SUPPORTING INFORMATION

TRACING URBAN WASTEWATER CONTAMINANTS INTO THE ATLANTIC OCEAN BY NONTARGET SCREENING

Pablo A. Lara-Martín1*, Aurea C. Chiaia-Hernández2,3, Miriam Biel-Maeso1, Rosa M. Baena-Nogueras1, Juliane Hollender2,4

1Physical Chemistry Department, Faculty of Marine and Environmental Sciences, University of Cadiz, Campus of International Excellence of the Sea (CEI-MAR), 11510 Puerto Real, Spain
2Eawag, Swiss Federal Institute of Aquatic Science and Technology (Eawag), 8600 Dübendorf, Switzerland
3Institute of Geography and Oeschger Centre for Climate Change Research, University of Bern, Bern, 3012 Switzerland
4Institute of Biogeochemistry and Pollutant Dynamics, IBP, ETH Zurich, 8092 Zurich, Switzerland

Corresponding Author: Campus Río San Pedro s/n, Campus de Excelencia Internacional del Mar (CEI-MAR), 11510 Puerto Real, Cádiz, Spain. Tel.: +34 956 016159; fax: +34 956 06040. E-mail address: pablo.lara@uca.es

Number of pages: 14
Number of figures: 7
Number of tables: 4
Number of files: 1
Extended information for Materials and Methods

**Water sampling.** All samples were collected in summer 2015 using 2.5 L amber glass bottles rinsed with solvents (acetone, methanol, and water) and baked at 450 °C prior to sampling. Coastal surface water samples were taken from Cadiz Bay, a semi-enclosed water body located in SW Spain and inhabited by 640,000 people (Fig. 1 in the manuscript). Anthropogenic activities include maritime traffic, shipyards, aquaculture, and agriculture. The mouth of Guadalete River is located at the north of the bay. This river drains a basin of 3.677 km² inhabited by 750,000 people, being the third longest river in Andalusia (165 km length). Within Cadiz Bay, three transects of different lengths (14-20 km) were sampled, one in the river estuary (Guadalete Estuary, G1-10) and two in tidal creeks (Sancti Petri, S1-9, and Rio San Pedro, R1-8). Twenty-four-hour composite wastewater samples (influent and effluent) were also collected on three different weekdays from the WWTP serving Jerez de la Frontera, the most populated city in the area (215,000 inhabitants), and its industrial area. The sewage treatment at this plant consists of a primary physicochemical treatment followed by a secondary biological treatment comprising nitrification and denitrification zones. The plant discharges into Guadalete Estuary at sampling point G1, resulting in the most relevant sewage pollution source in Cadiz Bay (1). The rest of the cities by the bay area discharge treated wastewater into the ocean via marine outfalls.

Surface and bottom (up to 612 m depth) oceanic water samples were collected at the Gulf of Cadiz (NE Atlantic) using Niskin bottles mounted on a Rosette on-board the r/v “Ángeles Alvariño” (Spanish Institute of Oceanography) during the sampling campaign STOCA 201509. Three transects of approximately 50 km length were sampled (Fig. 1): Cadiz Bay coast (CC1-4), Guadalquivir River mouth (GD1-4) and Trafalgar Cape (TF1-5). The Guadalquivir River is the longest river in Andalusia and the fifth in Spain (657 km length), covering an area of 57 071 km² with over 4.5 million inhabitants. Populated cities such as Seville (700,000 inhabitants) and Cordoba (327,000 inhabitants) discharge their sewage here after treatment. The Trafalgar Cape is adjacent to the confluence between the Atlantic Ocean and the Mediterranean Sea, resulting in a water exchange through the Strait of Gibraltar where saltier and denser Mediterranean water flows towards the Atlantic Ocean at deeper levels and less saline and dense surface Atlantic water flows towards the Mediterranean Sea. The interface between both layers corresponds to the isohaline 37.5, located at 100-150 m depth. Here, untreated wastewater is dumped directly into the ocean by small settlements (e.g., Caños de Meca) not connected to the sewer network.

**Sample extraction and mass spectrometry determination.** All the water samples (n = 59) were immediately filtered through previously combusted Whatman glass fiber filters (1 μm pore size) and the filtrate kept at 4 °C until extraction in the laboratory (performed within 24 h after collection). Sample aliquots (100 mL for wastewater, 200 mL for surface water, and 500 mL for oceanic water) were extracted in duplicate using 200 mg solid phase extraction (SPE) Oasis HLB cartridges (Waters Corp., Milford, MA). Ten mL of methanol and 10 mL of HPLC water (Merck, Darmstadt, Germany) were used as solvents to activate the cartridges. The extracts (10 mL of methanol) were evaporated to dryness and redissolved in 1 mL of methanol:water 1:1. A mix of internal standards (n = 24) dissolved in the same solvent was added at 100 µg L⁻¹ (Table S1). Experimental blanks using HPLC water were extracted using the same protocol. This protocol was previously developed by our group for analysis of pharmaceuticals and other polar contaminants (e.g., artificial sweeteners and surfactants) in seawater, showing acceptable recoveries (> 70%) for most target compounds and allowing their determination at parts per trillion (2-4). Further information on SPE conditions and recovery efficiencies for target compounds can be found in these works.

The detection of sewage-derived contaminants was performed by high performance liquid chromatography coupled to high resolution mass spectrometry (HPLC-HRMS) using a Q Exactive hybrid quadrupole-Orbitrap mass spectrometer (Thermo Fisher Scientific, San Jose, CA) equipped
with an electrospray (ESI) source (5). Twenty microliters of sample were injected in a 2.1 × 50 mm × 3.5 µm particle size X-bridge C18 column (Waters Corp., Milford, MA) at 35 °C. Different mixtures containing analytical reference standards (>95% purity) sorted by class were also injected for identification of contaminants (n = 629) (Table S1). The mobile phase consisted of HPLC water (A) and methanol (B), both with formic acid (0.1% v:v), and isopropanol (C). Separation of compounds was achieved in 30 minutes at 200 µL min⁻¹ using the following gradient: 95% A: 5% B between 0 and 1 min, 50% A: 50% B at 4 min, 100% B at 18 min, 100% C at 22 min, and kept at 100% C until the end of the run. Data-dependent (DDA) and data-independent acquisition (DIA) analyses (9 per scan cycle) were performed separately in the positive and negative ionization modes. Full range mass spectra were recorded over a mass range of 100 to 800 m/z with a nominal resolving power of 140,000 referenced to m/z 400 and with a mass accuracy < ±5 ppm. High-resolution product ion spectra were acquired in MS/MS experiments with a nominal resolving power of 17,500. Signal intensities were reported as peak areas previously normalized by considering the areas of internal standards (an average ion suppression percentage was calculated for each sample) and the amount of sample extracted. For additional confirmation of prioritized compounds, samples were reinjected and measured at different energy collision dissociation (HCD) values for targeted MS/MS fragmentation (DDA). Quality control accounted for about 30% of the samples injected, including pure methanol:water 1:1 samples that were used as instrument blanks to control contamination within the instrument and two standard mixtures at 250 µg L⁻¹ and 500 µg L⁻¹ concentrations in vial that were used to control the stability of the instrument during the measuring time in terms of mass shift and signal intensity. Both instrumental blanks and standard mixtures were run before and after a batch of 8 water samples. Sample replicates were measured for all samples. Further information on LC-HRMS instrumental settings, DIA mass isolation windows, and quality control can be found in the ref. 5.

Selection of sewage-derived compounds. A flowchart of the workflow used and the different nodes involved is presented in Fig. 2 in the manuscript. Some elements of this workflow (external nodes) were implemented with the openly accessible statistical software R (version 3.4.3), whereas the rest were part of the HPLC-HRMS vendor software Compound Discoverer 2.1 (Thermo Fisher Scientific, San Jose, CA). The first 5 nodes (from Input Files to Group Unknown Compounds) were aimed to select all the MS features from the HPLC-HRMS files and group them into possible compounds. First, MS and MS/MS spectra of those features presenting chromatographic peaks were selected and aligned among different samples. Compounds were then defined by grouping all the different features such as molecular ions, salt adducts, and isotopes. Possible in-source fragments were not considered here and were discerned manually later on from the component list. Compounds in the experimental blanks having a relative signal intensity greater than 20% compared to that in water samples were removed from the analysis (Mark Background Compounds node, Fig. 2). All the information obtained was finally grouped and presented in a table showing the molecular mass, retention time, relative signal intensity, and peak area of every compound. Some of the most relevant parameters optimized in the nodes used were: Maximum peak width = 0.8 min, Maximum shift in retention time = 0.25 min, Mass tolerance = 5 ppm, Minimum peak intensity = 10,000, Ions detected = [M+H]+1, [M-H]-1, [M+Na]+1, [M+NH4]+1, [2M+H]+1, and [2M-H]-1, Preferred ions = [M+H]+1, [M-H]-1, and [M+Na]+1, and Use Isotope Pattern in Precursor Reevaluation = True. The optimized NTS workflow was capable to detect all the compounds we had standards for (Table S1) prepared in pure solvents (methanol:water 1:1). Internal standards were also detected in all spiked wastewater and seawater samples, as well as selected pharmaceuticals (e.g., carbamazepine) previously measured using target methods (2).

In a second stage, statistical analyses of the dataset were performed, including principal component analysis (PCA) and agglomerative hierarchical cluster analysis (HCA). To define main axes of contaminant variation among the study sites, a PCA on the selected compound intensities was carried out, using the R package “prcomp”. Signal intensities were mean-centered and scaled to agree with...
PCA assumptions. Common compounds from wastewater influent and effluent, G1 and G2 samples \((n = 11,256)\) were selected to create sewage-derived contaminant lists (Table S2). To classify the wastewater compounds into different groups across study sites G1 to G10 (Fig. 1), HCA was performed on the dissimilarity matrix created from these lists using Ward's minimum variance agglomeration method, which is based on variance analysis. We used the functions “daisy” and “hclust” from the R package “cluster” to perform the HCA (5). Those compounds grouped into clusters with progressively decreasing signal intensities from G1 to G10 and also detected in oceanic samples were prioritized and selected for further structural elucidation and identification.

**Identification of prioritized contaminants.** Selected water samples were reanalyzed to perform DDA measurements in the negative and positive ionization modes for those compounds included in Table S2. MS and MS/MS spectra are stored in File S1. Identification of the molecular structure of these substances was carried out through a combination of different nodes (Fig. 2), which were complementary and provided different confidence identification levels (6) between 1 and 4:

**Level 1:** the structure was confirmed via appropriate measurement of a reference standard with MS, MS/MS, and retention time matching (see Table S1 for standards available during analysis). This information was also stored in the node Search Mass Lists.

**Level 2:** a probable structure was proposed by matching MS and MS/MS experimental data with library spectrum data (minimum score was set to \(> 65\%\)). Two different nodes were used to this aim. The first one (Search mzCloud) was included in Compound Discoverer 2.1, whereas the second one involved converting HPLC-HRMS data files into mzXML format using ProteoWizard 3.0 and online search using the R package “RMassBank” (7).

**Level 3:** top-ranked structures from in silico fragmentation of candidates from compound database searches were proposed. Two different mass lists accounting for \(> 15,000\) possible contaminants in the environment (Environment and Food Safety, or EFS, included in Compound Discoverer 2.1, and NORMAN, ref. 8) were used (Table S3), as well as on-line chemical databases (ChemSpider and PubChem). This information, together with retention time, MS and MS/MS spectra, was used to feed the R package “MetFragR” with up to several thousands of possible candidates per component for evaluation (9). The output resulted in lists of chemicals prioritized according to a score based on the linear combination of the number of peaks explained by the algorithm, number of references of candidates in ChemSpider and PubChem, and correlation of retention time with log Kow values of the candidates and reference standards (Fig. S7) in Table S1. An additional in silico fragmentation algorithm included in Compound Discoverer 2.1, FISh (Fragment Ion Search), was used to evaluate the top candidates proposed by MetFrag, providing a score in terms of percentage of theoretical fragments that matched those in the MS/MS spectra.

**Level 4:** no structural information was available, only molecular formula. In this case the MS/MS spectra were either uninformative or could not be acquired, usually caused by too low intensity. Molecular formula was calculated by the node Predict Compositions and those compounds in EFS and NORMAN lists, as well as the top-ranked chemicals in ChemSpider and PubChem in terms of number of references, matching the molecular formula were proposed as possible candidates.

All the relevant information regarding the identification of sewage-derived compounds found in both coastal and oceanic waters \((n = 537)\) is provided in Table S4 for measurements in negative and positive ionization modes. Those compounds detected under both ionization modes are also indicated.
References


8. NORMAN Suspect list exchange (https://www.norman-network.com/nds/SLE/)

Figures:

**Fig. S1.** Principal component analysis (PCA) of normalized signal intensities of all organic compounds detected in water samples from Cadiz Bay (G1-10 = Guadalete Estuary transect, R1-8 = Rio San Pedro transect, and P1-9 = Sancti Petri transect) (A) and the Gulf of Cadiz (GD1-4 = Guadalquivir River mouth transect, CC1-4 = Cadiz Coast transect, and TF1-5 = Trafalgar Cape transect) (B). Proportion of the variance explained by the two main components (1 and 2) is also indicated, being between 9 and 26%.
Fig. S2. Mass vs retention time plots for detected compounds corresponding to different sample types: influent and effluent wastewater (Jerez WWTP), estuarine water (G1 and G10) and oceanic water (CC1 and CC4). Shared sewage-derived compounds are plotted in red.
Fig. S3. Number of sewage-derived compounds detected in water samples from Cadiz Bay (G1-10 = Guadalete Estuary transect, R1-8 = Rio San Pedro transect, and P1-9 = Sancti Petri transect) and Gulf of Cadiz (CC1-4 = Cadiz Coast transect, TF1-5 = Trafalgar Cape transect, and GD1-4 = Guadalquivir River mouth transect).
**Fig. S4.** Examples of contaminants identified at confidence levels 1 (sucralose, an artificial sweetener) (A), 2 (sulfurol, a food additive) (B), and 3 (17-amino-3,6,9,12,15-pentaoxaheptadecan-1-ol, an industrial chemical) (C). Chromatograms showing their presence in water samples and mass spectra containing their molecular ions and isotopic patterns used for molecular formula assignment are presented. Identification is completed by comparison of their MS/MS fragments to those from standards (A), online libraries (B), and in-silico fragmentation algorithms (C). Blue bars in the mass spectra indicate the molecular ions and green bars related isotopes.
Fig. S5. Kendrick mass defect plots showing the occurrence of homologous series (-CH$_2$) for CHNOS sewage-derived contaminants (A) and of polyethylene (PEG) and polypropylene (PPG) glycol derived compounds (-C$_2$H$_4$O and -C$_3$H$_6$O, respectively) (B) in water samples from the Gulf of Cadiz. Colored dots are compounds confirmed to belong within the same series (horizontally aligned for -CH$_2$ and PEG series, and diagonally oriented for PPG series). More information on specific series can be found in Table S4.
Fig. S6. Normalized signal intensities of potential sewage markers (benzotriazole, carbamazepine, atraton, and sucralose) (A) and of the pharmaceutical antipyrine and its main degradation products (B) in Guadalete Estuary (G1-10) and adjacent oceanic waters (Cadiz Coast, CC1-4, S = surface, D = deep).
Fig. S7. Linear correlation between retention time (RT) and log P for all the standards in Table S1.
Tables (enclosed as Excel files):

**Table S1.** List of reference standards for identification of contaminants and internal standards used in this work.

**Table S2.** List of tentative sewage-derived compounds detected by HPLC-ESI-HRMS (both negative and positive ionization modes are covered).

**Table S3.** Environmental and food safety (EFS) (Compound Discoverer 2.1) and NORMAN network (24-05-2017 version, https://www.norman-network.com/) compound databases used in this work for suspect screening.

**Table S4.** Prioritized sewage-derived contaminants identified by HPLC-ESI-HRMS (both negative and positive ionization modes are covered). Possible formulas were ranked (1 = best candidate) by considering the difference between the theoretical and experimental masses and the spectral similarity score between the measured and theoretical isotope pattern (SFit). Identification based on MS/MS spectra was given a score between 0 and 100% for mzCloud library search and FISh, and between 0 and 7 for MetFrag. Homologous series analysis was performed by plotting Kendrick mass defect plots (14,01565 for CH2, 44,02622 for C2H4O, and 58,04187 for C3H6O) and those compounds that were observed to belong to one of these series were assigned to the same series ID number. Compounds detected in both ESI+/− are highlighted in yellow. The total number of compounds is given at the end of the table.

Files (Zip format):

**File S1.** MS and MS/MS spectra of sewage-derived compounds in selected wastewater effluent (Sample_EffluentA and B) and coastal samples (Sample_G1A to G2C). Text file name indicates m/z (m/z), retention time (rt) and collision energy (HCD) for each compound, whereas MS features (molecular ions, isotopes, and adducts) and MS/MS fragments and their abundances are compiled within the file. The PDF files are a graphical representation of the content of the MS/MS text files, showing the MS/MS spectra for each compound.