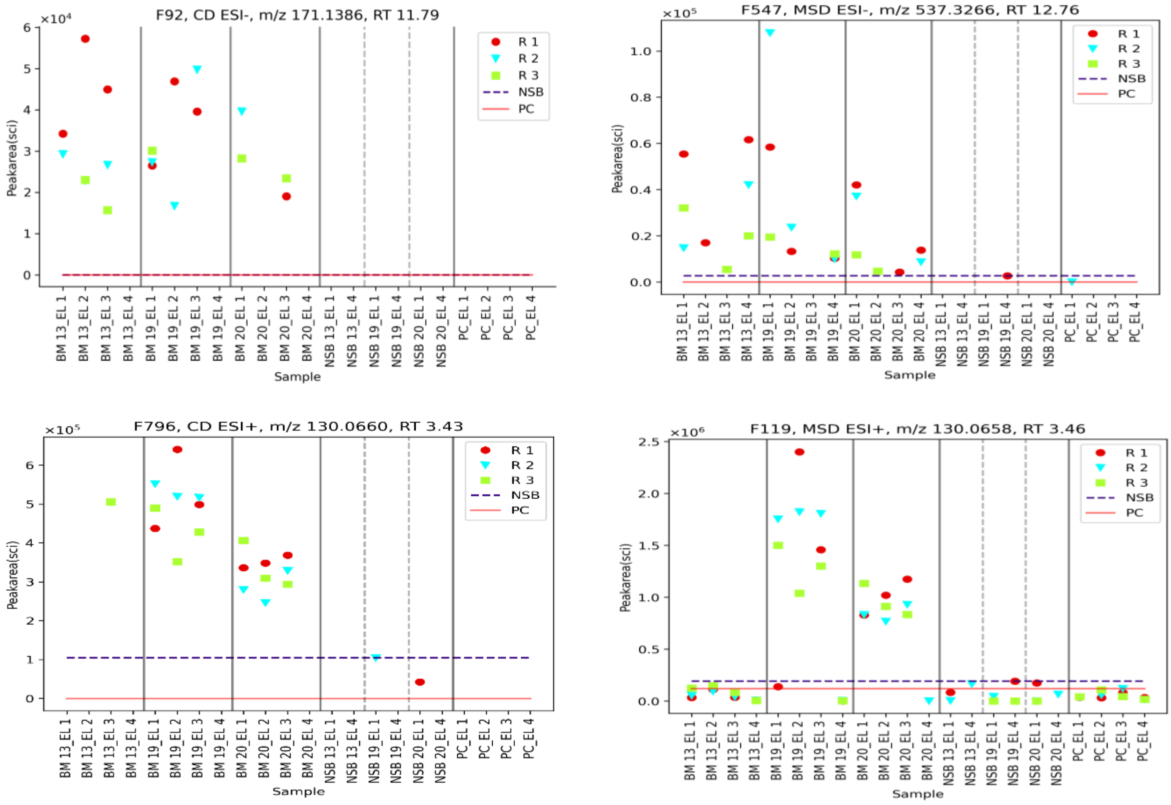

Figure 6: Subset of level 1 scatter graphs for selected features after filtration and checks. The F number in the title refers to the feature number assigned after processing for ease of identification, while CD (Compound Discover) and MSD (MSDial) are the programs used to deconvolute the data files. The m/z are rounded to 4 sf and the retention time (RT) rounded to 2 sf, are presented in each graphs title. The horizontal lines are maximum peak areas for non-specific binding (NSB) and protein (PC) controls. For inclusion in the final scatter graphs each feature in any elution had to be detected in at least two replicates (R) above the maximum peak area of PC or corresponding NSB. Also, within elutions (EL), for each BM an RSD < 50% for at least one elution was required.

3.2.4 Tentative Feature identification

Of the 306 features identified as binding to the LBD of the RAR, 191 features had MS^2 spectra. These were screened through online and inhouse libraries and tentative identification assigned to each feature. The chromatogram scoring was subsequently taken into account and those with a score of >3, indicating a poor peak with high background were excluded leaving 97 features. Consequent database searches led to no prediction for 75 features, 7 were identified as target compounds, 4 were suggested as substructures of other compounds and 3 were similar to known compounds. The remaining 8 were identified only by Chemspider search for MS1, MSDial or CD library with low confidence (Table S6).

The identified features were further manually inspected for their structural probability to bind RAR LBD. The first, presence of carboxy group on the molecules was considered as the necessary functional group of the molecules binding to RAR LBD as shown earlier(28). The second important features was presence of aliphatic chain linked to carboxy group or linkage of carboxy group to cyclohexane by aliphatic chain.

Based on such investigation, we prioritized 4 compounds as the most promising RAR binding molecules – trigonelline, decanoic acid (DA), azelaic acid (AA) and docosahexaenoic acid (DHA). All of them were consequently analysed for their relative potency to bind to RAR compared to ATRA (Figure 7). Results showed that trigonelline and DHA bind to RAR while azelaic acid and decanoic acid did not showed any significant effect. Interestingly, in pull-down-based study Jing et al. (2022) focused on chemicals in dust with ability to bind to RAR LBD and authors also tentatively identified azelaic acid as possible compound that binds to RAR LBD. This information together with no detected effect of azelaic acid in bioassay indicate presence of compound with the same elementary composition but different structure. Further, presence of trigonelline in samples of freshwater blooms biomass is not surprising since this alkaloid is very well known for its presence in many plants, and especially, marine diatoms (29). The trigonelline has been also detected in studies of algae from across the globe, although at relatively low concentrations (29,30). The compound docosahexaenoic acid (DHA) is an omega 3 fatty acid, found all over the human body and essential in fetal development. Importantly, DHA is also intensively studied in microalgae and cyanobacteria for industrial application (31). Therefore, the origin of this compound can be also assigned to the phototrophic members of freshwater blooms. Moreover, the DHA is described endogenous ligand of RXR (32) and since spectra of compounds that can bind both receptors overlay, it is very likely that DHA represent one of these compounds.

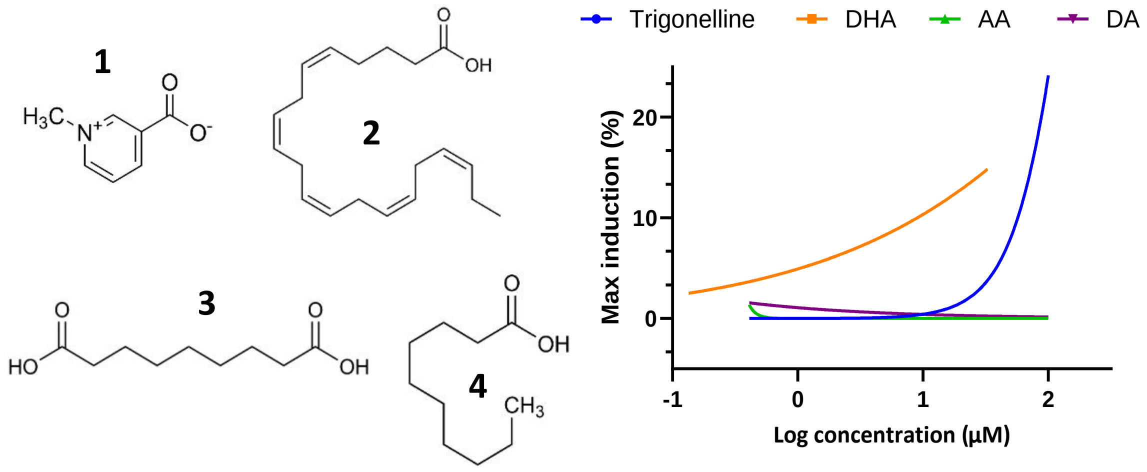

Figure 7. Relative potencies of compounds detected in freshwater bloom biomasses by non-target MS analysis using pull-down assay expressed as percentage of ATRA max induction. 1 – trigonelline, 2 – docohexaenoic acid (DHA), 3 – Azelaic acid (AA), 4 – decanoic acid (DA).

3.2.5 Comparison of CD and MSDial features

The same raw data were processed by two software packages (CD and MSD). For comparison, the primary m/z was rounded to 3 significant figures (sf) and the retention time to 1 sf and the features plotted as a Venn diagram along with plotting all points as scatter graphs ((Figure S12 (ESI+) and Figure S13 (ESI-)). The Venn diagram compared features observed at the three stages of the filtration process, from initial data output to final candidate features. Initially, 0.7 and 1.3 % of features overlapped between MSD and CD in ESI + and ESI- respectively. However, from the scatter graphs (Figure S12a and S13a) a large proportion of features are within the exclusion window of mass and retention time. There are also many features in CD of greater m/z at retention times between 4 and 11 minutes not observed in MSD data and a large proportion of low molecular mass features eluting after 11 minutes and may be noise or in-source fragmentation. The large numbers of features captured by CD processing is likely an artefact of the settings.

In the final data, 6.8 % and 10.7 %, for ESI + and ESI- respectively, fit the criteria of overlapping based on retention time rounded to 1sf and m/z to 3sf. However, an examination of the m/z against the RT (Figure S12 c and S13c) would imply that where the greatest variability occurs is where the retention time is >11 minutes but the m/z is < 200. It is also likely that more features overlapped than were identified by the Venn diagram approach and the simple method used to align likely generated some false positive differences than occurred. Had more surrogate compounds or marker features been identified some corrections between methods could have been applied.

A further comparison is also made between ESI+ and ESI- features by adjusting each to the molecular mass, assuming that the ionization is M+H or M-H and rounding the m/z to 3 sf. There are three features for CD, and two for MS-Dial after final filtration that are in both ionization modes (Figure S14). The two MS-Dial masses also match two of the CD masses. This comparison is helpful as the same feature identified in both modes can provide additional confirmation. A visual inspection of the pattern of features by comparing m/z against time for the final data set (Figure S15) does show that there are potentially more cross over features particularly between 11-12 min, but these were not captured in the Venn models using the filtration settings applied and again the application of retention time correction methods may improve the sensitivity of the method.

4. Future recommendation

Here, the authors would like to recommend using the EnDiTrap pipeline in advanced settings to improve results and facilitate consequent LC-MS analysis. The basic setup represents the default pipeline version applied within this study to show applicability even in basic set-up. The advanced set-up provides suggestions for further improvements that lead to more robust and informative results.

Table 1. Overview of a basic and advanced setup of EnDiTrap pipeline.

	Protein Preparation		
Basic	expression in 2xLB media	purification using resin/column-based affinity chromatography	reaction done in plastic tubes
Advanced	expression in optimized cultivation media	fast protein liquid chromatography (FPLC) protein purification	the reaction can be done in glass vials
	Target analysis		
Basic	3 replicates	no positive control spike in original samples	

Advanced	each replicate is prepared as a mix of 3 replicates	spike positive control in an amount 2.5 -5x higher than is usual LOQ	
		Non-target analysis	
Basic	mass range set from 70-900 m/z	no sample pre-clean step	the original sample analyzed only by target MS
Advanced	reducing the mass range to 600 m/z - improve the sensitivity and prevent noise	use of size exclusion chromatography to remove larger protein fragments	the original sample injected alongside the pull-down sample to confirm features presence and quantifiable results

5. Conclusions

The developed EnDiTrap pipeline represents an emerging approach for the identification of contaminants of concern. Based on well-established methods for bioactivity testing and target analysis of retinoids, retinoic acid receptor alpha from *Homo sapiens* and its native ligands retinoids were used as a model protein-ligand system for EnDiTrap pipeline development. The presented research provides a novel and more informative design of pull-down assay that significantly reduce number of unspecific features detected. Moreover, the developed data processing by Python pipeline provided the opportunity to compare the effect of different software on the outcome and brought interestingly contrasting results pointing out importance of human-driven data curation and processing. Above that, our case study with freshwater blooms biomass confirmed the successful applicability of EnDiTrap pipeline. In targeted analysis, we confirmed majority of previously reported retinoids associated with teratogenic effects of freshwater blooms. In non-targeted analysis, we identified trigonelline and DHA as agonists of RARα LBD. Therefore, the EnDiTrap pipeline is effective tool to study EDCs in environmental samples and can be considered for purposes of eCPIN framework.

Acknowledgement

The work was supported by the Czech Science Foundation Grant No. 20-04676X. G. Codling was funded through Horizon 2020 MSCA project No. 839243 (PullED-MS). Authors thank the RECETOX Research Infrastructure (No LM2023069) financed by the Ministry of Education, Youth and Sports for supportive background. This work was supported from the European Union’s Horizon 2020 research and innovation program under grant agreement No 857560 (CETOCOEN Excellence). This publication reflects only the author's view, and the European Commission is not responsible for any use that may be made of the information it contains.

Conflict of interest

Authors declare no conflict of interest.

Reference

1. Sehnal L, Procházková T, Smutná M, Kohoutek J, Lepšová-Skácelová O, Hilscherová K. Widespread occurrence of retinoids in water bodies associated with cyanobacterial blooms dominated by diverse species. Water Res. 2019 Mar 15;156:136–47.

2. Nováková Z, Novák J, Bittner M, Čupr P, Přibylová P, Kukučka P, et al. Toxicity to bronchial cells and endocrine disruptive potentials of indoor air and dust extracts and their association with multiple chemical classes. J Hazard Mater. 2022 Feb 15;424:127306.

3. Chou PH, Liu TC, Lin YL. Monitoring of xenobiotic ligands for human estrogen receptor and aryl hydrocarbon receptor in industrial wastewater effluents. J Hazard Mater. 2014 Jul 30;277:13–9.

4. La Merrill MA, Vandenberg LN, Smith MT, Goodson W, Browne P, Patisaul HB, et al. Consensus on the key characteristics of endocrine-disrupting chemicals as a basis for hazard identification. Nature Reviews Endocrinology 2019 16:1 [Internet]. 2019 Nov 12 [cited 2023 Jan 20];16(1):45–57. Available from: https://www.nature.com/articles/s41574-019-0273-8

5. Neale PA, O'Brien JW, Glauch L, König M, Krauss M, Mueller JF, et al. Wastewater treatment efficacy evaluated with in vitro bioassays. Water Res X. 2020 Dec 1;9.

6. Fang Y, Chen X, Tan Q, Zhou H, Xu J, Gu Q. Inhibiting Ferroptosis through Disrupting the NCOA4-FTH1 Interaction: A New Mechanism of Action. ACS Cent Sci. 2021 Jun 23;7(6).

7. Jain A, Liu R, Xiang YK, Ha T. Single-molecule pull-down for studying protein interactions. Nat Protoc. 2012 Mar;7(3):445–52.

8. Peng H, Sun J, Alharbi HA, Jones PD, Giesy JP, Wiseman S. Peroxisome Proliferator-Activated Receptor γ is a Sensitive Target for Oil Sands Process-Affected Water: Effects on Adipogenesis and Identification of Ligands. Environ Sci Technol [Internet]. 2016 Jul 19 [cited 2023 Jan 20];50(14):7816–24. Available from: https://pubmed.ncbi.nlm.nih.gov/27340905/

9. Sun Y, Cui H, Li T, Tao S, Hu J, Wan Y. Protein-affinity guided identification of chlorinated paraffin components as ubiquitous chemicals. Environ Int. 2020 Dec 1;145.

10. Jia Y, Zhang H, Hu W, Wang L, Kang Q, Liu J, et al. Discovery of contaminants with antagonistic activity against retinoic acid receptor in house dust. J Hazard Mater. 2022 Mar 15;426:127847.

11. Gong Y, Yang D, Barrett H, Sun J, Peng H. Building the Environmental Chemical-Protein Interaction Network (eCPIN): An Exposome-Wide Strategy for Bioactive Chemical Contaminant Identification. Vol. 57, Environmental Science and Technology. American Chemical Society; 2023. p. 3486–95.

12. Eroglu A, Hruszkewycz DP, Dela Sena C, Narayanasamy S, Riedl KM, Kopec RE, et al. Naturally occurring eccentric cleavage products of provitamin A β-carotene function as antagonists of retinoic acid receptors. Journal of Biological Chemistry. 2012;287(19):15886–95.

13. Bernhofer M, Dallago C, Karl T, Satagopam V, Heinzinger M, Littmann M, et al. PredictProtein - Predicting Protein Structure and Function for 29 Years. Nucleic Acids Res [Internet]. 2021 Jul 7 [cited 2023 Jan 20];49(W1):W535. Available from: /pmc/articles/PMC8265159/

14. Drozdetskiy A, Cole C, Procter J, Barton GJ. JPred4: a protein secondary structure prediction server. Nucleic Acids Res [Internet]. 2015 Jul 1 [cited 2023 Jan 20];43(W1):W389–94. Available from: https://academic.oup.com/nar/article/43/W1/W389/2467870

15. Hon J, Marusiak M, Martinek T, Kunka A, Zendulka J, Bednar D, et al. SoluProt: prediction of soluble protein expression in Escherichia coli. Bioinformatics [Internet]. 2021 Apr 9 [cited 2023 Jan 20];37(1):23–8. Available from: https://academic.oup.com/bioinformatics/article/37/1/23/6070085

16. Käll L, Krogh A, Sonnhammer ELL. Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. Nucleic Acids Res [Internet]. 2007 Jul [cited 2023 Jan 20];35(Web Server issue):W429. Available from: /pmc/articles/PMC1933244/

17. Sehnal L, Smutná M, Bláhová L, Babica P, Šplíchalová P, Hilscherová K. The Origin of Teratogenic Retinoids in Cyanobacteria. Toxins 2022, Vol 14, Page 636 [Internet]. 2022 Sep 15 [cited 2022 Sep 15];14(9):636. Available from: https://www.mdpi.com/2072-6651/14/9/636

18. Adams KJ, Pratt B, Bose N, Dubois LG, st. John-Williams L, Perrott KM, et al. Skyline for Small Molecules: A Unifying Software Package for Quantitative Metabolomics. J Proteome Res [Internet]. 2020 Apr 3;19(4):1447–58. Available from: https://doi.org/10.1021/acs.jproteome.9b00640

19. Schymanski EL, Ruttkies C, Krauss M, Brouard C, Kind T, Dührkop K, et al. Critical Assessment of Small Molecule Identification 2016: automated methods. J Cheminform. 2017 Mar 27;9(1).

20. Koelmel JP, Li X, Stow SM, Sartain MJ, Murali A, Kemperman R, et al. Lipid annotator: Towards accurate annotation in non-targeted liquid chromatography high-resolution tandem mass spectrometry (LC-HRMS/MS) lipidomics using a rapid and user-friendly software. Metabolites. 2020 Mar 1;10(3).

21. Ivanova H, Wagner LE, Tanimura A, Vandermarliere E, Luyten T, Welkenhuyzen K, et al. Bcl-2 and IP3 compete for the ligand-binding domain of IP3Rs modulating Ca2+ signaling output. Cellular and Molecular Life Sciences. 2019 Oct 1;76(19):3843–59.

22. Pipal M, Priebojova J, Koci T, Blahova L, Smutna M, Hilscherova K. Field cyanobacterial blooms producing retinoid compounds cause teratogenicity in zebrafish embryos. Chemosphere [Internet]. 2020;241:125061. Available from: https://doi.org/10.1016/j.chemosphere.2019.125061

23. Javůrek J, Sychrová E, Smutná M, Bittner M, Kohoutek J, Adamovský O, et al. Retinoid compounds associated with water blooms dominated by Microcystis species. Harmful Algae [Internet]. 2015 [cited 2017 Apr 18];47:116–25. Available from: http://www.sciencedirect.com/science/article/pii/S1568988315000992

24. Smutná M, Priebojová J, Večerková J, Hilscherová K. Retinoid-like compounds produced by phytoplankton affect embryonic development of Xenopus laevis. Ecotoxicol Environ Saf. 2017;138(December 2016):32–8.

25. Pípal M, Novák J, Rafajová A, Smutná M, Hilscherová K. Teratogenicity of retinoids detected in surface waters in zebrafish embryos and its predictability by in vitro assays. Aquatic Toxicology. 2022 May 1;246.

26. Schaffer L v, Millikin RJ, Miller RM, Anderson LC, Fellers RT, Ge Y, et al. Identification and Quantification of Proteoforms by Mass Spectrometry. Proteomics [Internet]. 2019 May 1;19(10):1800361. Available from: https://doi.org/10.1002/pmic.201800361

27. Nesvizhskii AI, Aebersold R. Interpretation of Shotgun Proteomic Data. Molecular & Cellular Proteomics [Internet]. 2005 Oct 1;4(10):1419–40. Available from: https://doi.org/10.1074/mcp.R500012-MCP200

28. Borthwick AD, Goncalves MB, Corcoran JPT. Recent advances in the design of RAR α and RAR β agonists as orally bioavailable drugs. A review. Bioorg Med Chem [Internet]. 2020

Oct;28(20):115664. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0968089620304946

29. Dawson HM, Heal KR, Torstensson A, Carlson LT, Ingalls AE, Young JN. Large diversity in nitrogen- And sulfur-containing compatible solute profiles in polar and temperate diatoms. In: Integrative and Comparative Biology. Oxford University Press; 2020. p. 1401–13.

30. Gebser B, Pohnert G. Synchronized regulation of different zwitterionic metabolites in the osmoadaption of phytoplankton. Mar Drugs. 2013;11(6):2168–82.

31. Guedes AC, Amaro HM, Barbosa CR, Pereira RD, Malcata FX. Fatty acid composition of several wild microalgae and cyanobacteria, with a focus on eicosapentaenoic, docosahexaenoic and α-linolenic acids for eventual dietary uses. Food Research International. 2011 Nov;44(9):2721–9.

32. German OL, Monaco S, Agnolazza DL, Rotstein NP, Politi LE. Retinoid X receptor activation is essential for docosahexaenoic acid protection of retina photoreceptors. J Lipid Res. 2013 Aug;54(8):2236–46.