UAB
Universitat Autònoma de Barcelona

Environmental Proteomics in Wastewater-Based Epidemiology (EP-WBE)

Ester Sánchez Jiménez

Universitat Autònoma de Barcelona

Institute of Environmental Assessment and Water Research
Doctoral Program in Environmental Science and Technology

Doctoral Thesis

Environmental Proteomics in Wastewater-Based Epidemiology (EP-WBE)

Ester Sánchez Jiménez

Thesis supervisor: Montserrat Carrascal Pérez and Damià Barceló Culleres
Academic tutor: Montserrat Sarrá Adroguer

Bellaterra, July 2025

A mi familia, la de sangre y la elegida.

“No permitas que otros te desanimen o te digan que no lo puedes hacer.
A mí me dijeron que las mujeres no podían trabajar en química.
Yo no vi razón alguna por la que no pudiera hacerlo.”

Gertrude Elion (farmacóloga y bioquímica, 1918-1999)

“Puedes empezar en cualquier momento,
aunque te lleve toda una vida llegar a ser bueno.”

Esther Lederberg (microbióloga, 1922-2006)

Agradecimientos

Primero quiero agradecer a mis directores de tesis, Damià Barceló y Montse Carrascal. Gracias Damià por compartir tu conocimiento y por aportar tu experiencia a lo largo de este proceso. Tu acompañamiento ha sido clave para enriquecer y complementar este trabajo, aportando perspectivas que sin duda lo han fortalecido de manera significativa. Gracias Montse por toda tu confianza, ayuda y paciencia durante este largo proceso. Y gracias también por haber decidido hace ya casi 10 años que merecía la pena trabajar conmigo, a pesar de haber tenido que enseñarme todo desde cero, hemos recorrido un largo camino desde entonces. Aunque a veces tengamos desavenencias durante el trabajo por nuestras personalidades diferentes, admiro mucho que siempre pongas por delante la parte humana y personal. No puedo olvidarme de Joaquín y agradecerte por acogerme en el grupo cuando llegué a Barcelona, por los debates científicos con los que una nunca deja de aprender contigo y por dar siempre una visión diferente a los problemas que ayuda a mejorar el trabajo.

La parte de metabolómica de esta tesis no hubiera sido posible sin la ayuda desinteresada de Antoni, siempre dispuesto a leer, corregir y dar su opinión para mejorarlo. Gracias también al ICRA en Girona y a todos los colaboradores de las diferentes depuradoras protagonistas aquí.

Por supuesto agradecer a mis compis de laboratorio: Vanessa, Jaxi, Gustavo, Vir y Tere. Muchísimas gracias por vuestros ánimos, por vuestra ayuda, por los cafés, por las charlas, por las risas, por los congresos, por los cócteles (sin alcohol para mí), por los *retreats* y por mil cosas más. Puede que mis fuerzas hayan flaqueado alguna vez, pero ir a trabajar siempre ha sido un placer.

Aunque ya no esté con ellos, quiero dar las gracias a mis ex-compañeros del IRB por seguir acogiéndome con los brazos abiertos cuando voy. Sobre todo gracias por las meriendas, las sesiones de escaladas y los castellers. Fue un gusto trabajar con vosotros, pero más aún es la amistad de ahora.

Thanks to Sara for taking me under her wing during my time in Davis, for teaching me not only science but also culture, for taking me to visit new places and for the climbing sessions.

Este viaje no hubiera sido posible sin el gran apoyo moral de mis amistades, sobre todo en el último tramo. Ari, gracias por estar ahí siempre, por tu apoyo y tus consejos, por entenderme y guiarme en el proceso. Cristi, gracias por tu presencia en la distancia, por hacerme ver la parte positiva de las cosas, por acogerme en Malmö y enseñarme tu país de acogida. Ana, gracias por nuestra conversación 24/7 desde cualquier parte del mundo, por planear mil y un viajes y por ofrecerte a escribir parte de la tesis (aunque ambas sabemos que solo lo hacías para que pudiéramos irnos a uno de esos viajes). Fonti, Fortu y Maricarmen, gracias por aguantar y acceder a mis caprichos cuando bajo a Málaga, y por llevarme a comer pescaíto o migas. Mariel, gracias por nuestros desayunos y meriendas gordas para poder contarnos la vida durante tres horas seguidas. Y a todos, gracias por vuestra amistad todos estos años y los muchos que espero que queden.

No menos importante ha sido el apoyo incondicional de mi familia, al ser la primera tesis de la familia poco sabían sobre lo que se les venía encima. Gracias por apoyarme en todo, por no dejar que me rindiera, por ofrecerme un refugio en Málaga (o Córdoba) siempre que lo necesitaba, por subir cuando yo no podía bajar y sobre todo por acompañarme en mis locuras. Al final esta tesis es tan vuestra como mía.

El yoga y la lectura (no científica) también han tenido una gran importancia en este proceso. Y aunque sea poco convencional, quiero darme las gracias a mí, por mi perseverancia a pesar de haber querido tirar la toalla más veces de las que puedo contar y por demostrarme a mí misma que era capaz de conseguirlo.

GENERAL INDEX

Figures and tables index

Figures

4.3 FROM SOURCE TO STREAM: EVALUATING WASTEWATER TREATMENT PLANT PERFORMANCE

4.4 NON-TARGET PROFILING OF THE WASTEWATER METABOLOME USING A SUITE OF HRMS TOOLS: A STUDY AVROSS DIVERSE WASTEWATER TREATMENT PLANTS

Tables

3. SAMPLING

4.1 SHOTGUN PROTEOMICS TO CHARACTERIZE WASTEWATER PROTEINS

4.2 SEWAGE PROTEIN INFORMATION MINING: DISCOVERY OF LARGE BIOMOLECULES AS BIOMARKERS OF POPULATION AND INDUSTRIAL ACTIVITIES

4.3 FROM SOURCE TO STREAM: EVALUATING WASTEWATER TREATMENT PLANT PERFORMANCE

4.4 NON-TARGET PROFILING OF THE WASTEWATER METABOLOME USING A SUITE OF HRMS TOOLS: A STUDY ACROSS DIVERSE TREATMENT PLANTS

ABSTRACT

This doctoral thesis addresses the application of environmental proteomics in the field of wastewater-based epidemiology (WBE). WBE involves the analysis of pollutants and biomarkers to obtain qualitative and quantitative data on the activities and health of inhabitants within a given wastewater catchment. After excretion from the human body, biomarkers, together with many other compounds, enter the sewer system. Therefore, sewage integrates a universe of biochemical signals representing the collective biochemical signals of the community it serves. Since its emergence in the early 2000s, WBE has been used as a tool for monitoring population-level substance use, including illicit or therapeutic drugs, pharmaceuticals, personal care products, caffeine, tobacco, alcohol, pesticides, flame retardants, or plasticizers. Most recently it has gained widespread public exposure during the COVID-19 global pandemic, where SARS-CoV-2 monitoring using RT-PCR reaffirmed early-warning capabilities and the potential to reveal hotspots of infection. However, it was not until 2019 that proteins were proposed as complementary biomarkers to small molecules for near-real time, population wide, human biomonitoring of disease. Some protein biomarkers are already approved by the FDA (Food and Drug Administration) mainly for the study of cancer. They are usually found in blood serum or plasma, but two of them are detected in urine, which would expand the molecular epidemiology to provide information on the health of a community through the use of urban water fingerprinting.

This study aims to identify and characterize protein biomarkers in untreated and treated wastewater using environmental proteomics to support population health monitoring, strengthen environmental surveillance, and assess the performance of wastewater treatment plants (WWTPs). Changes in community health and behavior are reflected in sewage protein profiles, which environmental proteomics can effectively analyze within the framework of WBE.

We first developed a protocol to monitor proteins across different water matrices, focusing on both influent and effluent streams at WWTPs. This is the first step for the creation of test devices for the health and environmental monitoring based on proteins. We differentiated between the soluble and particulate fractions, with emphasis in the first group. A key advantage of the method is the use of the whole reviewed protein database to elucidate all potential biological contributors.

Next, we studied influent samples from 10 WWTPs serving different population sizes and industrial activities. This enables a broad characterization of the wastewater proteome and the identification of potential biomarker signatures. Our results showed that the soluble fractions

are rich in Eukaryotic proteins, while bacteria-related proteins are most abundant in the particulate. Among the soluble proteins, amylases and albumins were proposed as markers of human populations and industrial activities, respectively. Thus, proteins present in wastewater carry information about the activities in the catchment areas.

The objective of the treatment plants is the removal of the contaminants so the treated water can be discharged into the environment or used for other applications. While contaminant removal has been extensively studied for small molecules, the fate of macromolecules like proteins remains underexplored. We selected three treatment plants based on demographic and industrial characteristics. Samples from the inlet, the outlet and the receiving waters were analyzed. Findings revealed that most proteins were effectively removed during treatment, with the exception of some recalcitrant proteins, such as human keratins and amylases, or livestock albumin, which are usually the most abundant ones. In the receiving waters even those proteins were absent and the few that were identified do not pose environmental risks.

As mentioned before, small molecules have been typically studied in wastewater. However, most of the studies are carried out in a targeted way for a few interesting compounds. In contrast, we employed a metabolomics approach more similar to the proteomics one, this means using non-targeted methods to broaden the number of possible compounds to annotate. Annotated compounds were classified by physicochemical characteristics and grouped in superclasses, allowing cross-site comparisons for pattern detection. We found that sites with larger human populations present higher levels of long-chain fatty acyls, organoheterocyclic compounds and benzenoids. Notably, these sites also had higher levels of human amylases, reinforcing the potential synergy between metabolomic and proteomic data in environmental surveillance.

To date, most proteomics studies in water matrices have concentrated on specific organisms, mainly in bacteria. These studies represent the first comprehensive effort to elucidate the complete proteome across multiple water matrices, encompassing all possible biological sources. This led to the identification of proteins from livestock and other species for the first time in such contexts. Protein profiles were found across the different catchment areas and we attempted to correlate these with the metabolite's profiles. This is a challenging task that has yet to overcome some obstacles, for example the lack of standardized quantification using internal standards as used for small molecules. This project opens new possibilities for the environmental surveillance using the wastewater proteome. Potential applications include pest monitoring, population size estimation, detection of illegal discharges, and assessment of habit- and health-related biomarkers.

ABBREVIATIONS

5-HIAA	5-hydroxyindoleacetic
APCI	Atmospheric-pressure chemical ionization
bbCID	Broadband collision-induced dissociation
BOD	Biological oxygen demand
BSA	Bovine serum albumin
CE	Capillary electrophoresis
CI	Chemical ionization
CID	Collision-induced dissociation
COD	Chemical oxygen demand
CTS	Chemical translation service
DAA	Differential abundance analysis
DBP	Disinfection byproduct
DOM	Dissolved organic matter
ECD	Electron capture dissociation
EDC	Endocrine disrupting compound
EI	Electron ionization
EM	Electron multiplier
ESI	Electrospray ionization
ETD	Electron transfer dissociation
FASP	Filter-assisted sample preparation
FC	Faraday cup
FDR	False discovery rate
GC	Gas chromatography
GFF	Glass fiber filter
GO	Gene ontology
HCD	Higher energy collisional dissociation
HILIC	Hydrophilic interaction liquid chromatography

HR-LC/MS	Liquid chromatography coupled to high-resolution mass spectrometry
HRMS	High-resolution mass spectrometry
ICP-MS	Inductively coupled plasma mass spectrometry
Ig	Immunoglobulin
ILIS	Isotopic labelled internal standards
InChIKey	International chemical identifier
iTRAQ	Isobaric tags for relative and absolute quantification
kDa	kiloDalton
LC	Liquid chromatography
LC-HRMS/MS	Liquid chromatography coupled to high-resolution tandem mass spectrometry
LFQ	Label-free quantification
MALDI	Matrix-assisted laser desorption/ionization
MoNA	Mass bank of North America
MP	Microchannel plate
MS	Mass spectrometry
MSI	Metabolomics standards initiative
mTRAQ	Amine-modifying tags for relative and absolute quantification
NMR	Nuclear magnetic resonance
NMWL	Nominal molecular weight limit
NOM	Natural organic matter
NSC	Normalized spectral count
PCA	Principal component analysis
PCR	Polymerase chain reaction
PhACs	Pharmaceutically active compounds
PMT	Photomultiplier tube
PSM	Peptide-spectrum match
PTM	Post-translational modification
PVDF	Polyvinylidene fluoride

RPLC	Reversed-phase liquid chromatography
RT-PCR	Reverse transcription polymerase chain reaction
SCIM	Sewage chemical-information mining
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
SIM	Selected ion monitoring
Sm	Summer
Sp	Spring
SPE	Solid-phase extraction
SRM	Selected reaction monitoring
TFA	Trifluoroacetic acid
TMT	Tandem mass tags
TP	Transformation product
UHPLC	Ultra-high performance liquid chromatography
UVPD	Ultraviolet photodissociation
WBE	Wastewater-based epidemiology
Wn	Winter
WWTP	Wastewater treatment plant

1. INTRODUCTION

Wastewater-based epidemiology (WBE)

Wastewater-based epidemiology (WBE) (Figure 1) analyzes pollutants and biomarkers to obtain qualitative and quantitative data on the activity and public health of inhabitants within a given wastewater catchment (Mao et al., 2021). This approach is based on the fact that molecules excreted by humans and animals, mostly in feces and urine, but also in saliva, sputum, mucus, vomit and phlegm (Zahedi et al., 2021), end up in the sewer system. These compounds reflect the entire population and carry information about human lifestyle and health, exposure to environmental pollutants, and industrial activities (Daughton, 2018; Rice & Kasprzyk-Hordern, 2019; Devault & Karolak, 2020).

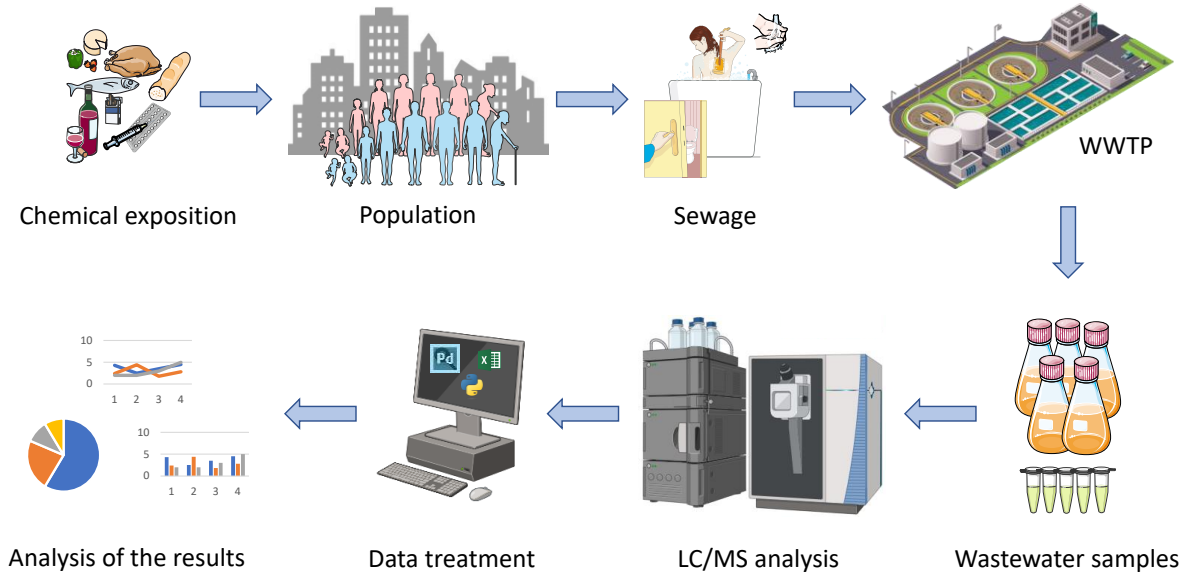


Figure 1. Wastewater-based epidemiology workflow.

Background and evolution of WBE

Every day people all over the world are exposed to a growing number of environmental pollutants. Pollution is considered a major environmental cause of illness and mortality in modern times. The primary sources of potentially hazardous chemicals, in addition to tainted air, food, and water, include a wide range of everyday items such as cosmetics, medications, food packaging, plastics, furniture, electronics, paints, lubricants, adhesives, and many more (Senta et al., 2020). Human biomonitoring (HBM) studies (Husøy et al., 2019) include the analysis of particular biomarkers in biological matrices from individuals in order to determine exposure to these substances. High costs, selection bias (the challenge of choosing individuals who are representative of the entire population), ethical approval requirements, and a lack of temporal dimension (individuals are sampled only once or occasionally over a 24-hour period) are some of the drawbacks of this approach. Because of this, it is challenging

to track temporal trends in contaminant exposure and extrapolate the findings of HBM studies to the entire population (Senta et al., 2020).

For these reasons, WBE has gained traction in recent decades. The first ever work presenting ideas for analyzing wastewater for illicit drug usage appeared in the 1970s (Hignite & Azarnoff, 1977). However, the idea of determining the consumption of illicit drugs on the basis of analysis of wastewater was introduced by Daughton and Ternes in 1999 (Daughton & Ternes, 1999) and it particularly garnered attention since 2001 due to using wastewater as a tool for tracking drug usage in communities (Daughton, 2001). Since the early 2000s, WBE has been applied to monitor population-level substance use, and has most recently gained widespread public exposure during the COVID-19 global pandemic, where SARS-CoV-2 wastewater monitoring reaffirmed early-warning capabilities and the potential to reveal hotspots of infection (Figure 2). These applications highlight the inclusive, minimally invasive and cost-effective benefits of WBE (Bowes et al., 2024). It is also a near real-time tool since the analysis can usually be carried out in the following 24 hours to 7 days after the sampling.

The optimization of the resources for preventing, avoiding, controlling, or mitigating human exposure risks, as well as for maintaining or promoting health, depends on timely evaluation of the overall health of small-area human populations. One new strategy is the ongoing monitoring of sewage for chemicals that indicate the general state of human health or any other long-term trends in community health. Initially termed Sewage Chemical-Information Mining (SCIM), this method monitors anthropogenic and natural chemicals entering sewers as a result of human activities, daily actions, and behaviors (Daughton, 2012; Daughton, 2018). SCIM takes advantage of the constant availability of raw or untreated sewage water, which, besides, does not require institutional review boards approvals since it is essentially anonymous. Therefore, SCIM is applied to the group of individuals served by a specific treatment facility. While the term “community” is used, its definition would vary through time, which could become a disadvantage for the interpretation of the data (Daughton, 2018). BioSCIM is a derived concept that is specifically used for continuously measuring the state or time-trends in community-wide health. While, it uses of the same matrix, the difference relies in the specific study of biomarkers related to human disease, stress, or health. This data could help to compare different populations, arising possible correlations between exposures and diseases (Daughton, 2012).

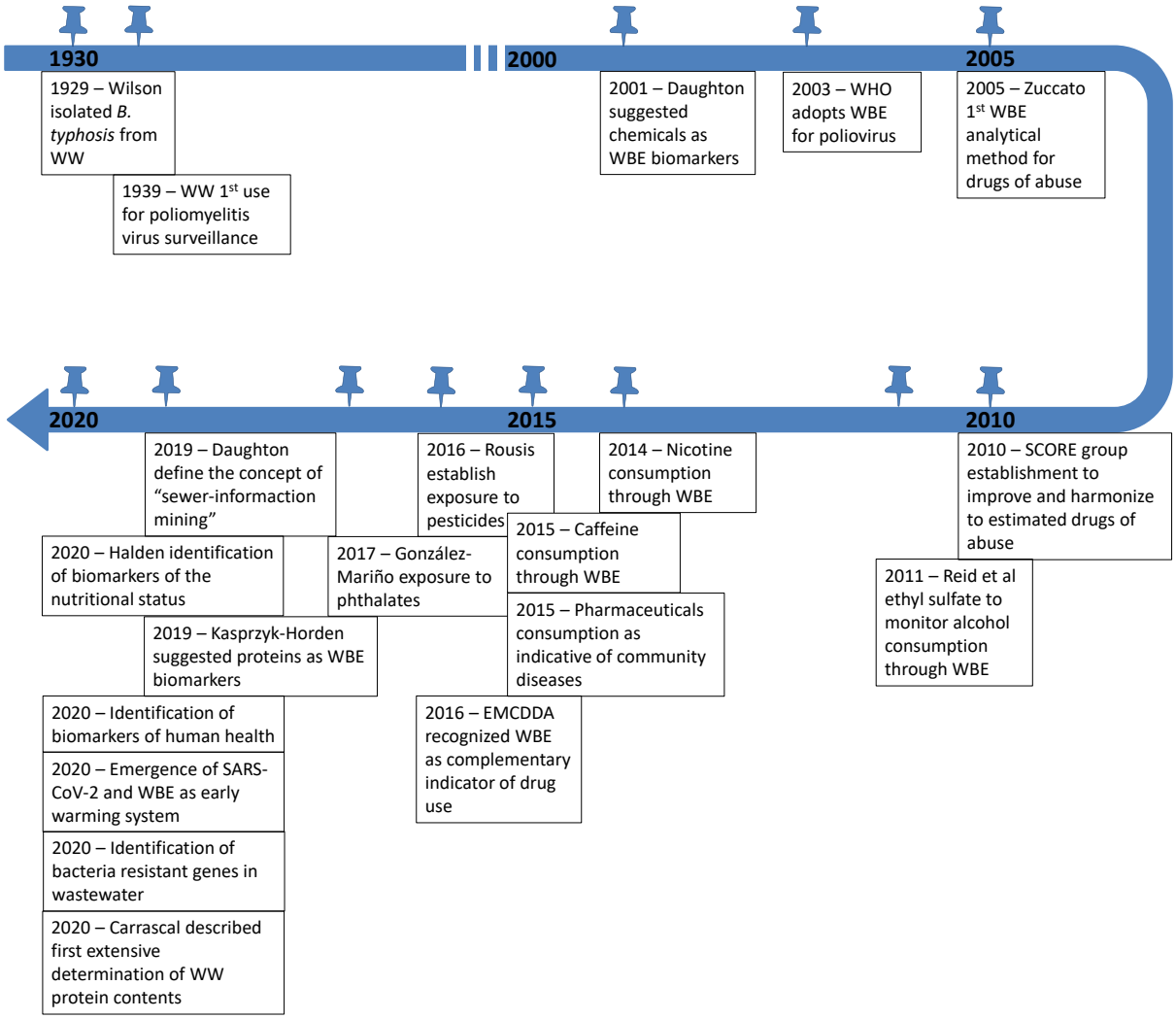


Figure 2. Chronological steps of wastewater-based epidemiology. Figure adapted from Picó et al., 2024.

WBE has been widely applied to chemical compounds to understand the use and misuse of illicit or therapeutic drugs within a community (Mastroianni et al., 2017; Boleda et al., 2009; Daughton, 2001; Tscharke et al., 2016; Zuccato et al., 2008), pharmaceuticals including antidepressants (Baker et al., 2014; Sheng et al., 2014), antibiotics (Subedi et al., 2017) and personal care products (Burgard et al., 2013; Gao et al., 2016; Nguyen et al., 2018), caffeine (Gracia-Lor et al., 2017), tobacco (Castiglioni et al., 2014; Mackuľak et al., 2015) and alcohol use (Ryu et al., 2016a), pesticide exposure (Rousis et al., 2017; Rousis et al., 2020), flame retardants (O'Brien et al., 2015), plasticizers (González-Mariño et al., 2017), pathogen (Choi et al., 2018; Fioretti et al., 2017), genetic biomarkers (Ahmed et al., 2020) and, recently, SARS-CoV-2 virus detection (Barcelo, 2020).

Composition and treatment of wastewater

The raw or untreated wastewater contains dissolved organic matter (DOM), which includes proteins, carbohydrates, fats, oils, and a range of chemicals such as surfactants and detergents, personal care products, pharmaceuticals, pesticides or industrial chemicals (Shon et al., 2006). Along with those, wastewater also contains natural organic matter (NOM), which comes from the breakdown and degradation of organisms, vegetables, and soil. The number and diversity of the compounds is increasing as organic molecules are continuously synthesized and introduced into the market. Additionally, there is a wide range of transformation products (TPs), including disinfection byproducts (DBPs) (Hertkorn et al., 2013).

Water quality is defined as the degree to which water is clean and whether it is suitable for a particular purpose such as drinking or disposal to the environment. It is often understood in a limited manner as a set of standards against which regulatory compliance can be assessed, usually achieved through water treatment. There are different treatments for the removal of pathogens and chemicals from wastewater, which can be used individually or in combination. In general, these treatments are classified in physical (sedimentation, filtration, inactivation by solar or UV radiation), biological (activated sludge, algae) and chemical (coagulation-flocculation, inactivation by oxidants such as chlorine) (Zahedi et al., 2021) (Figure 3). Urban WWTPs are generally constructed with an approach that depends on the catchment areas. This makes every WWTP optimized for different types of compounds, although the final objective is in all the cases to allow human and industrial effluents to be disposed of without danger to human health or natural environment (Lindblom & Samuelsson, 2022). The most common WWTP design includes primary treatment, biological secondary treatment, sand filtration and sometimes an advanced treatment with phosphorus and nitrogen removal. In some situations, the wastewater treated effluent may be disinfected prior to discharge to a natural environment (Zhao et al., 2023). While this general approach is accepted to fulfill guidelines and regulations in stable conditions, it is necessary to understand if the performance of the plants is still satisfactory under challenging situations as the ones engendered by global change with the rise of temperatures, more torrential rains or longer periods of drought.

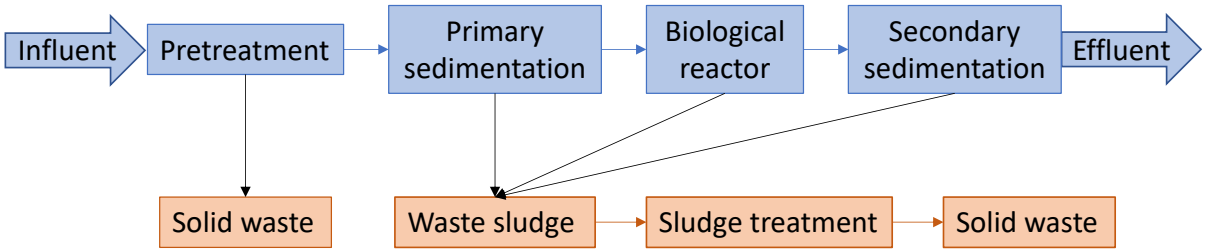


Figure 3. General overview of a wastewater treatment plant (WWTP).

Wastewater treatment efficiency and the quality of wastewater treated effluents is evaluated, on the routine basis through measurements of parameters such as the chemical and biological oxygen demand (COD and BOD, respectively), total suspended solids, or total nitrogen and phosphorus (*Directive - 91/271 - EN - EUR-Lex*, n.d.). Some additional specialized tools, such as the measurement of the aromaticity using the specific UV absorbance, especially relevant in on-line analytical techniques, or specific methods to analyze a small number of hand-picked contaminants, typically using gas or liquid chromatography coupled to mass spectrometry, are also employed to meet with governmental environmental standards (*Directive - 2013/39 - EN - EUR-Lex*, n.d.).

Proteins as biomarkers in wastewater

In 2019, Rice & Kasprzyk-Hordern proposed using proteins as biomarkers in wastewater alongside small molecules, suggesting that combining WBE with clinical proteomics could enable near-real-time, population-wide human disease monitoring. Many of the protein biomarkers approved by the FDA (Food and Drug Administration) are used in oncology. Most of them are present in blood serum or plasma, but a couple of them are in urine. This is of the most interest as it would allow the expansion of molecular epidemiology towards WBE, since urban water is considered a pooled sample of the population, being the urine a high proportion of this sample (Rice & Kasprzyk-Hordern, 2019). This would lead to the proteomic monitoring of the population using an anonymous and representative sample.

Rice & Kasprzyk-Hordern (2019) proposed five criteria (modified from Daughton (2012)) for selecting protein biomarkers: excreted in urine, specific for a disease, available biomarker-disease information, stability and high urinary concentration. Using these criteria, the authors suggested the following biomarkers from the literature as initial targets of interest: prostate specific antigen (PSA), C-reactive protein (CRP), interleukin-6 (IL-6), interleukin-8 (IL-8), podocin (PDC), anterior gradient protein 2 (AGR2) and uromodulin (URM). They are all present in urine and the relationship biomarker-disease have already studied.

Nevertheless, it was not until 2020 that the first experimental study of proteins within wastewater was made by Carrascal et al. (2020). These authors were the first to successfully assess the wastewater proteome and, as a result, a significant number of human proteins were detected in this media. Using polymeric devices with polycaprolactonediol homopolymer cap units placed for 11 days in the influent water, the anoxic reactor (denitrification) and the effluent water, they reported the identification of 690 proteins from bacteria, plants and animals, including humans. The most represented taxonomy was bacteria with proteins like elongation factor Tu (EF-Tu), 60 kDa chaperonin (GroEL) and ATP synthase. However, the species with the highest contribution to the total number of proteins was *Homo sapiens*. Among the human proteins they identified uromodulin (Garimella & Sarnak, 2016), already proposed by Rice & Kasprzyk-Hordern (2019), and α-amylase (Mattes et al., 2014), and S100A8 (Wang et al., 2018), both of them previously proposed as human health markers. Furthermore, they concluded that the proteome profile was changing along the different sites; the influent site was dominated by human proteins, some of them persisting in the anoxic reactor, while the effluent was dominated by bacterial proteins presumably from the sludge (Carrascal et al., 2020).

Environmental sciences

The environment continually receives chemicals from agricultural production, industrial processes, and other human activities. This leads to the presence of complex mixtures of human-made chemicals and their transformation products in these spaces without knowing their effects on organisms and ecosystems (Nesatyy & Suter, 2007). In general, to assess the pollution status of a specific ecosystem, it is necessary to chemically analyze soil samples, water, sediment, and biota (Kolpin et al., 2002). However, there are some challenges to do this process: (i) the wide variety of chemicals, with different physicochemical properties, (ii) the low effect levels, concentration additivity or synergism, and (iii) the weather variability which influence the concentrations. The study of the biota present in an ecosystem is especially important because the effects on aquatic organisms are driven by internal concentrations of the chemicals bioavailable. For example, lipophilic compounds tend to accumulate in soil and lower organisms, leading to significant bioconcentration in top predators of the food chain (van Lipzig et al., 2005). The analysis of an organism's proteome has arisen as a strategy to overcome these challenges. It would be possible to detect subtle changes in some proteins levels as a response to those environmental contaminants. This way new biomarkers of exposure could be discovered, helping to unmask mechanisms of toxicity. This is the principle of environmental proteomics. Over the years, this science has

investigated many organisms (microorganisms, plants, invertebrates, vertebrates) (Nesatyy & Suter, 2007).

The investigation of a proteome allows the association of individual proteins or groups of proteins with disease or toxicity, making them useful as biomarkers. Therefore, the identification of proteins that are induced or suppressed allows to draw conclusions regarding the molecular mechanisms underlying stress response. This approach represented a paradigm shift in molecular biology as it was focused on generating an overview of the proteome (both qualitative and quantitative), instead of focusing on a single protein or protein family. In this way, the information could give a full understanding of an organism, especially when completing with genomic, metabolomic or histopathology data. Moreover, the focus of environmental proteomics can range from global to targeted protein analysis. In the first case the goal is to identify as much proteins as possible in the whole organism, while in the second case a sub-proteome (e.g. organelle or pathway) is elected to be studied (Nesatyy & Suter, 2007).

Analytical tools for wastewater analysis

The advances in environmental sciences and more specifically in the analysis of wastewater are closely related to the advances of the available tools. Most biological samples mainly consist of a highly complex mixture of polar and non-polar biomolecules, which are required to be separated prior to their analysis. Common separation techniques include gas chromatography (GC) and liquid chromatography (LC) (Figure 4) (Noor et al., 2020). GC is based on the separation of the molecules while they move through a column filled with a gaseous mobile phase (called carrier gas) with a temperature gradient, which will affect the volatility of the molecules. This technique is mostly used for identification and quantification of volatile non-polar or slightly polar compounds, which makes it a good option for the analysis of some small molecules (Schauer et al., 2005). On the other hand, LC or high-performance liquid chromatography (HPLC) is used for a larger range of molecules, including non-volatile compounds and polar molecules, which it is suitable for small molecules, peptides and proteins. In both cases, chromatographs can be coupled to a mass spectrometer, which will analyze the compounds as they are released.

Mass spectrometry (MS) is an analytical technique used for the identification and quantification of sample analytes in a gaseous state based upon their mass-to-charge ratio (m/z) in a vacuum environment. Mass spectrometers have the following parts: a source for the ionization, one or more mass analyzers, and a detector (Figure 5).

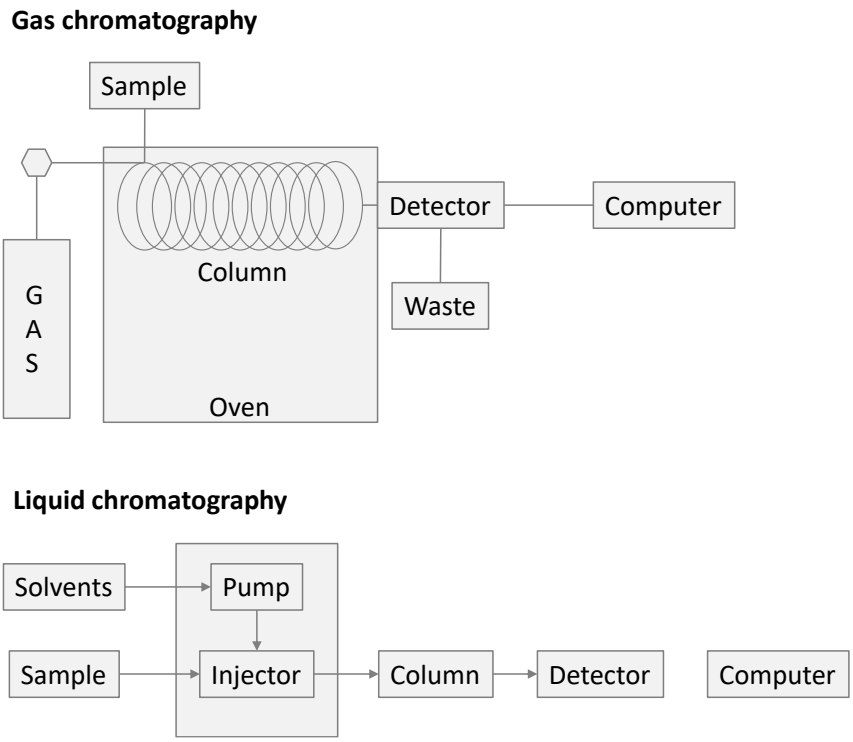


Figure 4. General schematic for gas and liquid chromatography.

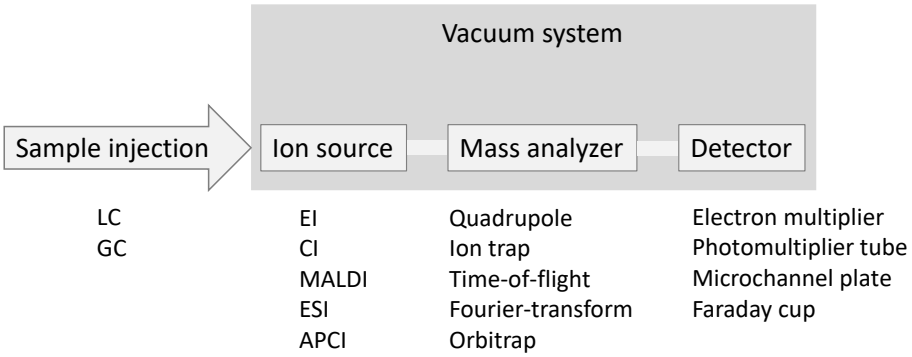


Figure 5. General schematic of a mass spectrometer. Figure adapted from Jiang et al., (2024).

As a gaseous state of the sample is required, ionization is necessary to pass from liquid to gas. This occurs in the source and is performed pumping the liquid sample through a high voltage (1-4 kV) tip, where the stream is broken off in miniscule charged droplets and producing mono-charged ($[M]^+$, $[M+H]^+$) or multiple-charged ($[M+nH]^{n+}$) ions (Sinha & Mann, 2020). Depending on the sample type and target molecules, there are different types of ionization: electron ionization (EI) (Grilo & Mantalaris, 2019), chemical ionization (CI) (Bradbury & Plückthun, 2015), matrix-assisted laser desorption/ionization (MALDI) (Shah & Maghsoudlou, 2016), electrospray ionization (ESI) (Huang et al., 2012) or atmospheric-

pressure chemical ionization (APCI) (Zhang et al., 2010). Among all the available ionization methods, ESI and MALDI are currently widely favored for the ionization of peptides and proteins because they are soft techniques and have high sensitivity for the types of molecules.

The mass analyzer is the main part of a mass spectrometer. It carries out the separation of ions according to their mass-to-charge ratio using an electrical or magnetic field. In order for the ions to go through the mass analyzer without intermission but with efficiency, the process occurs in a vacuum environment at low (10^{-3} to 10^{-5} Torr) or very low (10^{-7} to 10^{-10} Torr) pressures. The most common types of mass analyzers include quadrupole (Q), ion trap, time-of-flight (ToF), Fourier-transform (FT), and orbitrap. These mass analyzers can be used independently or combined. Some of the most frequent combination are quadrupole-time-of-flight (Q-TOF), quadruple-Orbitrap (Q-Orbitrap), quadrupole-Orbitrap-linear ion trap (Q-Orbitrap-LIT), triple quadrupole (QqQ), and time of flight/time of flight (ToF/ToF) (Noor et al., 2020; Peters-Clarke et al., 2024). Quadrupoles and ion traps can isolate ion populations; however, ion traps can also accumulate these ions. TOF devices accelerate the ions through an electric field, and are based on the inverse relationship between the arrival time and the m/z (the faster the ion arrives to the detector, the lower the m/z, and vice versa). These three mass analyzers are usually used for small molecules. In proteomics, orbitrap is the main mass analyzer of choice. Orbitraps work as a trap, where the motion of the trapped ions is measured in frequency and then Fourier-transformed to m/z (Peters-Clarke et al., 2024).

Fragmentation is the breaking of precursors or parent ions from the ion source into smaller products or fragment/daughter ions. For the fragmentation to occur additional energy is provided to the precursors in a collision cell (Noor et al., 2020). There are several techniques of fragmentation, the most common are: collision-induced dissociation (CID), electron capture dissociation (ECD), electron transfer dissociation (ETD) and higher energy collisional dissociation (HCD). These approaches differentiate in the activation method: collisional for CID and HCD, and electron-based for ECD and HTD (Peters-Clarke et al., 2024). Nowadays, HCD is built in the majority of the mass spectrometer and, therefore, is the fragmentation technique used in many studies.

As a final step, the ions are sent to the detector. Here the m/z values and abundances are measured, however the quantity of ions is quite low, so an additional step is needed. This is the amplification of the signals using different multipliers. Some detectors are electron multiplier (EM), photomultiplier tube (PMT), microchannel plate (MP), and Faraday cup (FC); all of them are based on the counting of the ions (Jiang et al., 2024). Electron multipliers, including multichannel plate detectors, are commonly used detectors to identify proteins and

peptides and commercially used with quadrupole, ion trap and time of flight analyzers (Liu et al., 2014). However, orbitraps do not used any of the detectors mentioned before. In this case, detectors measure a voltage proportional to the current, from which frequency is extracted and then Fourier-transformed into m/z (Jiang et al., 2024).

Mass spectrometry for small molecules

Recent advances in water chemistry have improved our knowledge about the genesis, composition, and structure of DOM. High-resolution mass spectrometry (HRMS) is currently the election technology for its analysis. When coupled with nano-chromatography, HRMS produce consistent results (Hawkes et al., 2016) allowing both direct infusion and chromatographic separation. Research employing HRMS technology has investigated the changes in DOM across wastewater treatment plants (WWTPs), advanced wastewater treatment plants (AWTPs) and drinking water treatment plants (DWTPs) (Sanchís et al., 2021; Gonsior et al., 2014; Lavonen et al., 2015; Phungsai et al., 2016; Maizel & Remucal, 2017), proving the suitability of this tool with the defined objective of performance assessment.

The polar/ionic analytes require specific chromatographic separation. Separation modes are usually reversed phase (RP) chromatography and hydrophilic interaction liquid chromatography (HILIC). In RP, a hydrophobic stationary phase is used for retention and a mobile phase consisting of a mixture of organic modifier and a water phase for elution. In HILIC, the analytical column is polar combined with a highly organic mobile phase in which water is introduced as the eluting solvent. The most commonly reported ionization source is electrospray ionization (ESI). Triple quadrupole mass analyzers are used the most and considered the reference technique to quantify illicit drugs and their metabolites. This technique is both sensitive and suitable for quantitative analysis, and selective for identification, and an excellent tool for targeted analysis (Huizer et al., 2021).

The use of LC-MS/MS allows the detection of illicit drugs at the ppb-ppt. In this case, LC-MS/MS identification is aided by preconcentration steps used to pull out trace amounts of drugs from dilute solutions. Techniques such as solid phase extraction (SPE) use chemical attractions such as lipophilic and/or ion exchange interactions to extract analytes from up to a liter of solution or more onto resins that can then be eluted in small volumes of solvent, effectively concentrating analytes by 2-4 orders of magnitude (Burgard et al., 2013).

One of the last reviews about the techniques used for the study of small molecules was made by Senta et al. (2020). The authors resumed the analytical methods used for the determination

of human biomarkers of parabens, UV filters, phthalates, phosphorus flame retardants (PFRs) and bisphenol A (BPA) in WBE and similar studies. Except for two GC-MS methods for PFRs, all other methods were based on LC-MS/MS, which is nowadays the most common technique for the analysis of polar environmental contaminants. In most of these methods, electrospray ionization (ESI) and multiple reaction monitoring (MRM) mode on triple quadrupole (QqQ) or quadrupole-ion trap (QTRAP) mass spectrometers were employed for the detection and quantification of the target compounds. Q/ToF was also used in several configurations depending on the studied compounds. Except for one method on UV filters, which employed direct injection, all other methods included sample treatment step using SPE. Polymeric reversed-phase and mixed-mode strong cation-exchange sorbents (Oasis HLB and MCX, respectively) were used in most methods, however other types of sorbents, including mixed-mode strong and weak anion exchange sorbents (Oasis MAX and WAX, respectively), as well as hydrophobic, end capped silica phase (Bond-Elut C18), were also employed.

Mass spectrometry for proteomics

In standard proteomics, RPLC is the most used LC (Zhang et al., 2010) for peptide and protein separation. At MS level, the proteome can be investigated either with top-down analysis of the intact proteins or by focusing on their digested peptides (bottom-up) (Figure 6). These techniques are complementary because none is able to provide complete information about a protein of interest on its own (Nesatyy & Suter, 2007; Rice & Kasprzyk-Hordern, 2019).

As showed in 6, the workflow for bottom-up proteomics is composed of the following steps: protein extraction from cells or tissues, digestion of these proteins into peptides, analysis of the peptide extract by LC-MS/MS, and lastly peptides are identified from the MS/MS spectra and proteins inferred from the peptides identifications with the assistance of a protein database. There are different versions of each steps depending on the purpose of the experiment (Jiang et al., 2024). Some of the drawback of this approach are incomplete proteolytic digestion, peptides non-suitable for ionization, non-useful fragmentation spectra, or lack of sequence information in databases. This can limit the ability of bottom-up proteomics to examine issues that are important for biological functions, such as PTMs or site-specific mutations of individual proteins. On the contrary, top-down proteomics measures intact proteins following a similar but different workflow: proteins are extracted from cells or tissues and directly analyzed by LC-MS/MS, afterwards spectra are identified by comparing with databases (Figure 6). One advantage of this approach is the capacity of identifying proteoforms, but the protein coverage is lower than in bottom-up (Jiang et al., 2024). Other limitations of the top-down approach are: low sensitivity due to the high number of charges and large m/z of intact proteins that can be out of the mass spectrometer range, less

separation resolution which leads to coelution complicating MS spectra and decreasing ion intensities (Cupp-Sutton & Wu, 2020), the need of higher activation methods such as ETD, ECD or UVPD which are not available in all the instruments (Lanzillotti & Brodbelt, 2023), low solubility of certain proteins and the lack of robust, user-friendly data processing tools (Melby et al., 2021; Brodbelt, 2022). The selection of the approach will depend on the study's objective, however top-down has not been used in wastewater so far.

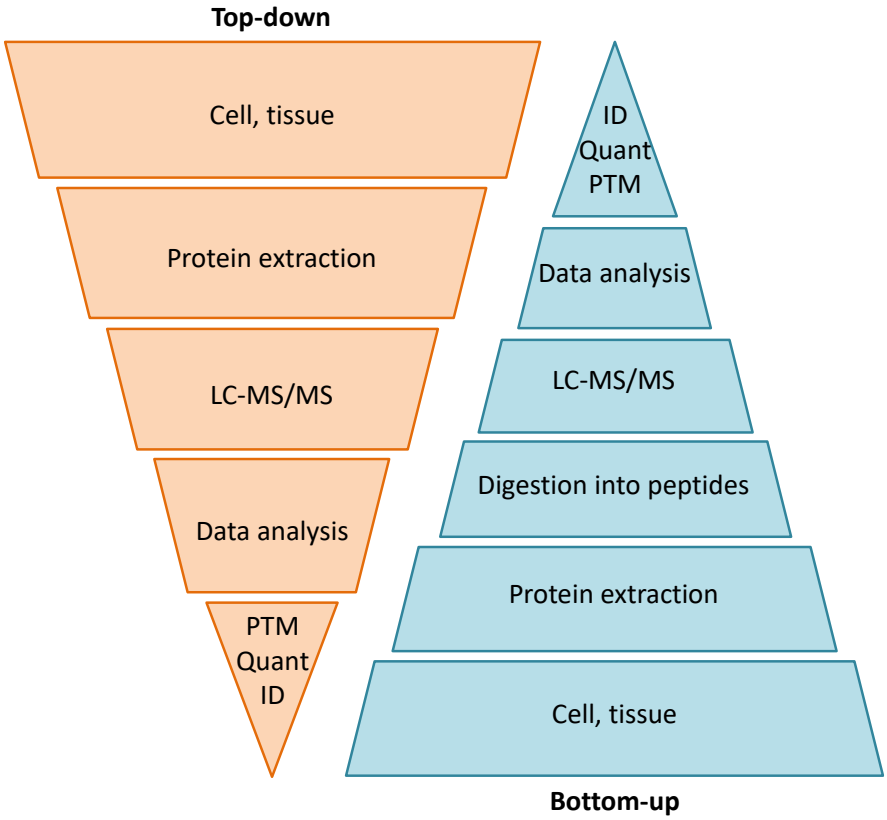


Figure 6. Top-down and bottom-up proteomics workflows.

ESI is the ionization methods of choice for the identification, quantitation and characterization of a protein. After ionization, there are two acquisition strategies: data-dependent acquisition (DDA) and data-independent acquisition (DIA) (Figure 7). In DDA, an MS1 is acquired iteratively, and from that spectra some signals are chosen (based mostly on the charge and the intensity) to be fragmented and measured an MS2. The drawback is that the analysis are not always reproducible as the chosen signals for MS2 can be different each time even if the sample is the same, moreover low abundance signals are less likely to be chosen. In DIA, the MS2 are independent of the charge and intensity of the MS1 signals. In this approach, the mass range is divided in sections and each section subjected to fragmentation. Therefore, each MS2 will be a mixed of different signals. This process is repeated in each cycle time (Jiang et al., 2024; Peters-Clarke et al., 2024).

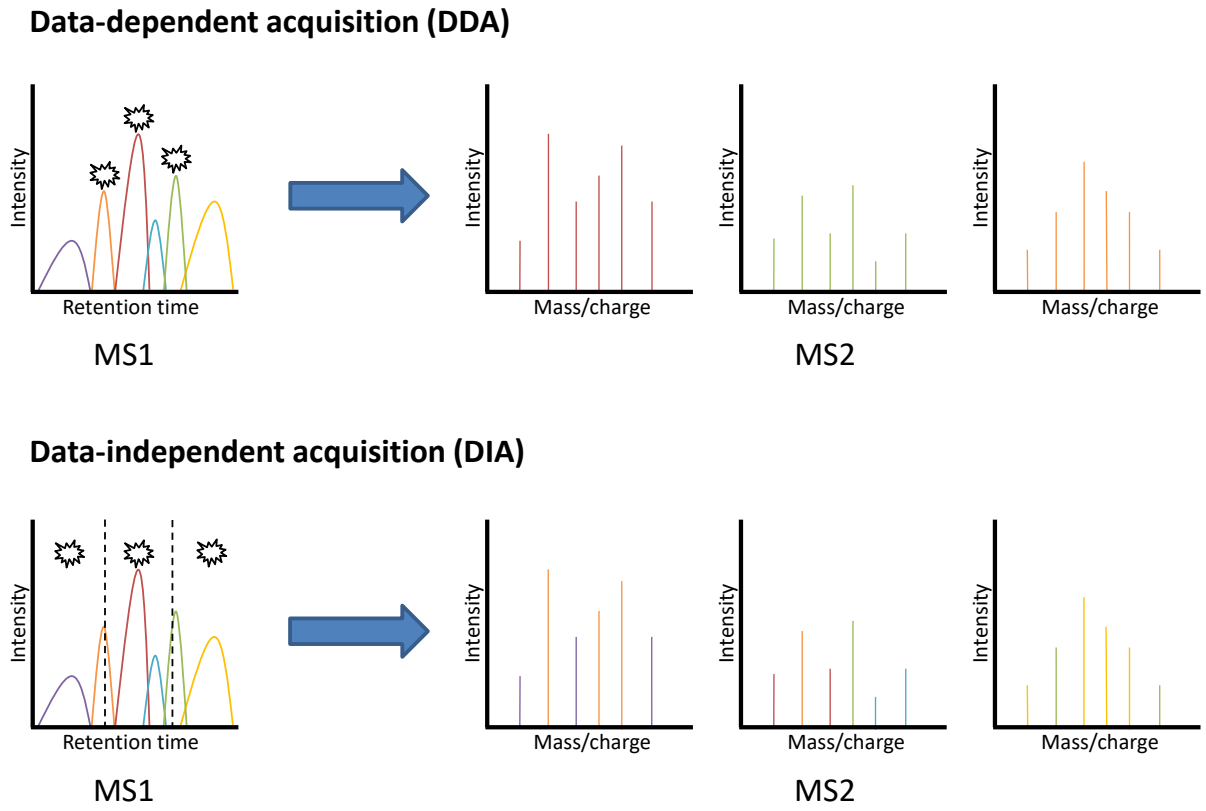


Figure 7. Mass spectrometry acquisition strategies for shotgun proteomics.

There are a number of protocols for the quantification of peptides. One of the simplest is the label-free quantification (LFQ), since no modification has to be made in the workflow. LFQ consists in the calculation of the area under the extracted ion chromatogram of each peptide-specific signal. The disadvantage of this protocol is that it provides a relative quantification, and not an absolute one (Jiang et al., 2024). This makes it unsuitable when studying public health in wastewater, as this matrix is very complex and variable, thus it is a challenged to decide a standard that will not be already present in wastewater (Rice & Kasprzyk-Hordern, 2019). Other quantification methods used in proteomics, that could be applied to wastewater, are stable isotope labeling (mTRAQ or dimethyl labeling), peptide labeling with isobaric tags (TMT and iTRAQ), labelled peptides (AQUA), labelled proteins (PSAQ), and concatemers formed by linking several peptides of interest together like a synthetic protein (QConCat) (Rice & Kasprzyk-Hordern, 2019; Jiang et al., 2024). However, as the study of proteins in wastewater is very recent, these methods have not been applied for now.

Proteomics of wastewater

The technological aspects of environmental proteomics involve protein isolation, fractionation, data acquisition, and identification based on available sequence information and gene ontology (Figure 8). Current proteomic techniques would allow a large-scale characterization of peptides and proteins in sewage and treated water. The detection approach based on LC-MS/MS is well-established and has already been applied to the discovery and validation of biomarkers of kidney disease (Bringans et al., 2017), Alzheimer's (Korecka & Shaw, 2021), and Parkinson's disease (Cilento et al., 2019) in clinical studies. However, one of the main issues that can be encountered is the sample preparation, including protein digestion, which is relatively time-consuming and may be unsuitable for real-time surveillance in early warning systems (Devianto & Sano, 2023).

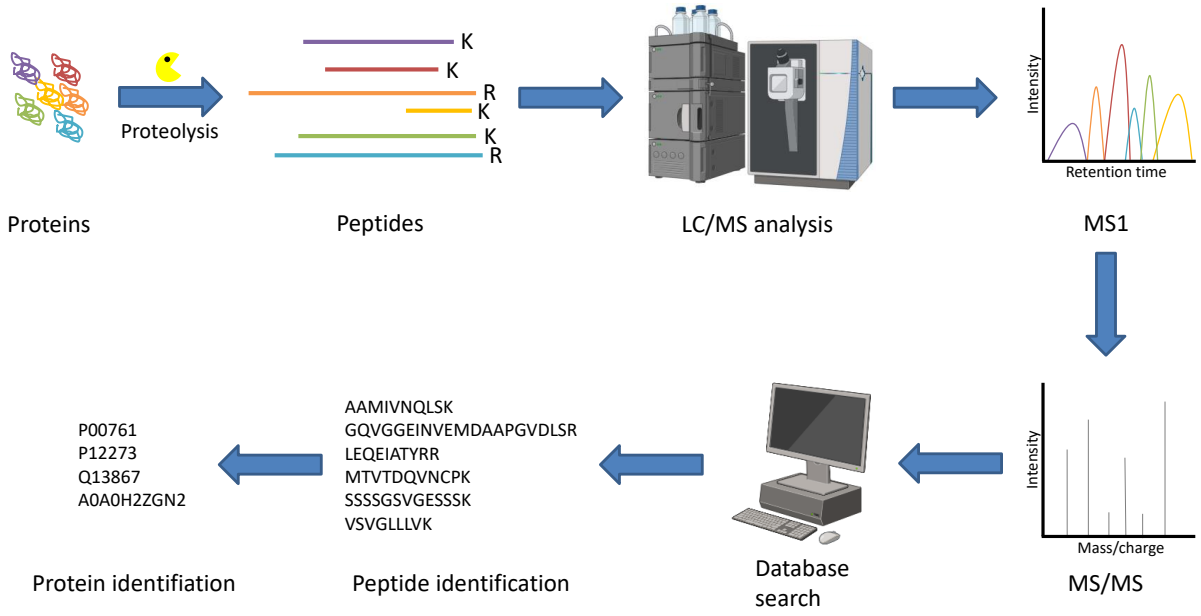


Figure 8. Workflow of a bottom-up protein identification using mass spectrometry.

There are some drawbacks on the use of proteins from wastewater as biomarkers. For example, quantification of expected protein levels in wastewater might prove challenging because protein levels vary from ng to mg/l in urine. If a theoretical 100 times dilution of urine in wastewater is assumed, the estimated levels of proteins should range between pg and µg/l (Rice & Kasprzyk-Hordern, 2019). Another disadvantage is the stability of proteins under wastewater conditions. In aqueous environment, most proteins are present in a folded state stabilized by various interactions, such as hydrogen bonding, disulfide bonding, van der Waals interaction, and salt bridge. The secondary structure of a protein has charged groups on the surface, residues with hydrophilic chains facing water and hydrophobic chains buried in the folded protein (Ahmad, 2022). Nevertheless, the fluctuations of temperature and pH may lead

to the denaturation or unfolding of proteins by the disruption of the secondary and tertiary structures (Butreddy et al., 2021). Protein unfolding and denaturation may lead to the degradation of proteins by proteases already present in the wastewater from urine and feces, or produced by microorganisms. The decay rate of protein markers in wastewater is an essential aspect, and the degradability of protein markers should be understood upon the implementation of WBE.

Sludge proteomics

Until now, the proteomics studies made in wastewater samples have been related to the sludge and their composition. The sludge is a byproduct originated from the bacteria used in the biological reactor(s), which is the primary method used for treating municipal wastewater in industrialized countries before its discharge to receiving waters. This byproduct requires disposal. However, it needs to be stabilized before, frequently by biological treatments such as anaerobic and aerobic digestion. The optimization of the treatments previous to the disposal has made necessary a better understanding of the origin and composition of this sludge (Park et al., 2008a).

The sludge originates from microbial metabolism, cell lysis, and organic matter adsorbed from influent wastewater, providing a polymeric matrix in which microorganisms are embedded. While various organic compounds constitute it, such as polysaccharides, proteins, DNA, humic acids, etc., proteins are known to be the most abundant organic matter in this extracellular matrix (Park et al., 2008a). These extracellular proteins mostly consist of enzymes and structural proteins. Their major functions in wastewater treatment include the formation of microbial aggregate, pollutant migration and transformation via adsorption and catalysis, and resistance to toxic substances by strongly binding heavy metals, organic matter, and nanoparticles. On the other hand, extracellular enzymes supply an external digestion system, playing an important role in the degradation of the organic small molecules, organic colloidal fraction and particulate biomass, which are taken up and utilized as carbon and energy sources by microorganisms. In addition, extracellular proteins can reduce the effects of toxic substances because hydroxyl, carboxyl, phosphate and amide groups in extracellular proteins provide binding sites for antibiotics, heavy metals, and nanoparticles (Zhang et al., 2015; Zhang et al., 2019).

In 2015, Zhang et al. took samples from anaerobic, anoxic and aerobic sludges of a wastewater treatment plant (WWTP) to characterize the extracellular polymeric substance. As mentioned, one of the main components of this substance were proteins, which were involved

in biological processes, such as catalytic activity, binding activity, structural molecule activity, transporter activity, and so on. The original localizations of these proteins were all around the cell (membrane, organelles, macromolecular complex; meaning that they originated from cell lysis) and the extracellular region. Some of the groups were further classified: catalytic proteins had transferase and hydrolase activities, while binding proteins were implicated in organic cyclic compounds, heterocyclic compounds, ion, small molecules, ions, and carbohydrate derivatives binding, among other (Zhang et al., 2015).

Other researchers have tried to identify the proteins constituting the sludge. One of the studies indicated that some proteins have amyloid-like properties, being rich in cross β-sheet and extremely resistant to chemical and thermal denaturation. This type of proteins represents a 5-40% of the sludge biovolume. Although the biological roles of amyloids are still poorly understood, it is proposed they increase the overall hydrophobicity of biofilms, increase their stiffness and even be an antimicrobial peptide. Another find is that some proteins, including those amyloid-like proteins, are glycosylated with carbohydrates containing vicinal hydroxyl groups (e.g., neutral sugar), carboxyl (e.g., sialic acid) and sulfate ester groups. Glycosylation seems to make the proteins more chemical and thermal resistant (Lin et al., 2018).

In 2019, Zhang et al. focused on the bacterial proteins identified to the moment. They classified the proteins in 3 functions: (1) microbial attachment and aggregation; they are important for resisting environmental stress and for sludge-water separation; some important proteins are: chaperonins, binding proteins, metabolism-related proteins, a-synthesis proteins, HesF from *Anabaena sp.*, adhesion and penetration protein autotransporter, type I pili and curli amyloid fibers from *Escherichia coli*; (2) biodegradation; enzymes such as cellulosomal proteins (degradation of any substrate), non-cellulosomal proteins (degradation of particular substrates) and Clocel_3197 (degradation of soft biomass) from *Clostridium cellulovorans*; and (3) response to environmental stress (Zhang et al., 2019).

Limitations of wastewater proteomics

Short-term events (such as work-commuters, tourists, festivals or rainfall) and long-term events (e.g., seasonal changes with wet and dry weather) affect the molecules' presence in wastewater either by diluting/concentrating them or by affecting directly the treatment plant performance. Those changes cannot be controlled and sometimes not even predicted (mostly the short-term ones) (Huizer et al., 2021). Because of this, how samples are collected is a very important first step, as it could alter the conclusions of the study.

The most common sampling in WBE studies is 24-hour composite. This has the advantage to attenuate the influent flow and composition variations that can occur during the time of sampling. This kind of sampling can be done in three ways: 1) flow-proportional: a subsample volume proportional to the flow in the sewer at a constant time interval is taken and then subsamples are weighted individually to form a composite sample; it is the more reliable and the chosen method most of the times; 2) volume-proportional: it takes samples more frequently during higher flows and less frequently during lower flows remaining the sampling volume constant, however, this cannot provide a true average concentration since only the frequency changes and individual samples are not weighted properly according to the flow in the sewer; 3) time-proportional: both frequency and sampling volume are constant and therefore does not provide a true average concentration as well (Huizer et al., 2021).

Once the data is acquired, it has to be analyzed. The most common objective is to determine the concentration of the molecules and then link them with the population. The data about the population usually comes from surveys, drug abuse treatment centers or hospital visits, telephone hotlines, arrests, seizures or trafficking. Nevertheless, in all this data there are sampling bias and time lags, and it also relies in the people's honesty in the case of the surveys, which are also expensive to do (Burgard et al., 2013). Additionally, the per capita estimate is used to know the number of individuals present in the catchment area, but once again this data is usually out of date, and the catchment area and the political boundary do not have to be the same (Daughton, 2012; Burgard et al., 2013).

Data analysis

Protein identification is relatively straightforward with different vendors or freely available search engines. It starts with the comparison of the experimental MS/MS spectra against the theoretical ones obtained from the chosen reference database. The spectra associated with a peptide will be called peptide spectrum match (PSM) and will be paired with a score of how good the fitting is (Jiang et al., 2024). Afterwards, proteins are inferred by grouping together the PSMs that explain them better, and are also associated a fitting score. However, some proteomes are very complex and can include many homologous proteins; there can also be several proteomes implicated and those homologous proteins can come from the same organism or a different one. In result, this can lead to misidentification during the search (Feng et al., 2022). As a general practice, those proteins with the same peptides are reported as protein groups, so at least one of the members will be the correct one (Peng et al., 2023).

The composition of wastewater can vary widely. Moreover, this composition is unknown as it can change from treatment plant to treatment plant and even from day to day in the same

facility. Because the organisms present in the wastewater are not known it is almost mandatory to use in the search a database with all the known organisms in order to not miss identifications. This carries the problem of the homologous proteins, which give place to degenerate peptides. These are those peptides shared among multiple proteins and, therefore, cannot be uniquely attributed to any protein (Feng et al., 2022). This is a problem that is also encountered in metaproteomics (study of microbial communities' proteins). Since the curation of microbial databases seems impossible due to the high number of different but similar microorganisms, some authors are exploring other solutions. Feng et al. (2022) developed an algorithm, called MetaLP, where the taxonomic abundance from metagenomics sequencing is used as prior information for the search. However, apart from microorganisms, animals, plants and other organisms can be found in wastewater, so taxonomic abundance may not be that simple to comprehend in this matrix.

Nowadays, the way to control the quality of PSMs and proteins is the associated score of fitting. There are various statistical approaches, but the most used is the false discovery rate (FDR), which corresponds to the expected fraction of false positive matches. For the calculation of the FDR a decoy database (shuffled or reversed sequences) is used against the target one (database of our choice). There are two approaches: Käll's method (Equation 1a), and Elias and Gygi method (Equation 1b), whose difference is the way they treat the databases and calculate the FDR. FDR is measured in percentage and is usually applied in the range 1-3% (Jiang et al., 2024; Uszkoreit et al., 2024).

$$a)\ FDR = \frac{Decoy\ PSMs + 1}{Target\ PSMs}$$

$$b)\ FDR = \frac{2\ \times Decoy\ PSMs}{Target + Decoy\ PSMs}$$

Equation 1. Calculation of the FDR by Käll's method (a) and Elias and Gygi method (b).

Apart from the database search, there are another two approaches to carry out the identification of MS/MS spectra. The first one is the library matching, where the spectra identified in database searched in previous experiments are joint in a spectral library. This library, instead of the sequence database, is then used to match the experimental spectra. The disadvantage of this approach is that if a peptide does not have any associated spectra in the library, it cannot be identified (Noor et al., 2020). The second approach is used when there is no or limited information in databases and it is called *de novo* sequencing. Here,

peptides' sequences are interpreted from the spectra and then subjected to sequence alignment to find the organism of origin or the closest one (Noor et al., 2020).

WBE is widely applied in the study of small molecules, however the study of the proteins in wastewater is an emerging field. In this aspect, there is a lack of specific tools for this purpose, as well as barely information about the proteins in this matrix, except for the studies made in sludge. The thesis aims to start filling these gaps, combining the development of proteomics' techniques and databases with the need to find new biomarkers for human health and lifestyle, or environmental surveillance. First, a method to analyze different types of water matrixes will be developed following already stablished proteomics methods. Then, this method is used to characterize samples from different treatment plant, first in their entry and, afterwards, in their exit to understand the performance of the WWTPs. Samples from the receiving waters are also analyzed. Finally, one of the long-term objectives in the application of proteomics to wastewater is to find biomarkers, which could complement the information already given by small molecules. For this reason, samples were used for the study of both small molecules and proteins. Overall, this project opens new paths for the study of wastewater using proteomics techniques.

2. OBJECTIVES

The **primary objective** of this study is to identify and characterize protein biomarkers of exposure to possible hazardous compounds in untreated and treated wastewater through environmental proteomics, with the aim of monitoring population health, enhancing environmental surveillance, and evaluating the performance of wastewater treatment plants (WWTPs). Variations in population health and behavior are reflected in the protein profiles of sewage, which can be effectively assessed using environmental proteomics within the framework of wastewater-based epidemiology (WBE).

This issue was addressed through the following **specific objectives**:

1. Development of novel non-target strategies for protein monitoring in different water matrices including urban sewage, wastewater treatment plants (WWTPs) effluents (treated water) and water at different stages of its treatment:

   Sánchez-Jiménez E, Abian J, Ginebreda A, Barceló D, Carrascal M. *Shotgun proteomics to characterize wastewater proteins.* MethodsX. 2023 Sep 26;11:102403. doi: 10.1016/j.mex.2023.102403

Wastewater has been analyzed for small molecules for several decades. In consequence, the techniques used have evolved over time, and today there is a wide spectrum of protocols depending on the sample type and the analysis' aim: filtration and separation methods, use of internal standards, and targeted and non-targeted mass spectrometry, among others. However, there is a lack of a methodology to perform a large-scale characterization of the wastewater proteome, both in solution and in the particulate. In this context, there is a pressing necessity for the development of methods for the study of proteins in wastewater. These methods should include procedures for the separation of the soluble and particulate fractions of wastewater, the clean-up of the sample, the digestion of the extracted proteins, and performing high-resolution mass spectrometry. Since some of the small molecules' analysis are made in the treatment plants themselves with basic instrumentation, the methodology developed for the study of the proteins should be simple and prompt to automatization in order to also be carried out in the facilities.

2. Characterization of potential protein biomarker signatures for early epidemiological alerts and event follow up by correlating population and human health, habits, and activities with sewage protein profiles at different geographical sites:

   Carrascal M, **Sánchez-Jiménez E**, Fang J, Pérez-López C, Ginebreda A, Barceló D, Abian J. *Sewage Protein Information Mining: Discovery of Large Biomolecules as Biomarkers of Population and Industrial Activities.* Environ Sci Technol. 2023 Aug 1;57(30):10929-10939. doi: 10.1021/acs.est.3c00535

Wastewater-based epidemiology has been revealed as a powerful approach for surveying the health and lifestyle of a population. Classically, the characterization of wastewater has been restricted to the measurement of indirect parameters (chemical and biological oxygen demand, total nitrogen, among others) and small molecules of interest in epidemiology or for environmental control. Despite the fact that metaproteomics has provided important knowledge about the microbial communities in wastewater, practically nothing is known about other non-microbial proteins transported in this media. In this regard, Carrascal et al. (2020) was the first study in identifying non-microbial proteins, being the human the species with the higher number of identified proteins. However, this study focused on the proteins attached to a polymer probe and not on the ones soluble in the wastewater. Proteins have been proposed as potential biomarkers that complement the information provided by other available methods. Despite this, little is known about the range of molecular species and dynamics of proteins in wastewater, and the information hidden in these protein profiles is still to be uncovered.

3. Assessment of the efficiency of wastewater treatment by monitoring samples at WWTP influent, effluent and receiving waters:

There is scarce knowledge about how treatments used for the processing of the wastewater affect big biomolecules like proteins. Until now, the efficiency removal has been studied for small molecules, such as illicit drugs, pharmaceutics, personal care products, alcohol or tobacco. The reason is that wastewater treatment plants are built according to the population size to which they serve and to enhance the efficiency in the removal of the contaminants that are regular in that serving site. It is necessary to analyze the proteins that enter (influent) and exit (effluent) the treatment plants serving different sizes of populations and industrial activities to assess the efficiency of wastewater treatment in this kind of biomolecules. It is also important to investigate the receiving bodies of the treated water to elucidate the possible effect in the environment or the organisms.

4. Expanding the Omics toolbox: Complementary metabolomic profiling of wastewater influent:

Commonly only few metabolites are studied in the so-called targeted approach. The identification of new compounds using non-targeted methods can be difficult and time-consuming, and it requires information that sometimes is not available in the public databases. A halfway and recent strategy is to use a suspect list to prioritize some compounds before acquiring the rest in non-targeted mode. Because of the mentioned pitfalls there are few non-targeted studies in metabolomics. Despite being time-consuming it is essential to carry out this type of strategy to enlarge the identification of new compounds which are constantly produced by the industry. Furthermore, the metabolite profile of sites with different population sizes and industrial activity can complement the proteomics studies, allowing the determination of the compounds' origin due to the species information from the proteins. This objective was accomplished through an internship in the University of California Davis with a grant from the Spanish National Research Council (CSIC).

3. SAMPLING

Samples collected from various wastewater treatment plants (WWTPs) were used along the studies conducted in this thesis. The sampling sites and times varied depending on each objective (Table 1).

Table 1. Sampling sites used in the different objectives of this work.

WWTP	Objective			
	1	2	3	4
Banyoles		✓		✓
Besòs	✓	✓	✓	✓
Figueres		✓		
Girona	✓	✓	✓	✓
Granollers		✓		
Igualada		✓		
Manresa		✓		
Mataró		✓		
Olot		✓		✓
Vic	✓	✓	✓	✓

The methods and protocols used for the preparation and analysis of the samples are specified and explained in detail in each Results chapters.

Sampling sites.

Objective 1.

In order to optimize the method used for the analysis of the samples three sites were selected: Besòs and Vic from Barcelona province, and Girona from the province with the same name. Samples were collected on the same day in three different seasons: winter (14th of December 2020), spring (19th of April 2021) and summer (26th of July 2021). The sampling was carried out at the entrance of each WWTP.

Objectives 2 and 4.

For the characterization of the proteins present in the soluble fraction of the wastewater, 10 municipalities were studied, all located in the provinces of Barcelona and Girona. The sampling sites, shown in Figure 9 and detailed in Table 2, were Besòs (Barcelona), Granollers, Igualada, Manresa, Mataró and Vic from Barcelona province, and Banyoles, Figueres, Girona and Olot from Girona. The sampling was conducted at the entrance of each WWTP in three dates: 14th of December 2020, 19th of April 2021 and 26th of July 2021, corresponding to winter, spring and summer campaigns, respectively. Samples were collected simultaneously across all sites during each campaign.

Figure 9. Location of the sampling sites marked with a yellow star.

For the objective 4, five WWTPS of this collection were used: Banyoles, Besòs, Girona, Olot and Vic (see Table 1). These locations were selected based on the human population (Table 2) and the different industrial activities.

Table 2. Code, equivalent population, served population and designed flow per each sampling site.

WWTP	Code	Equivalent population[1]	Served population[2]	Designed flow (m^3/day)[3]
Banyoles	DBAY	53400	28000	12000
Besòs	DBSS	2843750	1502000	525000
Figueres	DFIG	110640	53000	17000
Girona	DGIR	206250	159000	55000
Granollers	DGRA	121500	100000	30000
Igualada	DIGU	285666	67000	20000
Manresa	DMAS	196167	93000	53500
Mataró	DMAT	451250	190000	57000
Olot	DOLO	99166	46000	17000
Vic	DVIC	340000	55000	25000

[1] The equivalent population is the maximum population the treatment plant was designed for. (Source: Agència Catalana de l'Aigua)

[2] The served population is the actual number of individuals living in the site the treatment plant served. (Source: https://www.epdata.es/)

[3] The designed flow is the maximum flow the treatment plant was designed for. In practice, this flow will vary every day. (Source: Agència Catalana de l'Aigua)

Objective 3.

Once the protein composition of the wastewater influent was uncovered, three sampling sites were chosen to address the efficiency of the WWTPs in removing proteins. The sampling sites were Besòs (representing a highly populated city), Vic (an industrial city) and Girona (a city with both a high human population and some industrial activity). For this purpose, three sampling campaigns were conducted, with samples collected over three consecutive days at each site:

- Spring: 25th to 27th of April 2022 (Besòs), 7th to 9th of April 2022 (Girona), and 8th to 10th of May 2022 (Vic).
- Summer: 14th to 16th of September 2022 (Besòs), 30th of August to 1st of September 2022 (Girona) and 5th to 7th of September 2022 (Vic).

- Winter: 13th to 15th of February 2023 (Besòs), 24th to 26th of January 2023 (Girona) and 30th of January to 1st of February 2023 (Vic).

In these campaigns, both the influent and the effluent were collected. The effluent samples were taken according to the hydraulic retention time: 16 hours for Besòs, 24 hours for Girona and 48 hours for Vic.

To investigate whether the proteins detected in the effluent are present in the rivers Girona and Vic sampling sites were selected. In both cases, the treated water goes to a river, Ter and Riera de Rimentol (which joins later to the Gurri river) for Girona and Vic, respectively (Figures 10 and 11). For each site samples were collected at three key points: upstream (before the WWTP), effluent (directly from the WWPT outlet) and downstream (after the WWTP discharge point). The collection was made on May 31st 2023 in Girona and June 21st 2023 in Vic.

Figure 10. Collection points of the upstream (yellow), effluent (blue) and downstream (green) in the Girona WWTP.

Sample collection.

An automated refrigerated collector was used. This equipment allows the collection of samples every several minutes during a certain time keeping the samples refrigerated through all the

time at 4 ºC to limit biological activity. Samples were collected every 15 min during 24 hours, mixing all the collected water. This way the samples used in this study are twenty-four-hours composite samples representing a whole day.

After collection, samples were transported at 4 ºC to the laboratory within the following 24 hours. There, part of the samples were immediately filtered or ultracentrifuged previous to the storage in the freezer (-40 ºC) for further analysis in the following week. The original samples that were not processed were kept at 4 ºC.

Figure 11. Collection points of the upstream (yellow), effluent (red) and downstream (green) in the Vic WWTP.

4. RESULTS

4.1 SHOTGUN PROTEOMICS TO CHARACTERIZE WASTEWATER PROTEINS

Objective 1: Development of novel non-target strategies for protein monitoring in different water matrices including urban sewage, wastewater treatment plants (WWTPs) effluents (treated water) and water at different stages of its treatment.

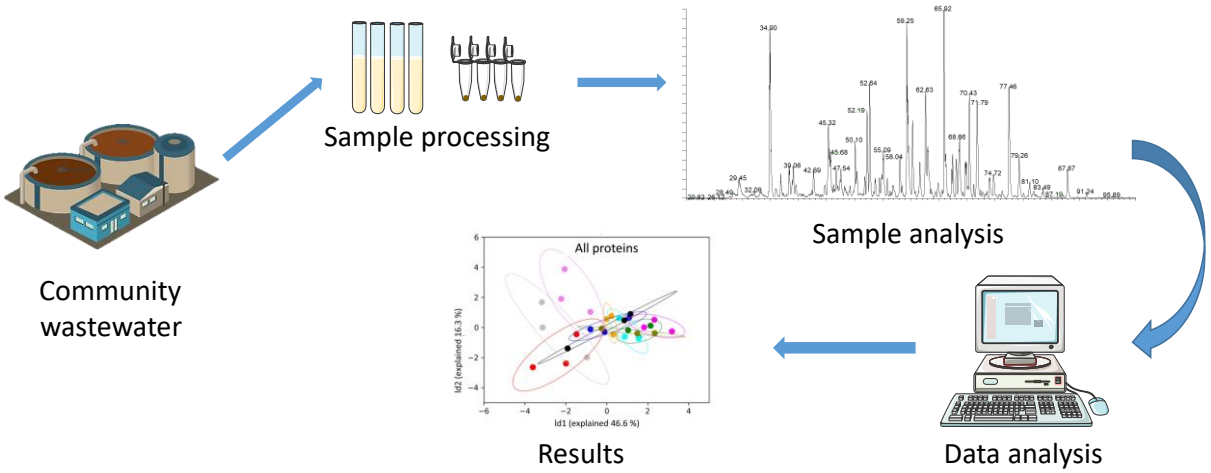
There was not an established method for the study of proteins in wastewater samples before this study commenced. Previous approaches were aimed at the analysis of the sludge inside the treatment plants and only focused on bacterial proteins. The first characterization of the wastewater proteome was carried out by Carrascal et al. (2020) using polymeric probes. In this section, a protocol is developed for the study of both fractions found in wastewater: soluble and insoluble or particulate. There are modifications to adapt the method depending on the fraction. This protocol was published in the journal MethodsX on September 26 2023 with doi: 10.1016/j.mex.2023.102403 (see supplementary material Wastewater_Protocol.pdf). Next, the article is reproduced as published.

ABSTRACT

Classically, the characterization of wastewater components has been restricted to the measurement of indirect parameters (chemical and biological oxygen demand, total nitrogen) and small molecules of interest in epidemiology or for environmental control. Despite the fact that metaproteomics has provided important knowledge about the microbial communities in these waters, practically nothing is known about other non-microbial proteins transported in the wastewater. The method described here has allowed us to perform a large-scale characterization of the wastewater proteome. Wastewater protein profiles have shown to be very different in different collection sites probably reflecting their human population and industrial activities. We believe that wastewater proteomics is opening the doors to the discovery of new environmental and health biomarkers and the development of new, more effective monitoring devices for issues like monitorization of population health, pest control, or control of industry discharges. The method developed is relatively simple and combines procedures for the separation of the soluble and particulate fractions of wastewater and their concentration, and conventional shotgun proteomics using high-resolution mass spectrometry for protein identification (Graphical abstract).

- Unprecedented method for wastewater proteome characterization.
- Proteins as new potential biomarkers for sewage chemical-information mining, wastewater epidemiology and environmental monitoring.
- Wastewater protein profiles reflect human and industrial activities.

Keywords: Wastewater, Proteomics, Compartmentalized proteome, Wastewater-Based Epidemiology (WBE), Sewage Chemical-Information Mining (SCIM), High-Resolution Mass Spectrometry (HRMS).



Graphical abstract. Overview of the process followed for the development of the method.

BACKGROUND

The analysis by mass spectrometry of small molecules in wastewater is a widely adopted approach for population monitoring by governmental institutions. Typical applications are, for example, monitoring the consumption of tobacco, alcohol or drugs by a community and, more recently, the COVID prevalence using polymerase chain reaction (PCR) (Picó & Barceló, 2021). Current development of mass spectrometry allows extending this approach to large molecules such as proteins, as it has been stressed by several authors (Picó & Barceló, 2021; Rice & Kasprzyk-Hordern, 2019). This would open the window to the monitorization of human health biomarkers already known in clinics, as well as environmental biomarkers. All of this in the context of the sewage chemical-information mining (SCIM) (Daughton, 2018).

Previously, we reported for the first time on the peptide and protein components absorbed in a support submerged in wastewater at a wastewater treatment plant (WWTP) (Carrascal et al., 2020; Perez-Lopez et al., 2021). We detected proteins from prokaryotic to higher eukaryotic organisms, covering plant, animal, and human proteomes as well. We were able to identify not only major components (albumins, keratins), but also less abundant ones that are known as disease biomarkers. In this study, we used polymer probes to trap the proteins; although an effective method, it required many days to collect the samples and the proteins could be biased by the polymer affinity and the biofilms' formation. Therefore, we centered on developing a strategy that allowed the study of the proteome directly in wastewater samples collected in the inlet of WWTPs (Carrascal et al., 2023). Here we offer an improved method based on it, where we establish the recovery of the soluble and particulate fractions and the process followed for obtaining the peptides.

LABORATORY EQUIPMENT

Apparatus

- Bullet Blender Homogenizer: model Storm 24, Next Advance.
- Centrifuge: model Allegra 25R, Beckman Coulter.
    + Centrifuge rotors: TS-5.1-500 (swinging) and TA-15-1.5 (fixed angle), Beckman Coulter.
- Centrifuge: model MiniSpin, Eppendorf.
    + Centrifuge rotor: F-45-12-11 (fixed angle), Eppendorf.
- Digestor: model Digest Pro-MS, Intavis.
- Electrophoresis gel tank: model Mini-Protean Tetra Cell, Biorad.
- Electrophoresis power supply: model PowerPac 1000, Biorad.
- High resolution mass spectrometer (HRMS): Q Exactive TM HF, Thermo Scientific.
- Thermomixer: model F2.0, Eppendorf.
- Ultracentrifuge: model 90SE, Sorvall Discovery.
    + Ultracentrifuge rotor: SW-28 (swinging), Beckman Coulter.
- Ultra High-Performance Liquid Chromatographic (UHPLC) system: UltiMate 3000 RSLCnano, Thermo Scientific TM.
- Vacuum concentrator: Savant SpeedVac SPD130DLX, Thermo Scientific.

Reagents

- 0.1 mm Zirconium silicate beads (11079101z, BioSpec).
- 10% Sodium dodecyl sulfate (SDS, 71,736–500ML, Sigma-Merck).
- 30% Acrylamide/Bis Solution 37.5:1 (2.7% crosslinker) (161-0158, Biorad).
- Acetonitrile (ACN, 10,001,334, Fisher Scientific).
- Ammonium bicarbonate (ABC, A6141-25 G, Sigma-Merck).
- Ammonium persulfate (APS, GE17-1311-01, Sigma-Merck).
- β-mercaptoethanol (M3148-25ML, Sigma-Merck).
- Bovine serum albumin (BSA, A8022-50 G, Sigma-Merck).
- Bromophenol blue (GE17-1329-01, Sigma-Merck).
- Dithiothreitol (DTT, D5545-1 G, Sigma-Merck).
- Glycerol (GE17-1325-01, Sigma-Merck).
- Glycine (G8898-500 G, Sigma-Merck).
- Iodoacetamide (IAA, I1149-25 G, Sigma-Merck).
- Methanol (MeOH, 15,624,740, Fisher Scientific).

- MilliQ H_2O (Millipore).
- Phosphate-buffered saline (PBS, P4417-100TAB, Sigma-Merck).
- SimplyBlue SafeStain (LC6060, Thermo Fisher).
- Sodium hydroxide (NaOH, S8045-500 G, Sigma-Merck).
- Tetramethylethylenediamine (TEMED, 10,549,960, Fisher Scientific).
- Trifluoroacetic acid (TFA, 302,031-100ML, Sigma-Merck).
- Tris (GE17-1321-01, Sigma-Merck).
- Trypsin (V5280, Promega).
- Tween-20 (P1379-500ML, Sigma-Merck).

Materials and buffers

- 0.5 ml LoBind tubes (10,316,752, Fisher Scientific).
- 1.5 ml LoBind tubes (10,708,704, Fisher Scientific).
- Amicon Ultra-15 Centrifugal Filter Unit 10 kDa cut-off (UFC901024, Sigma-Merck).
- Analytical column: 25 cm x 75 µm, C18, 1.6 µm, Odyssey (DY-25075C18A, Ionopticks).
- Elution plate (40.020, Intavis).
- Reaction plate (40.010, Intavis).
- Ultracentrifuge tubes 38.5 ml (082,326,823, Izasa Scientific).
- Trap column: Acclaim PepMap100 100 µm x 2 cm nanoViper C18 5 µm 100 Å (164,564, Thermo Scientific).
- Extraction buffer: ACN/milliQ H_2O (1/1, v/v) 0.25% v/v TFA.
- Lysis buffer: 4% SDS, 0.1 M DTT, 100 mM Tris-HCl pH 7.5.
- Reconstitution buffer: 5% MeOH 0.5% TFA.
- Resolving buffer (4x): 1.5 M Tris-HCl pH 8.8.
- Running buffer (10x): 0.25 M Tris-HCl, 1.92 M Glycine, 1% SDS pH 8.3.
- Sample buffer (5x): 0.1 M Tris-HCl, 44% Glycerol, 6% SDS, 0.03% bromophenol blue.
- Stacking buffer (4x): 0.5 M Tris-HCl pH 6.8.

PROCEDURE

Sampling

Twenty-four-hour composite water samples were collected at the WWTP inlet using refrigerated automatic wastewater samplers that took water samples every 15 min during this period. Samples are collected in the 24 h prior to the beginning of the experiment and stored at 4 °C.

Separation of the soluble and particulate fractions

- Place 30 ml of the water sample in an ultracentrifuge tube.
- Ultracentrifuge at 24,000 rpm (~100,000 x g) for 60 min at 4 °C, Accel = 9, Decel = 9.
- Separate the supernatant from the pellet by pouring the supernatant in a new tube.
- At this point, the procedure can be stopped to continue later on. In this case, store both the supernatant and the pellet at -70 °C.

Protein extract preparation

Concentration of the soluble fraction (Casanovas et al., 2017)

- Add 2.5 ml of 0.1 M NaOH to the Amicon filter unit and centrifuge in the Beckmann centrifuge (TS-5.1-500) at 4000 x g 13 °C for 10 min.
- Add 2.5 ml of milliQ H_2O to the filter and centrifuge at 4000 x g for 15 min at 13 °C.
- Immerse the filter in Tween-20 5% overnight.
- Wash extensively the Tween-20 from the filter by immersing it in milliQ H_2O for 1-2 min and gently shaking. Discard the water and repeat this wash four more times. Make sure there is no Tween left (no bright soapy bubbles when the filter is shaken to remove the water). Otherwise, repeat the wash.
- Add 12 ml of milliQ H_2O to the filter and centrifuge at 4000 x g for 15 min at 13 °C. (x 2)
- Transfer 10 ml of the supernatant from the ultracentrifugation (Step 2.1) to the Amicon filter unit.
- Centrifuge at 4000 x g for 15 min at 13 °C.
- Transfer the concentrated sample left in the filter (about 500 µl) to a 1.5 ml LoBind tube.
- Evaporate the sample in the vacuum concentrator until dry.
- At this point or before evaporation, the procedure can be stopped to continue later on. In this case, store samples at -70 °C.

Lysis of the particulate fraction (Casas et al., 2017)

- In case the pellet from the ultracentrifugation is stored at -70 °C (Step 2.1), thaw it at 4 °C.
- Reconstitute the pellet in 30 ml of cold PBS.
- Ultracentrifuge at 100,000 x g for 60 min at 4 °C, Accel = 9, Decel = 9.
  Note: For the Sorvall ultracentrifuge, we use 24,000 rpm (103,864 x g).
- Remove the supernatant to waste.

- Add 250 µl of lysis buffer to the pellet in the ultracentrifuge tube and sonicate in a bath for 5 min.
- Transfer to a new 1.5 ml LoBind tube.
- Add 250 µl of lysis buffer to the ultracentrifuge tube and sonicate in a bath for 5 min.
- Transfer to the previous 1.5 ml LoBind tube.
- Incubate in the Thermomixer for 60 min at 95 °C and 800 rpm.
- Let the sample cool down at room temperature.
- Add 500 µl of the sample to 250 µl of Zirconium Silicate beads.
- Let it in the Bullet Blender 3 min at level 8.
- Centrifuge in the Beckman centrifuge (rotor TA-15-1.5) for 10 min at 18,000 x g and 22 °C.
- Transfer the supernatant to a new 1.5 ml LoBind tube (approx. 400 µl).
- At this point, the procedure can be stopped to continue later on. In this case, store lysed samples at -70 °C.

Concentration of the soluble and particulate fractions with SDS-page gels

- Prepare two SDS-PAGE gels (12% resolving, 5% stacking). Use 10 and 5-well combs for the soluble and the particulate fractions, respectively.
- Add 40 µl of the sample buffer 1x with 5% of β-mercaptoethanol to the dried soluble sample.
- Add 7 µl of glycerol and 2 µl of bromophenol blue to a 25% aliquot (approx. 100 µl) of the particulate sample.
- Incubate samples in the Thermomixer at 95 °C for 10 min.
- Let samples cool down at room temperature and spin down.
- Load the samples in the gels:
    + Soluble fraction: 40 µl per well in a 10 wells gel.
    + Particulate fraction: 110 µl per well in a 5 wells gel.
    + Standard: 1 µg of BSA per well.
- Run the electrophoresis at 50 V until samples are stacked at the head of the gel (approx. 50 min).
- Wash the gel with milliQ H_2O for 5 min. (x 3)
- Stain the gel with SimplyBlue for 1 hour.
- Wash the gel with milliQ H_2O for 1 hour.
- At this point, the procedure can be stopped to continue later on. In this case, store the gel at 4 °C.

In-gel digestion with trypsin (Casanovas et al., 2009)

- Excise the bands of concentrated proteins taking as a reference the beginning of the BSA band with a scalpel as shown in Figure 13, cut them into small pieces (ca. 1 mm^3) and place them in different wells of the Digestor reaction plate (use more than one well per sample if necessary).
- Place the reaction plate, the elution plate and the reagents into the Digestor.
- Run the method with the following steps:
    + Wash gel slices with 100 µl of 20 mM ABC pH 7.8 and incubate for 15 min. (x 2)
    + Wash with 100 µl of ACN and incubate for 15 min. (x 2)
    + Reduce with 60 µl of 10 mM DTT for 50 min.
    + Alkylate with 60 µl of 55 mM IAA for 30 min.
    + Wash with 100 µl of 20 mM ABC pH 7.8 and incubate for 15 min. (x 3)
    + Wash with 100 µl of ACN and incubate for 15 min. (x 3)
    + Add 40 µl of trypsin 3.1 ng/µl and incubate for 10 min.
    + Add 60 µl of 20 mM ABC pH 7.8 without removing the trypsin and incubate for 7 h at 37 °C.
    + Extract peptides with 50 µl of ACN/milliQ H_2O (1/1, v/v) 0.25% v/v TFA incubating for 15 min. (x 2)
    + Add 40 µl of ACN/milliQ H_2O (1/1, v/v) 0.25% v/v (TFA) and incubate for 15 min.
    + Add 40 µl of ACN and incubate for 15 min.
- Transfer digested samples to new 0.5 ml LoBind tubes and evaporate to dryness.

Analysis by liquid chromatography coupled to high resolution mass spectrometry (LC-HR-MS)

- Reconstitute samples in 50 µl of 5% MeOH 0.5% TFA.
- Inject 10% of each sample.
    + LC parameters:
        · Sampler temperature: 8 °C.
        · Column temperature: 35 °C.
        · Solvents: 0.1% formic acid (A) and acetonitrile 0.1% formic acid (B).
        · Gradient:

Time (min)	Flow-rate (µl/min)	%B
0	15	3
3	0.3	3
93	0.3	45
95	0.3	95
99	0.3	95
101	0.3	3
105	Stop run	

+ MS parameters:
  · Spray voltage: 1500 V.
  · Data-dependent mode: 10 MS/MS scans of the 10 most intense signals detected in the MS scan.
  · Fragmentation: HCD (27%).
  · Full MS (range 375-1600) acquired in the Orbitrap with a resolution of 60,000.
  · MS/MS spectra (range 200-2000) obtained in the Orbitrap with a resolution of 15,000.

Database search with proteome discoverer 3.0.1.27

- Processing workflow parameters:
  + Min. precursor mass: 350 Da.
  + Max. precursor mass: 5000 Da.
  + Database: Swiss-Prot (February 2023).
  + Min. peptide length: 6.
  + Precursor mass tolerance: 20 ppm.
  + Fragment mass tolerance: 0.02 Da.
  + Dynamic modifications: Acetyl in N-term, oxidation in M, Met-loss in M and Met-loss + Acetyl in M.
  + Static modifications: Carbamidomethyl in C.
  + Percolator (FDR targets):
    · Target FDR (Strict): 0.001 (0.1%).
    · Target FDR (Relaxed): 0.005 (0.5%).
- Consensus workflow parameters:
  + Precursor ions quantifier:
    · Peptides to use: Unique.

· Precursor abundance based on: Intensity.

· Normalization mode: Total peptide amount.

+ Peptide validator:

· Target FDR (Strict) for PSMs: 0.001 (0.1%).

· Target FDR (Relaxed) for PSMs: 0.005 (0.5%).

· Target FDR (Strict) for Peptides: 0.001 (0.1%).

· Target FDR (Relaxed) for Peptides: 0.005 (0.5%).

+ Minimum peptide length: 6.

+ Protein FDR validator:

· Target FDR (Strict): 0.01 (1%).

· Target FDR (Relaxed): 0.05 (5%).

For the complete list of parameters see Supplementary Material (4.1_PD30_SearchParameters.docx).

Table 3. Collection sites. Population equivalent, population served and water treated at the different WWTPs.

WWTP	Population (thousands)		Water treated (m^3/d)
	Equivalent[a]	Served[b]	
Besòs	2944	1502	322,238
Girona	206	159	43,556
Vic	340	55	22,384

[a] The equivalent population is the population the wastewater treatment plant was designed for. Agència Catalana de l'Aigua (21/11/2022), https://aca.gencat.cat/ca/laigua/infraestructures/estacions-depuradores-daigua-residual/.

[b] The served population is the actual number of people that live in the area whose wastewater the treatment plant receives. https://www.epdata.es/.

[c] https://sarsaigua.icra.cat/.

APPLICATION

Data presented here correspond to the analysis of 24-h composite wastewater samples collected between April and May 2022 at the inlet of 3 different wastewater treatment plants (WWTPs) serving cities of different population and diverse industrial activities (Table 3).

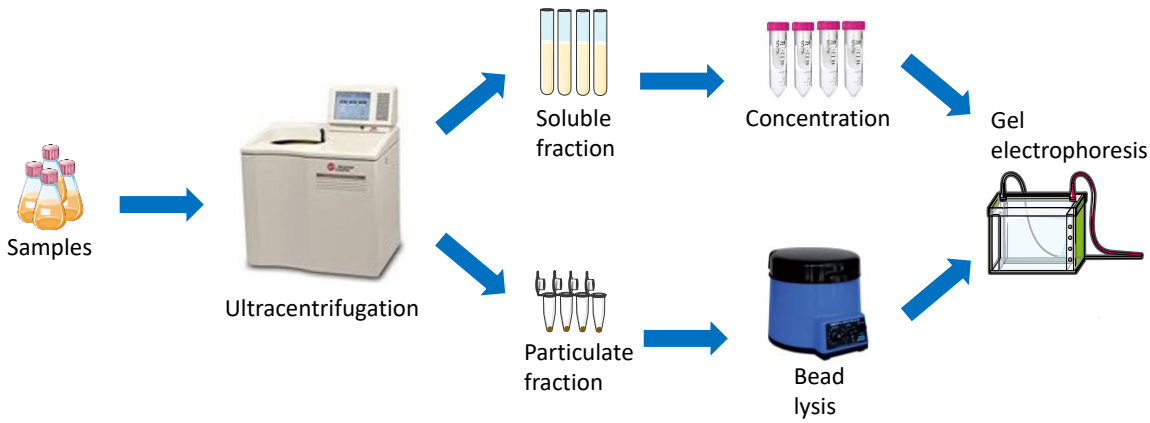


Figure 12. Outline of the sample fractionation and fraction processing.

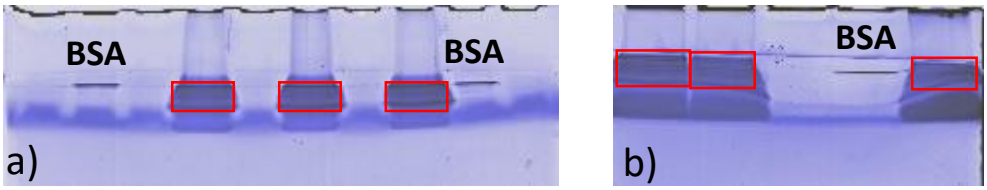


Figure 13. Bands excised taking the BSA reference for the soluble fraction (a, 10 wells gel) and the particulate fraction (b, 5 wells gel).

The soluble and particulate fractions were separated by ultracentrifugation and the particulate fraction was disrupted and homogenized mechanically. Solubilized proteins were concentrated in a small band at the head of a polyacrylamide gel (Figure 12). This band was submitted to the conventional procedures in shotgun proteomics, namely, in-gel digestion and LC-MS/MS analysis followed by database identification (Figure 14). The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (Jarnuczak & Vizcaíno, 2017) partner repository with the dataset identifier PXD042445.

The number of identifications ranged between 244 and 386 for the soluble part and between 488 and 935 in the particulate (Table 4). The proteins and the origin of these proteins in each fraction is very different (Figure 15). While in the particulate fraction Bacteria proteins dominate, in the soluble fraction the number of Eukaryotic proteins increases up to 42% of total protein. In the case of the soluble fractions, the 3 most abundant proteins are amylase enzymes from humans with the double of Peptide-Spectrum Matches (PSMs) than the next most abundant protein. These human amylases are followed by amylase enzymes from murids, human albumin and by albumins from livestock (cow, pig, sheep or rabbit). In contrast, in the particulate fraction, though the most abundant protein is a human elastase (with 4 times less PSMs than the most abundant soluble proteins), it is followed by a mix of proteins from human and different kinds of bacteria.

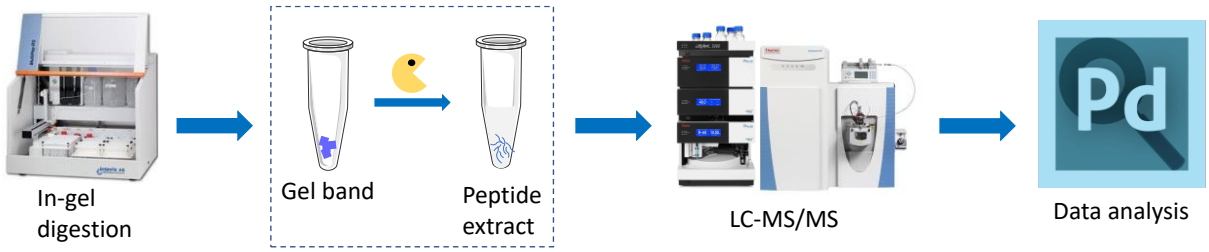


Figure 14. Outline of the protein digestion and peptide analysis.

Table 4. Number of proteins and peptides identifications by site and fraction (see Supplementary Material 4.1_Particulate_Identifications.xlsx and 4.1_Soluble_Identifications.xlsx for detailed information).

WWTP	Proteins		Peptides	
	Soluble	Particulate	Soluble	Particulate
Besòs	265	935	1183	1894
Girona	244	791	1024	2030
Vic	386	488	1277	1091

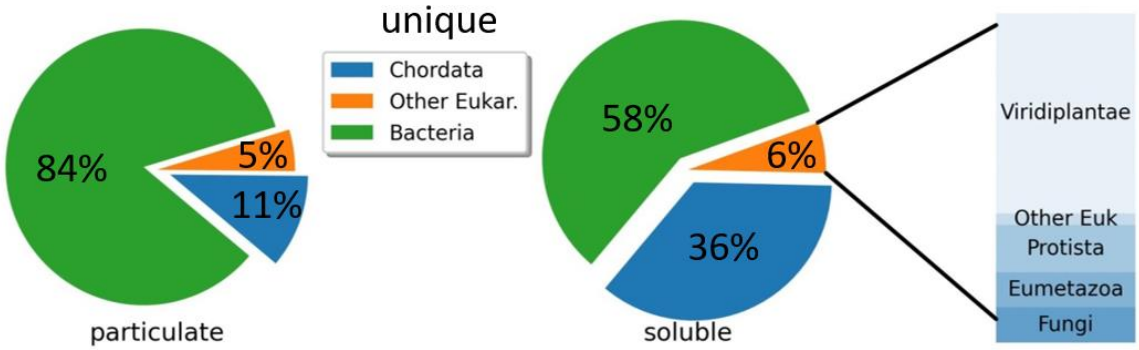


Figure 15. Distribution of the number of Bacteria and Eukaryotic proteins in the particulate and soluble fractions.

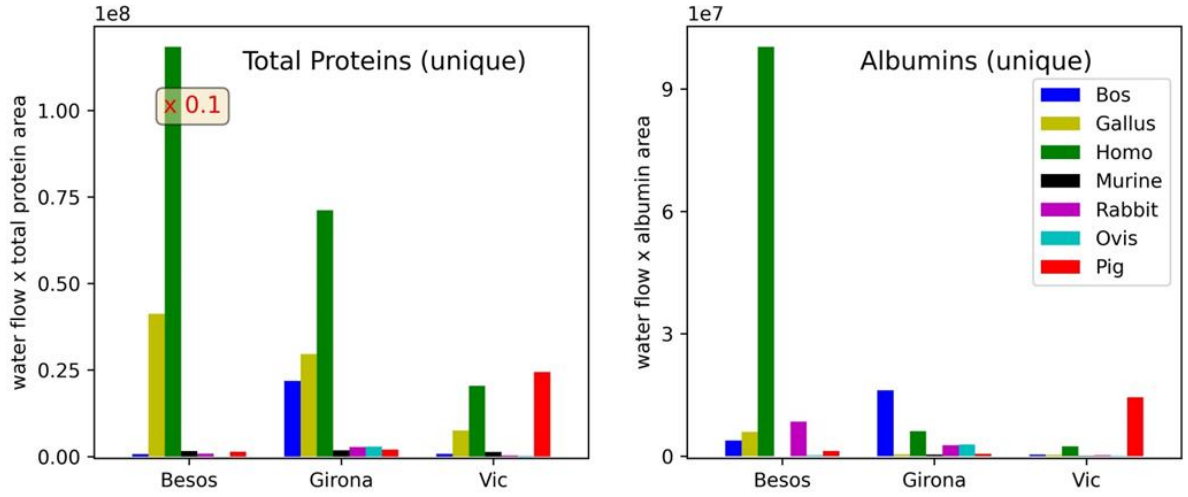


Figure 16. Distribution by sites of total proteins (right) and albumins (left) from Human, murids and different livestock species. Abundance calculation was made considering only protein unique peptides.

The data allows to characterize the contribution of different species to the wastewater proteome at each site (Figure 16). In cities with a high population density like Besòs (Barcelona) human proteins are dominant, while in more rural areas, such as Vic, livestock proteins can be more abundant than human proteins. The major contributor to livestock proteins is albumin as reflected in the small difference between total protein and albumin abundances in Figure 14 for farm animals. This albumin probably comes from the spillage of animal blood and tissue in the wastewater, thus reflecting the activity of the meat industry in these areas.

CONCLUSIONS

Wastewater proteomics is an emerging field within sewage chemical information mining and wastewater-based epidemiology. Until now, the study of wastewater was focused on small molecules. To our knowledge, this is the first method developed for the large-scale characterization of proteins in this kind of matrix, overcoming issues such as the heterogeneity and complexity of the wastewater, and the interferences from other molecules. We have shown that proteins in wastewater transport information on the human and industrial activities in the urban and rural areas from which these discharges originate.

The method described here has slight modifications relative to our previous works (Carrascal et al., 2023) in order to optimize the efficiency and simplify the resources needed for the analysis. Work is in progress to extend the approach to quantitative targeted methodologies.

We believe this method is an effective discovery tool in the search of community biomarkers and a first step to the development of specific test devices for health and environmental monitoring.

4.2 SEWAGE PROTEIN INFORMATION MINING: DISCOVERY OF LARGE BIOMOLECULES AS BIOMARKERS OF POPULATION AND INDUSTRIAL ACTIVITIES

Objective 2: Characterization of potential protein biomarker signatures for early epidemiological alerts and event follow up by correlating population and human health, habits, and activities with sewage protein profiles at different geographical sites.

Until now, wastewater studies were focused on small molecules, both elucidating which ones arrive at the treatment plant and which ones exit and go to the receiving waters. Moreover, as commented in the previous chapter, the protein studies were carried out inside the WWTP and did not include all the organisms. In this section, the wastewater proteome is characterized across ten wastewater treatment plants (WWTPs), describing for the first time the differences attributable to human activity in the various locations analyzed. This study was published in the journal Environmental Science & Technology on July 18 2023 with doi: 10.1021/acs.est.3c00535 (see supplementary material Wastewater_Proteome.pdf). The article is reproduced below.

ABSTRACT

Wastewater-based epidemiology has been revealed as a powerful approach for surveying the health and lifestyle of a population. In this context, proteins have been proposed as potential biomarkers that complement the information provided by currently available methods. However, little is known about the range of molecular species and dynamics of proteins in wastewater and the information hidden in these protein profiles is still to be uncovered. In this study, we investigated the protein composition of wastewater from 10 municipalities in Catalonia with diverse populations and industrial activities at three different times of the year. The soluble fraction of this material was analyzed using liquid chromatography high-resolution tandem mass spectrometry using a shotgun proteomics approach. The complete proteomic profile, distribution among different organisms, and semiquantitative analysis of the main constituents are described. Excreta (urine and feces) from humans, and blood and other residues from livestock were identified as the two main protein sources. Our findings provide new insights into the characterization of wastewater proteomics that allow for the proposal of specific bioindicators for wastewater-based environmental monitoring. This includes human and animal population monitoring, most notably for rodent pest control (immunoglobulins (Igs) and amylases) and livestock processing industry monitoring (albumins).

KEYWORDS: environmental proteomics, sewage epidemiology, water fingerprinting, mass spectrometry

1. INTRODUCTION

Sewage chemical-information mining (SCIM) (Daughton, 2018), of which wastewater-based epidemiology (WBE), also known as sewage epidemiology, is the more relevant branch, has arisen as a complementary alternative to provide comprehensive health and environmental

information on communities. Under this approach, sewage is regarded as an integrated pooled sample of the entire population served by a certain wastewater system; thus, its monitoring provides an average picture of its health status and activities (Daughton, 2018; Rice & Kasprzyk-Hordern, 2019; Sims & Kasprzyk-Hordern, 2020).

The success achieved through SCIM has been closely related to instrumental development, especially on mass spectrometry (MS) for the analysis of small and large molecules, and more recently by the introduction of techniques for the analysis of genetic material (Picó & Barceló, 2021). Some successful applications of SCIM include the consumption of illegal drugs (Mastroianni et al., 2017; Thomas et al., 2012), pharmaceuticals and personal care products (Burgard et al., 2013; Gao et al., 2016), tobacco (Castiglioni et al., 2014) and alcohol use (Ryu et al., 2016a), the exposure to toxicants like pesticides (Rousis et al., 2017), and Bisphenol A (Lopardo et al., 2019), and with regard to biological response, oxidative stress (Ryu et al., 2016b) or the monitoring of coronavirus prevalence during the recent COVID-19 outbreak (Alygizakis et al., 2021; Barceló, 2020).

In this context, several authors have stressed the potential relevance of proteins in wastewater as health and environmental biomarkers (Rice & Kasprzyk-Hordern, 2019; Picó & Barceló, 2021). Early studies already evidenced the presence of enzymatic activity in the effluent of wastewater treatment plants (WWTPs) (Westgate & Park, 2010), and human keratins and pancreatic elastase were identified among a few other bacterial proteins in sludge using the proteomic technology available at that moment (Park et al., 2008b). The presence of human proteins in sludge evidenced its resistance to degradation in wastewater and through the WWTP treatment and raised the question of their effect in the receiving waters (Westgate & Park, 2010). More recently, using ELISA analyses, quantitation of human immunoglobulins A and G in wastewater was reported and proposed as a tool for community serology (Agan et al., 2022). Besides these works, most sewage proteomic studies have focused on the characterization of the microbiome in either sludge (Kuhn et al., 2011) or wastewater (Westgate & Park, 2010; Zhang et al., 2019), and the information on other human, animal, or vegetal proteins remains scarce at best.

The current status of proteomics technologies allows sensitive and extensive analysis of very complex protein mixtures such those in wastewater. Disentangling the wastewater proteome would open the window to a new class of potential markers for SCIM purposes and would be the first step for developing new specific, targeted analytical methods to monitor anthropogenic activities and community health status in a non-intrusive way.

With this aim, in preliminary studies (Carrascal et al., 2020; Perez-Lopez et al., 2021), we used passive sampling polymeric devices and liquid chromatography coupled to high-resolution MS shotgun proteomic methods, to expand, for the first time, the proteomic profiling of wastewater beyond prokaryotes to eukaryotes higher organisms, covering plants, animals, and human proteomes. For the latter, we were able to identify not only the major proteome constituents, such as albumins and keratins, but also other less abundant proteins (for example, S100A8, uromodulin, and defensins), which are known as potential disease biomarkers. This seminal work can thus be regarded as a first attempt to disentangle the entire wastewater proteome, and, simultaneously, it highlighted the experimental and analytical challenges involved in its characterization.

In our previous work, the heterogeneity and complexity of the water samples drove us to use semi solid polymer probes in order to trap wastewater protein and allow their analysis minimizing interferences. While the method was effective, it required letting the probe submerged for many days. Further, the set of proteins trapped was very probably biased by the polymer affinity or the formation of biofilms in their surface. Consequently, we focused on developing strategies for the characterization of the proteome directly from wastewater using existing automatic infrastructure for water collection at WWTP entrances. Here, we present our results on the characterization of the soluble fraction of the wastewater proteome (filtered through 200 nm pore) from 10 different municipalities in Catalonia covering a wide range of population sizes and influent characteristics (relative contribution of domestic and industrial load).

The objectives of the present study were: (a) the deep proteomic characterization of the wastewater soluble fraction and (b) to describe the observed protein pattern and their possible correlation with human activity in order to identify potential biomarkers that could be validated for new applications for SCIM or WBE or become monitoring targets for the improvement of WWTP operation and management.

Consequently, first we will describe the collection of proteins identified, their origin, distribution and possible correlation with anthropogenic activities, and then we will discuss the potential utility of some of the more abundant and ubiquitous human and animal protein families identified in wastewater.

2. MATERIAL AND METHODS

2.1. Sample collection

Twenty-four-hour composite wastewater samples were collected at the inlets of 10 wastewater treatment plants (WWTPs) located in the Girona and Barcelona provinces in Catalonia (Figure 17). These WWTP receive influents from different municipalities with populations ranging 28,000 (Banyoles) to 1,500,000 (Besòs) as well as diverse activities (Table 5). Samples were collected in the framework of the Catalonian Net for SARS-CoV-2 surveillance (Guerrero-Latorre et al., 2022). An automatic water sampler was used at all sites. The samples were then transferred to the laboratory at 4 °C.

Figure 17. Location of the 10 WWTPs where samples were collected.

Three collection campaigns were conducted on the 14th of December 2020, and on the 19th of April and 26th of July 2021 (winter, spring, and summer campaigns, respectively). For the study of the particulate fraction, samples were collected on three different days at the entrance of WWTP Besòs and Vic in May 2022.

Table 5. Collection sites. Population equivalent, population served, and water treated at the different WWTPs.

WWTP	Population (thousands)		Water treated[1] (m^3/d)
	Equivalent[1]	Served[2]	
Banyoles	53	28	12,000
Besòs	2,844	1,502	525,000
Figueres	111	53	17,000
Girona	206	159	55,000
Granollers	122	100	30,000
Igualada	286	67	20,000
Manresa	196	93	53,500
Mataró	451	190	57,000
Olot	99	46	17,000
Vic	340	55	25,000

[1] Agència Catalana de l'Aigua (21/11/2022), https://aca.gencat.cat/ca/laigua/infraestructures/estacions-depuradores-daigua-residual/

[2] https://sarsaigua.icra.cat/ and https://www.epdata.es/

Data on water inflow measured on the day of collection were provided by the WWTP operators (see supplementary material 4.2_WWTP_Physicochemical.xlsx).

2.2. Sample preparation

Soluble Proteins. The collected samples were filtered immediately after arrival at the laboratory. For this, up to 100 ml of 24-h composite wastewater sample was centrifuged at 4000 × g (10 °C, 20 min), and the supernatant was filtered through 0.2-µm filters (VWR, North American, USA). The filtered samples were lyophilized using a freeze-dryer (TELSTAR LyoAlfa 6, PA, USA).

For the analysis, lyophilized samples were reconstituted in 20 ml MilliQ water and concentrated using a 10 kDa cutoff device (Amicon®, NMWL 10 kDa), with a filter that was previously passivated to minimize protein adsorption. Passivation was performed by washing the filter with 2.4 ml of NaOH 0.1M that was eliminated by centrifugation at 4000 × g (13 °C, 10 min), followed by a second wash with 2.5 ml milliQ water and centrifugation. Finally, filters were immersed overnight in Tween-20 (5% in MilliQ water), extensively washed with MilliQ

water, and centrifuged at 4000 × g (13 °C, 5 min). Samples were concentrated to approximately 400 µl, then evaporated to dryness using a SpeedVac. Proteins in the sample were cleaned and concentrated in the heads of sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) gels (5% stacking and 12% resolving) at 50 V for 40–50 min. Bovine serum albumin (BSA) was used as a reference marker. After electrophoresis, the gels were stained with Coomassie Blue and scanned. The bands of concentrated proteins were excised and digested with trypsin using an automatic device (DigestPro MS, Intavis), as previously described (Casanovas et al., 2009).

Proteins in Particulate. The three 24-h composite wastewater samples from each WWTP were combined, and up to 30 ml was processed. First, samples were centrifuged at 600 × g (15 °C, 15 min), and the supernatants were ultracentrifuged at 112700 × g (4 °C, 20 h, accel=9, desel=9) (Sorvall Discovery 90SE with rotor SW-28). Thereafter, the pellets were washed with phosphate-buffered saline using ultracentrifugation under the same conditions as described before.

After washing, the pellets were lysed with beads as described by Casas et al. (2016). Briefly, pellets were reconstituted in 500 µl of denaturing lysis buffer containing 4% SDS, 0.1 M DTT, and 100 mM Tris-HCl pH 7.5 through sonication in a bath (Ultrasons, J.P. Selecta) and incubated in a Thermomixer (Eppendorf, model F2.0) at 95 °C (800 rpm, 1 h). Samples were then homogenized in a Bullet Blender (Next Advance Storm, NY, USA) for 3 min at speed level 8 using 250 µl zirconium silicate beads (0.1 mm diameter, BioSpec, 11079101z). After homogenization, the beads were pelleted by centrifugation at 18000 × g (10 min) and pellet lysates were recovered from the supernatant.

Approximately 100 µl of each sample (25% of the total) was concentrated using SDS-PAGE gels (5% stacking and 12% resolving) at 50 V for 40 – 50 min. BSA was used as a reference marker. After electrophoresis, the gels were stained with Coomassie Blue and scanned. The bands with concentrated proteins were excised and digested with trypsin using an automatic device (DigestPro MS, Intavis), as previously described (Casanovas et al., 2009).

2.3. LC-HRMS/MS and database search

The LC-HRMS/MS system consisted of an Agilent 1200 Series Gradient HPLC (consisting of a capillary nanopump, binary pump, thermostatic micro injector, and micro-switch valve) coupled to an Orbitrap-Velos High-Resolution Mass Spectrometer (ThermoFisher) equipped with a nanoESI ion source.

For the analysis, the tryptic digests of the sample extracts were evaporated until dry and re-dissolved in 50 µl of 0.5% TFA 5% methanol with gentle agitation in a Thermomixer (5 min, at 22 °C, 900 rpm). Five microliters of this solution were injected into the HPLC system.

Separation was performed on a 15-cm long, 100 µm i.d. C18 column (Nikkyo Technos Co.) preceded by a C18 preconcentration cartridge (Agilent Technologies). Separation was done at 0.4 µl/min using a 120-min gradient from 3-35% solvent B (solvent A: 0.1% formic acid, solvent B: acetonitrile 0.1% formic acid).

The Orbitrap-Velos was operated in positive ion mode with a spray voltage of 1.7 kV. Spectrometric analysis was performed in data-dependent mode (Gao & Yates, 2019), acquiring a full scan followed by 10 MS/MS scans of the 10 most intense signals detected in the MS scan. Full MS (range 400-1650) was acquired in the Orbitrap with a resolution of 60.000. MS/MS spectra were obtained in a linear ion trap.

MS/MS spectra were searched using SEQUEST (Proteome Discoverer v2.5, ThermoFisher) with the following parameters: peptide mass tolerance, 20 ppm; fragment tolerance, 0.8 Da; enzyme, trypsin, and allowance of up to two missed cleavages; dynamic modification, methionine oxidation (+16 Da); and fixed modification, cysteine carbamidomethylation (+57 Da). Searches were performed using UniProt (rev. 10-21). Final results were filtered using peptide rank 1, peptide confidence high (0.1% FDR), and search engine rank 1 (Nesvizhskii, 2010).

The MS analysis of the particulate samples was performed as described above, except for the use of a different chromatographic system that consisted of a C18 column trap (nanoEaseTM M/Z Symmetry C18 100 Å, 5 µm, 180 µm × 20 mm, Waters Corporation) connected to a 25 cm long, 75 µm i.d. C18 column (nanoEaseTM M/Z HSS C18 T3 100 Å, 1.8 µm, 75 µm × 250 mm, Waters Corporation). The separation was done at 0.4 µl/min in a 180-min gradient from 2 – 40% solvent B (solvent A: 0.1% formic acid, solvent B: acetonitrile 0.1% formic acid). The HPLC system was composed of a µBinary Solvent Manager, µSample Manager, and Trap Valve Manager from the Acquity UPLC M Class (Waters Corporation).

MS/MS spectra of the particulate extracts were searched using SEQUEST (Proteome Discoverer v1.4, Thermo Fisher) software with the same parameters as above. The search included a reanalysis of the MS data from the soluble extracts and extracts from the material found in the polymeric probes used in our previous work (Carrascal et al., 2020). The database used for searching was UniProt (rev. 08-22). The mass spectrometry proteomics data have

been deposited to the ProteomeXchange Consortium via the PRIDE (Jarnuczak & Vizcaíno, 2017) partner repository with the dataset identifier PXD038781.

2.4. Data treatment and semiquantitative analysis

Overall descriptions of the soluble wastewater proteome were obtained from the protein identification output of a Protein Discoverer Multiconsensus analysis, including all protein identifications from the different sites and campaigns. For discussion purposes, only proteins assigned as master proteins, with at least two peptides pointing to them, were considered. Estimation of the relative abundance of proteins was based on normalized spectral counts (NSCs). NSCs correspond to the total peptide sequence matches (PSM) obtained using Protein Discoverer and normalized to the mass of the protein to consider that the number of tryptic peptides produced by a protein increases with its size, and thus also the total PSMs measured.

The comparison of the soluble and particulate proteomes of the material in this study and that of the material found in the polymeric probes (Carrascal et al., 2020) was performed as described above. They included all soluble extracts, two replicates of particulate samples from the Vic and Besòs WWTPs, and all the probe-derived sample analyses from the inlet of the WWTPs (site 1 of the three WWTP samples in our previous work). Owing to the high number of archives to process, a multiconsensus protein identification was performed considering five data groups: soluble data combined by campaign, combined particulate data, and combined probe data.

For the semiquantitative determination of proteins such as amylases and albumins, we selected peptides with an unambiguous match to the protein (no other proteins in the Protein Group) and with at least two PSMs. Unselected peptides pointing to a protein in the unambiguous set were then recovered and added to that set. Protein areas were calculated as the sum of all selected peptides pointing to the protein and were normalized to the wastewater flow measured at the WWTP inlet when the sample was collected.

To ensure reproducibility and traceability, Protein Discoverer output data were processed and documented using Jupyter notebooks and Python.

3. RESULTS AND DISCUSSION

3.1. Wastewater Proteome

The non-targeted shotgun proteomics study of these water samples allowed us to identify a total of 4318 peptides (1% FDR, >1 PSM) that indicated 827 proteins (1% FDR, >1 peptide) (see supplementary material 4.2_Soluble_Identifications.xlsx). The most abundant proteins were animal amylases and albumins (Table 6). Based on NSCs, eukaryotic proteins (mainly from mammals and birds) are the major components of wastewater, followed by bacterial proteins. Small amounts of viral proteins were also detected. Human proteins constituted 46% of the collection, followed by pig, chicken, cow, and rodent proteins (14, 9, 8, and 7%, respectively). Plant proteins made up >50% of all non-Chordata eukaryotes (Figures 18−20, Table 6).

It is noteworthy that due to the remaining uncertainty in the spectrum-peptide matching process and the limitations of the protein inference approaches in shotgun proteomics, some artifactual assignments are expected in our protein collection, which has not been manually curated. This may be, for example, the case of the suspicious assignment of a pancreatic α-amylase protein (AMYP) to the common ostrich (*Struthio camelus*, STRCA) protein, as shown in Table 6. Although the assignment cannot be discarded a priori (there are ostrich farms in Catalonia), this is the case for a protein assigned on the basis of two forms of the same unique peptide sequence (oxidized and not oxidized) with a 94% identity with the human sequence. The probability of an incorrect assignment is reduced as the number of unique peptides pointing to a specific protein increase.

The most represented phylum in the collection was Chordata, and the major contributors to the proteins in this phylum were humans and livestock. Humans were represented by 243 proteins. The most abundant human proteins were pancreatic enzymes headed by α-amylases, making these proteins the main markers of human presence in wastewater. Several blood proteins (albumin, immunoglobulins [Igs], and complement proteins) and skin-derived proteins were also present in notable amounts. A DAVID gene ontology analysis (Sherman et al., 2022) revealed several enriched functional terms such as those related to the immune response (Igs, calprotectin, lactoferrin, lipocalin, and dermcidin) or the anti-inflammatory response (meprin A, orosomucoid, and the serpin family). The most abundant non-human proteins detected in wastewater were albumin from cattle (and ovalbumin from poultry). Albumin from commensal rodents (rats and mice) was one of the most important proteins detected.

Table 6. The 20 most abundant proteins in the wastewater samples. STRCA: Ostrich; FELCA: Cat.

Access	Protein name	Entry name[1]		Coverage [%]	# Peptides	# Protein unique peptides	# NSCs
		Gene	Species				
P04746	Pancreatic alpha-amylase	AMYP	HUMAN	88	62	7	10924
P0DUB6	Alpha-amylase 1A	AMY1A	HUMAN	86	58	9	10470
P19961	Alpha-amylase 2B	AMY2B	HUMAN	88	60	1	9594
P08835	Albumin	ALBU	PIG	93	119	91	8846
P01834	Immunoglobulin kappa constant	IGKC	HUMAN	93	13	2	7712
P01012	Ovalbumin	OVAL	CHICK	75	32	22	7159
P02769	Albumin	ALBU	BOVIN	92	118	70	5480
P01846	Ig lambda chain C region	LAC	PIG	97	9	9	4773
P19121	Albumin	ALBU	CHICK	88	85	84	4393
P02768	Albumin	ALBU	HUMAN	90	87	14	4309
P83053	Pancreatic alpha-amylase	AMYP	STRCA	33	22	2	3013
P00687	Alpha-amylase 1	AMY1	MOUSE	23	15	2	2878
P0DOX7	Immunoglobulin kappa light chain	IGK	HUMAN	60	14	2	2786
P01009	Alpha-1-antitrypsin	A1AT	HUMAN	60	35	17	2675
P00690	Pancreatic alpha-amylase	AMYP	PIG	51	24	7	2627
P14639	Albumin	ALBU	SHEEP	88	71	28	2540
P00689	Pancreatic alpha-amylase	AMYP	RAT	28	14	2	2410
P07478	Trypsin-2	TRY2	HUMAN	63	12	8	2283
P07724	Albumin	ALBU	MOUSE	39	24	1	2173
P49064	Albumin	ALBU	FELCA	32	26	5	2172
P09571	Serotransferrin	TRFE	PIG	94	104	91	2159

[1] UniProtKB/Swiss-Prot entry name. The two terms of the entry name (gene_species) have been separated for convenience.

Overall, our data showed two main sources of proteins in wastewater: excreta (urine and feces) from humans, and blood and other residues from livestock.

3.2. Wastewater Proteome is Compartmentalized

In an earlier study in which we used polymeric probes to capture proteins from wastewater, we found high levels of bacterial proteins in the samples (Carrascal et al., 2020) (see supplementary material 4.2_Sol+Part+Probes_Identifications.xlsx). In contrast, in the filtered wastewater, bacterial proteins were relatively minor components, and the most abundant proteins differed from those found in the probes (Figures 18 and 21).

These differences can be explained by the formation of biofilms in the probes that become enriched in bacterial proteins. In addition, the fact that the samples were passed through a 0.2 µm filter suggests that most of the bacterial protein mass was transported in bacterial cells. A preliminary analysis of the particulate fraction of the wastewater samples confirmed this finding (see supplementary material 4.2_Sol+Part+Probes_Identifications.xlsx). We analyzed material from two different WWTP sites: a major urban area (Besòs) and a rural community (Vic). In both cases, bacterial proteins were the major components (Figure 18, bottom), although the distribution of the species was different. Although the bacteria-eukaryote distribution in the wastewater particulate was found to be similar to that of the polymeric probes, these two fractions showed some differences in the dominant proteins found in each of them (Figure 21, insert). A more in-depth study is required to confirm whether these differences are due to the potential selectivity of the polymeric probes or simply reflect the different origins of these samples.

Another interesting example of protein compartmentalization is human elastase 3A (CL3A), a protein that we are considering as a potential biomarker for the human population. Human elastase is a well-known component of sewage and WWTP sludge (Kuhn et al., 2011). This is a recalcitrant protein with a high concentration in feces (Westgate & Park, 2010), which we have described as the major component retained in our polymer probes. In contrast to most other Chordata proteins, which were located preferentially on the filtrates, CL3A was found in higher relative amounts in the particulate fraction, where it was the major component (second position in the particulates from Vic and Besòs). Other proteins found in large amounts, mostly in particulate fractions, were keratins. Whereas EFTU and 60 kDa heat shock proteins (CH60) would be the most abundant and pervasive markers of bacterial presence, for mammals this position would be occupied by keratins.

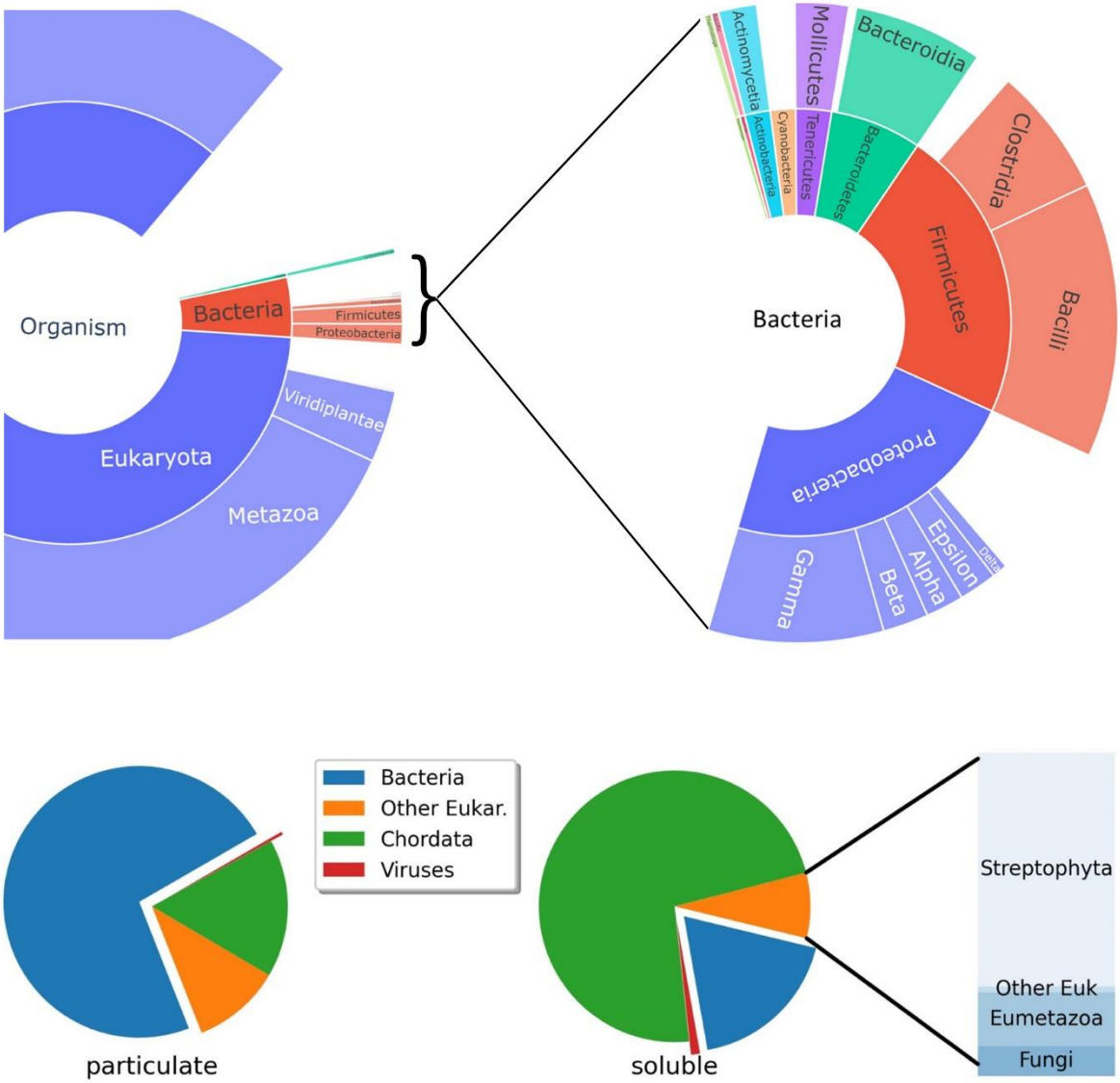


Figure 18. Distribution of Bacteria and Eukaryotic proteins in the soluble fraction of wastewater and comparison with the particulate fraction (bottom). The sunburst graphs were prepared from the Unipept analysis of all peptides identified in the soluble fraction of wastewater samples. Comparison of particulate and soluble fractions was obtained from the multiconsensus analysis of all available samples, as described.

Although we aimed to describe the characteristics of filtered wastewater in this work, these preliminary results on the particulate fraction reveal that this is only a partial view of the full wastewater proteome. Comprehensive and in-depth analysis of this proteome is, however, complicated. Our attempts to directly analyze the full wastewater composition (without the previous separation of the soluble and particulate fractions) were unsuccessful, likely due in part to inefficient trypsin digestion, probably caused by the interference of other compounds in the water. Therefore, we believe that parallel, separate analyses of the different wastewater

compartments using sample fractionation methods, as those described here, may be the best strategy for a complete description of the wastewater proteomes. Knowledge of the protein distribution between these two compartments will further aid in the future development of methods to monitor potential biomarkers.

3.3. Semiquantitative Analysis of the Wastewater Proteome Characterizes Human Activity Around the WWTPs

Wastewater collected from a community reflects its population and domestic and industrial activities (Daughton, 2018; Rice & Kasprzyk-Hordern, 2019). To test the potential of the wastewater proteome as a potential biomarker, we performed a comparative analysis of the protein composition at different collection sites and estimated the protein abundances considering the corresponding wastewater inflows.

For this purpose, protein semiquantitative data were calculated from the sum of the areas of their unique peptides in the ion chromatograms. Only those proteins unambiguously assigned by the search program (at least with a unique peptide pointing to it and with no other homologous members in the protein group) were considered. This procedure causes a bias in the set of proteins represented because those highly conserved between species produce only a few unique peptides or none in the worst case. Thus, highly conserved proteins are at risk of not being included in the set of quantified proteins. In return, this strict filtering prevents the contribution of peptides from similar proteins to the area of the one measured. As discussed previously, incorrect species assignment cannot be discarded for less represented species. This may be the case for the two proteins assigned to *Pongo abelii*, a species with high homology to humans, or the assignments to *Danio rerio* and *Dictyostelium discoideum*, two species that frequently appear in proteomic searches because of the assignment of proteins from other related species, but with genomes that have not been annotated to the same degree as those of these model species.

Using this approach, we obtained reliable quantitative data from a set of 489 proteins (between 350 and 401 proteins could be quantified in the different samples), representing a total of 112 species (26 from more than one protein Table 7).

Table 7. Species represented by at least two proteins in the set of proteins selected for semiquantitative analysis.

Species	Proteins	Species	Proteins
Homo sapiens	169	*Capra hircus*	3
Sus Scrofa	59	*Felis catus*	2
Bos taurus	55	*Prunus dulcis*	2
Gallus	37	*Bacteroides vulgatus*	2
Rattus norvegicus	9	*Bacillus amyloliquefaciens*	2
Mus musculus	9	*Lachnospira eligens*	2
Triticum aestivum	8	*Equus caballus*	2
Oryctolagus cuniculus	7	*Danio rerio*	2
Canis lupus familiaris	7	*Clostridioides difficile*	2
Solanum tuberosum	5	*Pongo abelii*	2
Ovis aries	4	*Malus domestica*	2
Pseudomonas aeruginosa	4	*Solanum lycopersicum*	2
Hordeum vulgare	3	*Dictyostelium discoideum*	2

The distribution of the major proteins in the different campaigns was highly variable (Figure 19). However, on average, they showed some characteristic traits that may indicate a relationship with the human population and industrial farming activities at each site, especially on the distribution profiles of pig, cow, and chicken proteins. A preliminary analysis of the data led us to select some protein groups (amylases, albumins, and Igs), with quantitative profiles that suggested that they could be potential biomarkers for further studies.

3.4. Amylases as Mammal Population Indicators

The most abundant protein in the wastewater soluble fraction was human pancreatic α-amylase. Rodent, pig, and chicken amylases were detected in lower amounts. Amylases are the major protein components in feces, together with elastases, and are present in minor proportions in urine (Candiano et al., 2010; Marimuthu et al., 2011). Pancreatic amylases and elastases are secreted in the pancreatic juice together with other lipases, nucleases, and proteases for the digestion of food. The main role of pancreatic enzymes in the intestine reflects their high stability against hydrolytic degradation. An example is pancreatic elastase,

which has been detected in WWTP sludge, evidencing its resistance to the wastewater environment and WWTP treatment. In blood and sera, α-amylase maintains its enzymatic

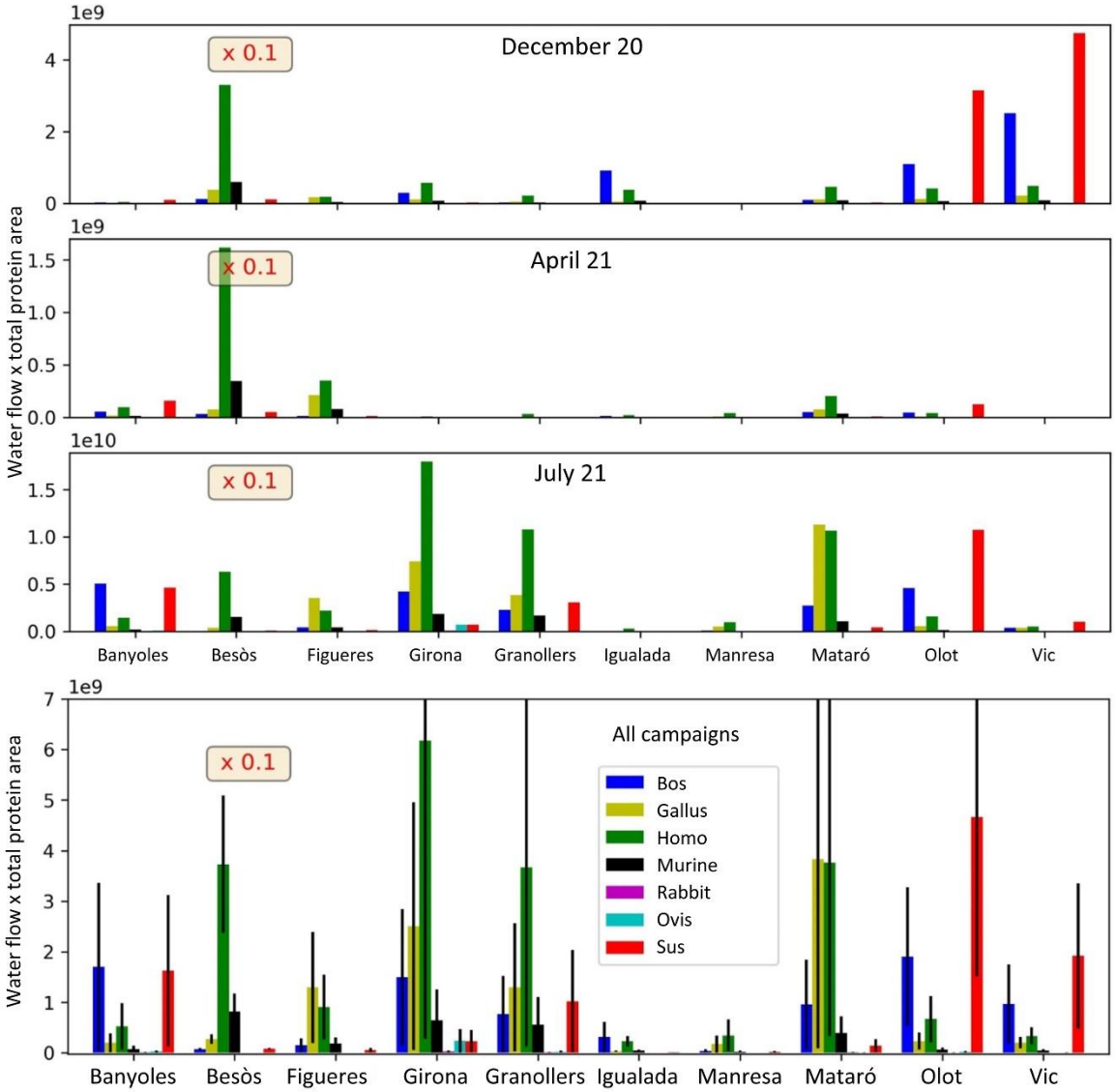


Figure 19. Distribution of proteins by species of origin in the different sampling sites by campaign (top) and with all campaigns combined (bottom).

activity unaltered for at least a week at room temperature (Foo & Rosalki, 1986), so that protein sequence information can be expected to be preserved still longer. Thus, due to their high abundance in wastewater, their probable resistance to protease action and the species-specific information carried in their sequences, amylases may be potential markers of human population and, as such, a potential tool to normalize the abundances of other biomarkers. One important uncertainty in WBE studies is the estimation of the number of inhabitants served by a WWTP (Daughton, 2012; Rico et al., 2017; Hsu et al., 2022). The wastewater physicochemical parameters frequently used for this purpose [chemical oxygen demand (COD), biological oxygen demand (BOD), total nitrogen or phosphorus] are highly unspecific

and census does not reflect population dynamics and is often outdated (Rico et al., 2017). In recent years, molecules excreted in human feces or urine have been evaluated and used as markers of population. Most of them are small molecules, such as creatinine, cholesterol, 5-hydroxyindoleacetic acid, caffeine, prostanol, or drugs widely used by the population. Human-specific protein forms such as amylases have the advantage of being virtually free of the contribution from other non-controlled exogenous sources and thus to provide more accurate measurements.

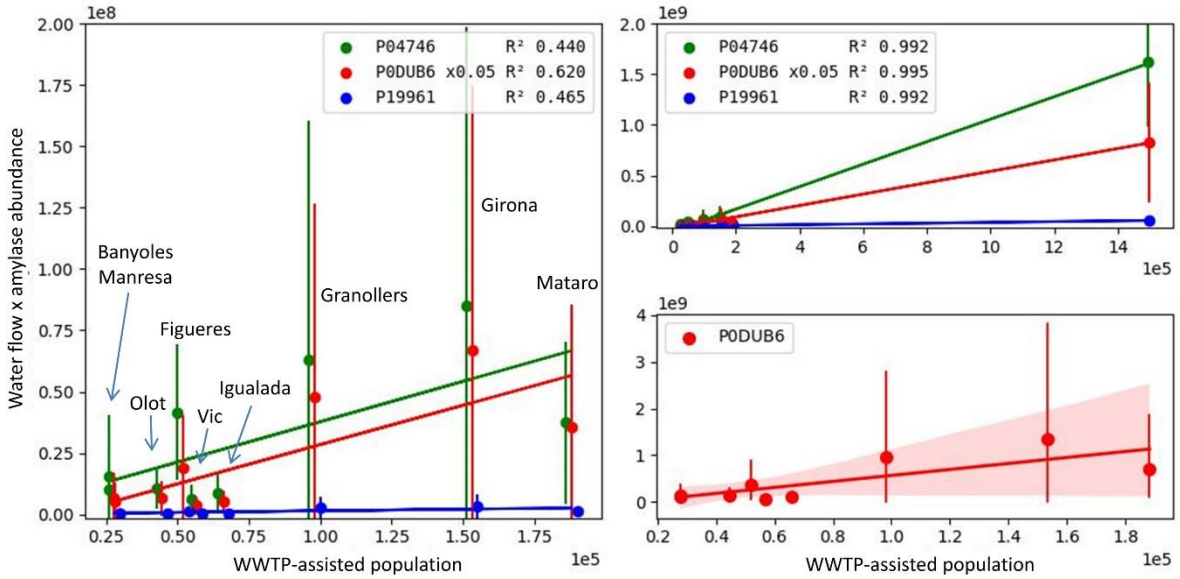


Figure 20. Human amylases represented versus the population assisted by the corresponding WWTP. Images include errors and regression lines obtained for the three amylases measured at all sites, except Besòs (left) and all sites (right, top), as well as the correlation for the major amylase (alpha-amylase 1A, P0DUB6) and the 95% confidence interval zone (Besòs not included) (right, bottom).

Amylase presence in wastewater greatly increased when comparing cities with large populations (Barcelona, Besòs WWTP) with small- and medium-sized cities (Figure 20, top-right). However, although a trend was observed between the amylase levels and the estimated population served by the WWTPs, the correlation was poor (Figure 20, left and bottom-right). Unfortunately, we did not have data from WWTPs serving populations in the middle range between Mataró and Besòs to obtain a more precise view.

The degree of correlation between population and amylase levels may be affected by inaccuracies in the population data. The population data used correspond to the available official figures (Table 5). However, this parameter is not exempt from uncertainty, as the actual population may be subjected to large fluctuations over time (for example, seasonal tourism) or may not reflect the actual population channeling their sewage into the WWTP. Another

factor to consider is the robustness of the biomarker as different protein degradation dynamics in the specific biological and physicochemical environments of the different zones could affect the measured levels of amylase in water. Further research is required to deconvolute these factors.

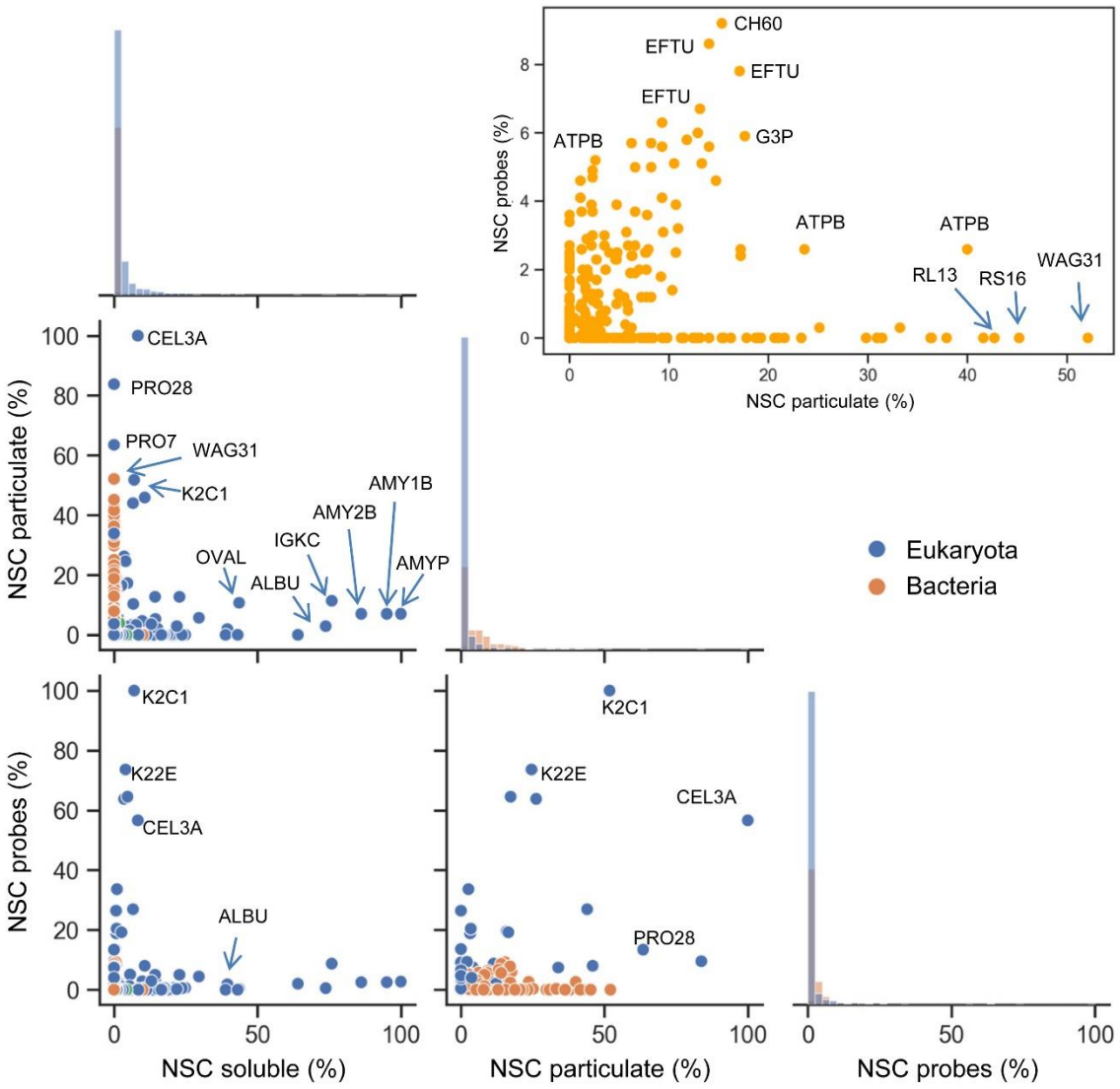


Figure 21. Comparison of the Eukaryote (blue) and Bacterial (orange) proteins in the wastewater soluble and particulate fractions and those found bound to the polymeric probes described in a previous work (Carrascal et al., 2020). Insert: details on the bacterial proteins. Abundance estimation (#NSC) was normalized to the most abundant component in each fraction. Diagonals are the histograms of the protein distribution in each fraction (soluble, particulate, and probes, from top-left to bottom right).

Since factors like changes in water flow due to precipitations that affect the levels of amylases in wastewater would be common to other proteins in wastewater, amylases could be useful to correct for these factors the amount of other human protein biomarkers or the amount of these pancreatic enzymes from other species. A potential application of great interest would be their use to monitor rodent populations in urban areas. Rat pests are a human health hazard because of the diseases they can transmit through the bacteria that infect them, and the transmission of fleas, ticks, and mites. In addition, they threaten the integrity of infested structures, and once established, their elimination is difficult. In large cities, rats live in the sewers. If no control action is taken, these animals can live between 2 and 3 years and procreate up to five times a year with an average of 4−8 offspring; thus, their number varies rapidly over a few months (Feng & Himsworth, 2013). Various strategies are currently being used to monitor these pests, generally based on animal counts and extrapolation to the total population (Byers et al., 2019). The number of animals in a large city is often referred to as the number of rodents per inhabitant. For example, it is estimated that in the city of Barcelona, there may be one rat per eight inhabitants (Ansede, 2018), and some estimates speak of up to 0.25 per inhabitant in the city of New York (Auerbach, 2014). However, there is no standardized method to determine their numbers, estimate population density, or understand population dynamics.

As in humans, rat amylases are secreted into the pancreas and excreted mainly in feces; therefore, their quantification in wastewater relative to human amylases may allow the detection of rodent population peaks. Murine amylases were found in water in 100−500-fold lower amounts than human amylases. The ratio of murine to human amylases varies with the site and is generally higher in small cities in predominantly rural areas (Banyoles, Olot, and Vic) and smaller in large urban industrialized areas (Besòs/Barcelona, Mataró) (Figure 22). Interestingly, a peak in the murine-to-human amylase ratio was observed in the July campaign in Igualada. This sample also showed a marked increase in the rat-to-mouse amylase ratio despite the fact that in all other samples the ratio remained relatively constant (Figure 22). Whether this could reflect an increase in the rat population at this site is difficult to determine from a single event but these preliminary findings have prompted us to conduct additional studies that are now underway for this particular application.

Currently, work is underway to develop a targeted-MS method that allows a more precise quantification of human and rodent amylases in wastewater. Knowledge of the best tracking subjects (those more abundant, unique, and species-specific peptides) should facilitate future development of immunoaffinity-based sensors for this purpose. The validation of this approach would be the source of new tools for pest surveillance that can provide integrated information

on the area of origin and conduction of the waste in parallel with other more local methods already in use (photo trapping, rat traps...).

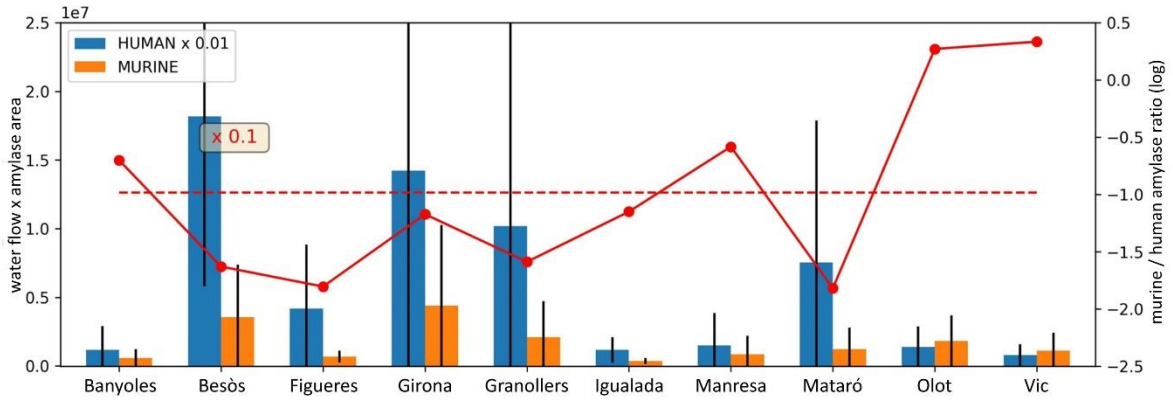


Figure 22. Murine (rat and mouse) and human amylases in the different sites. The red line indicates the murine to human ratio, and the red dashed line marks the mean ratio.

3.5. Albumins as Livestock Industry Markers

Albumins were found in high quantities in the analyzed water samples. The presence of albumin in wastewater probably results from industrial discharge of animal blood. Serum albumins are 60−70 kDa proteins with a high sequence homology among many primates (>90% identity) and other mammals (>50%). Differences between homologous albumins are widely distributed throughout their chains, resulting in significantly different sets of tryptic peptides after enzymatic digestion. Thus, considering albumins from humans, livestock, poultry, common human pets (cat and dog), and murids (rat and mouse), there are always between 24 and 42 different peptides potentially identifiable by our proteomics approach (>6 AA), which are unique to any pair of these species (Figure 23). Considering the full set, any of these albumins would theoretically yield between 21 and 38 unique canonical tryptic peptides, which can allow species-specific identification and quantification.

Feces and blood disposed of by slaughterhouses are of great concern as water pollutants (Savin et al., 2022; Bustillo-Lecompte & Mehrvar, 2015). Albumin is the largest protein component in sera (approximately 50−60% weight in humans). This high abundance and the differences in albumin sequences between species open the possibility of developing MS-based strategies for specific monitoring of the levels and sources of biological contamination downstream of the release point and at the WWTP. This could be a powerful monitoring tool not only for environmental studies assessing the status of a water body but also for regulatory agencies in the surveillance of controlled and uncontrolled discharges of animal residues in rivers and wastewater systems. Currently, occasional discharge can be indirectly detected by

routine monitoring of the organic load content in wastewater (for example, BOD, COD, and TOC); however, these methods do not provide information on the contributing molecules or their origin (Bustillo-Lecompte & Mehrvar, 2015).

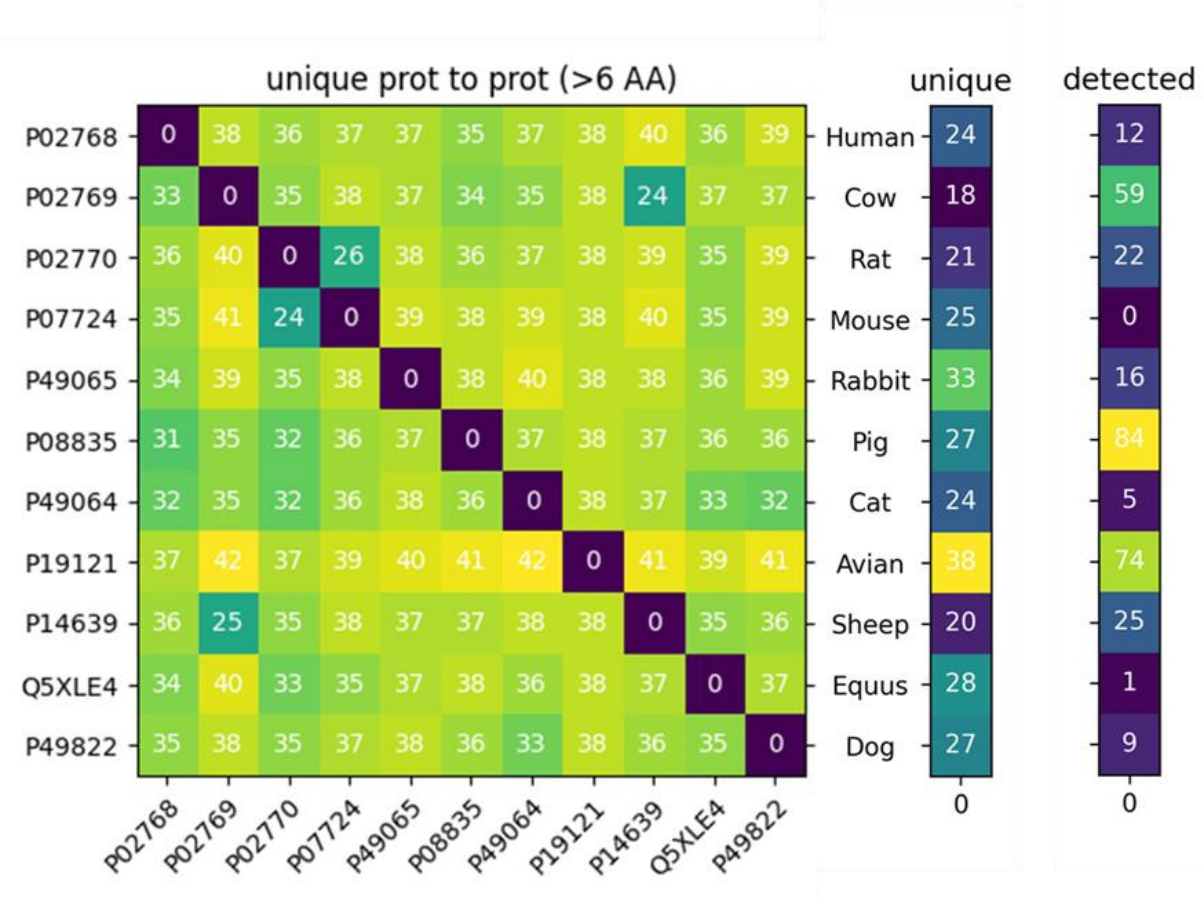


Figure 23. Number of tryptic sequences that are different between any pair of the represented albumins (left), that are different from any other (unique in this albumin set) (center) and unique sequences that were detected in our samples (right). Calculation of unique sequences considers only canonical tryptic peptides, whereas the experimental data includes sequences with missed cleavages.

In our study, we identified between 1 and 84 unique peptides for the different albumins considered above, which are, in some cases, higher than the expected canonical tryptic sequences because they also include peptides with missed cleavages (incompletely digested peptides). The number of unique peptides detected was highly dependent on the concentration of a specific protein in the sample. Thus, pig albumin is more representative, whereas no unique mouse peptide passed the data treatment quality filters (Figure 23).

As albumins are the major contributors to animal protein mass in the wastewater proteome, the albumin profile distribution was highly similar to that shown for the total protein distribution (Figure 19). In concordance, the profiles of farm animal albumins were found to be significantly

different among the sites (Figure 24). To determine the correlation between these albumins and the presence of the corresponding species at a given site, we compared the number of official livestock units in each area with measured albumin values (Figure 25). We found that livestock units and albumin abundance were significantly different from each other; for example, wastewater samples from areas with a relevant number of pig farms, such as Figueres, Igualada, and Manresa, contained significantly low amounts of pig albumin. As animal blood and tissues are the major containers of albumins, these proteins probably mark the presence in the sewage of animal residues from the meat industry (for example, slaughterhouses), whereas livestock units reflect the number of animals raised in the region. For example, in Mataró, where the most important Spanish company in the poultry processing sector is located, there is a significant difference between the poultry livestock units and the measured Gallus albumin levels. Similarly, pig albumin appears to be the main albumin in Vic, Olot, and Banyoles, cities in which the pork industry is of great importance.

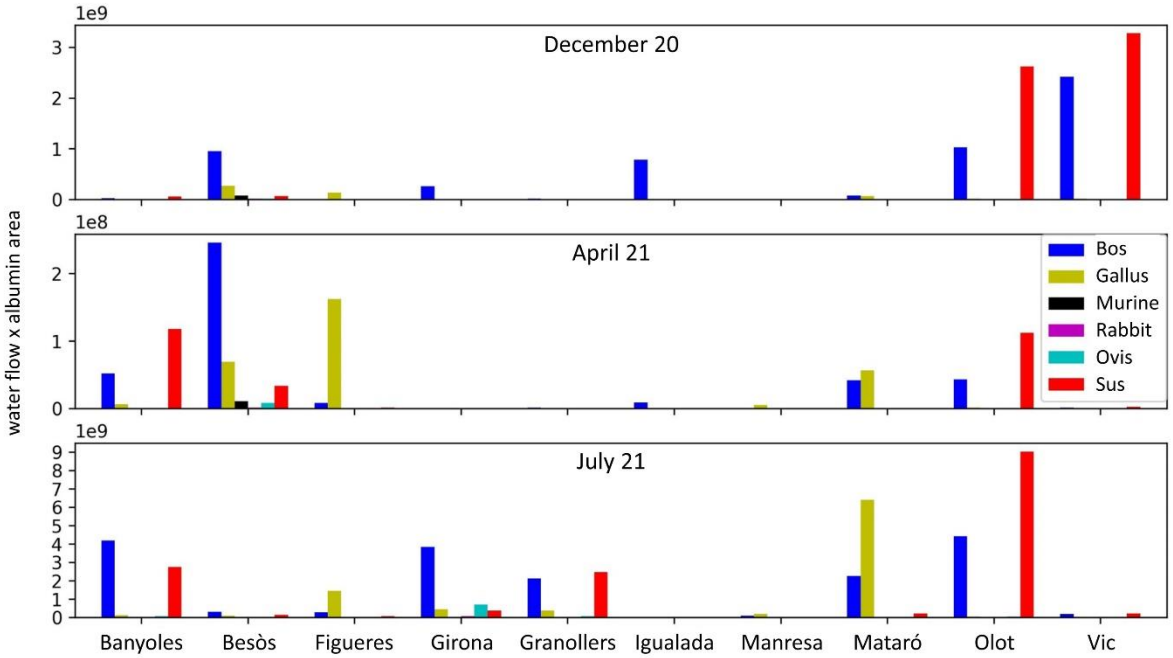


Figure 24. Albumin profiles from farm animals in the three campaigns.

3.6. Human Immunoglobulins

Another important family of proteins found in samples is human Igs. Recently, human Igs were proposed as health biomarkers, although their presence, distribution, and detectability in wastewater had not been assessed (Barceló, 2020). More recently, measurement of specific Igs in wastewater was proposed as a window for community serology and an ELISA method was developed in the context of COVID surveillance (Agan et al., 2022). Igs are large heterodimeric glycoproteins composed of two heavy and two light chains, each of which is a

combination of different variable and constant domains encoded by 176 genes. There are five Ig isotypes named based on their α, Δ, ϵ, γ, and μ heavy chains (IgA, IgD, IgE, IgG, and IgM, respectively), each containing one of the two classes of light chains (κ and λ). Both heavy and light chains were subdivided into highly homologous subtypes, each with a different entry in the UniProt database. This greatly complicates proteomic quantitation by measuring the areas of unique peptides. Thus, many of the Ig tryptic sequences identified in our samples indicated two or more different Ig sequence accessions; consequently, they were not selected for measurement. This led to a situation where we had no unique peptides to quantify the λ chain, or where the areas of the γ chain, which makes the major Ig in the blood, were relatively small and unreliable, as were calculated from a minor unique peptide. As multiple protein assignments of the identified peptides were always to subtypes of the same chain, we measured each Ig chain, considering all its subtypes together. This enabled us to measure three heavy chains and two light chains (Figure 26). Sequences pointing to the J-chain, a component of IgA and IgM, and variable sections of the heavy chains not related to a specific Ig were also measured.

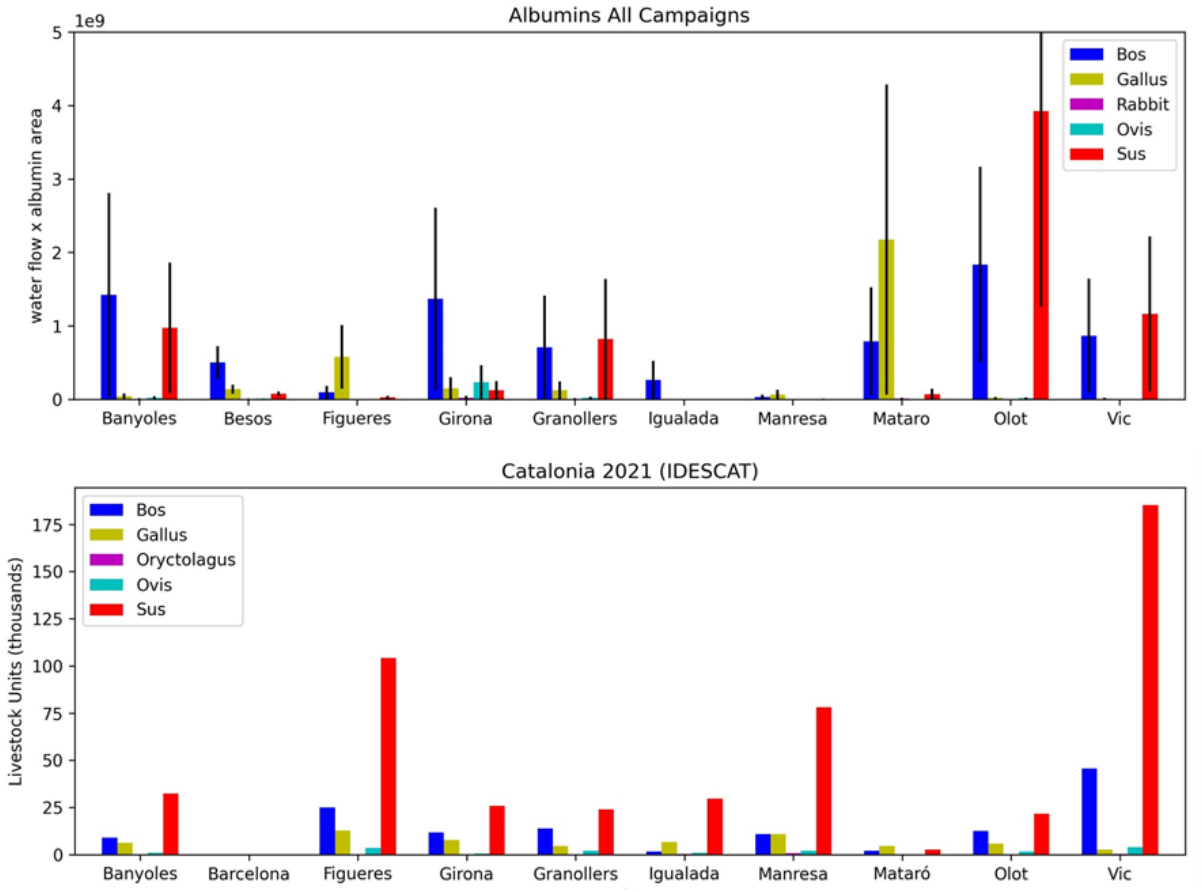


Figure 25. Comparison of the average albumin profiles with the livestock units in the county in which the WWTP is located (IDESCAT data for 2021, https://www.idescat.cat/).

The Ig chain profiles were similar among the different sites and through the different campaigns (not shown), whereas the areas varied greatly, likely correlating with the human population served by the corresponding WWTP site. As observed in the profiles for other proteins, Ig abundance changed significantly between campaigns, as reflected in the length of the corresponding error bars (Figure 26). Girona showed the highest difference, with an 8105-fold change between two campaigns, whereas Besòs (Barcelona) had the lowest maximum difference, with a fivefold change. As in the case of amylases, these variations could be partly a reflection of changes in the population at the site. This hypothesis is supported by the profiles of the human-Igs-to-amylases ratio, which, by comparison, were relatively constant between sites and campaigns (Figure 26, right). Thus, the greatest variation in the Igs/amylase ratio was 11-fold (Mataró), and the smallest maximum difference was only 1.4-fold (Besòs). On average, a 932 ± 2396 Ig maximum fold change was calculated between campaigns, whereas this average was 4 ± 3-fold for the Ig/amylase ratio, >2.5 orders of magnitude lower. These results further support the interest of abundant human proteins such as amylases to normalize the abundance of other human proteins in wastewater.

In summary, our shotgun analysis reveals the high abundance of antibody molecules in wastewater, and the capability to discern between different Ig types and chains as well as to determine their profile, thus providing new knowledge for further development on SCIM methods based on these molecules.

3.7. Wastewater Origin can be Differentiated by a Small Group of Biomarkers

We have shown that wastewater proteomes exhibit protein profiles that are characteristic of the site and time frame. Protein profiles contain information on human activities in a given area, as revealed when considering livestock activities. This opens a window for monitoring diverse activities or site statuses through the determination of these protein patterns. Based on these premises, we tested the possibility of automatically differentiating wastewater origin by employing its proteome profile. For this purpose, we performed different classification and clustering analyses, which produced poor outputs (Figure 27, top left). Wastewater is a highly complex matrix, and many of its components contribute little to distinguishing features that facilitate discerning between samples. Instead, they introduce noise, acting as confounding factors in the classification. In contrast, a linear discriminant analysis (LDA) using the albumin subproteome showed clear differentiation over the first component of the sites with dominant poultry farming (Figueres, Mataró, and Manresa) from the others. The latter, in turn, are

distributed along the second component, differentiating those with a predominance of cattle from those with a predominance of pigs (Figure 27, top-right).

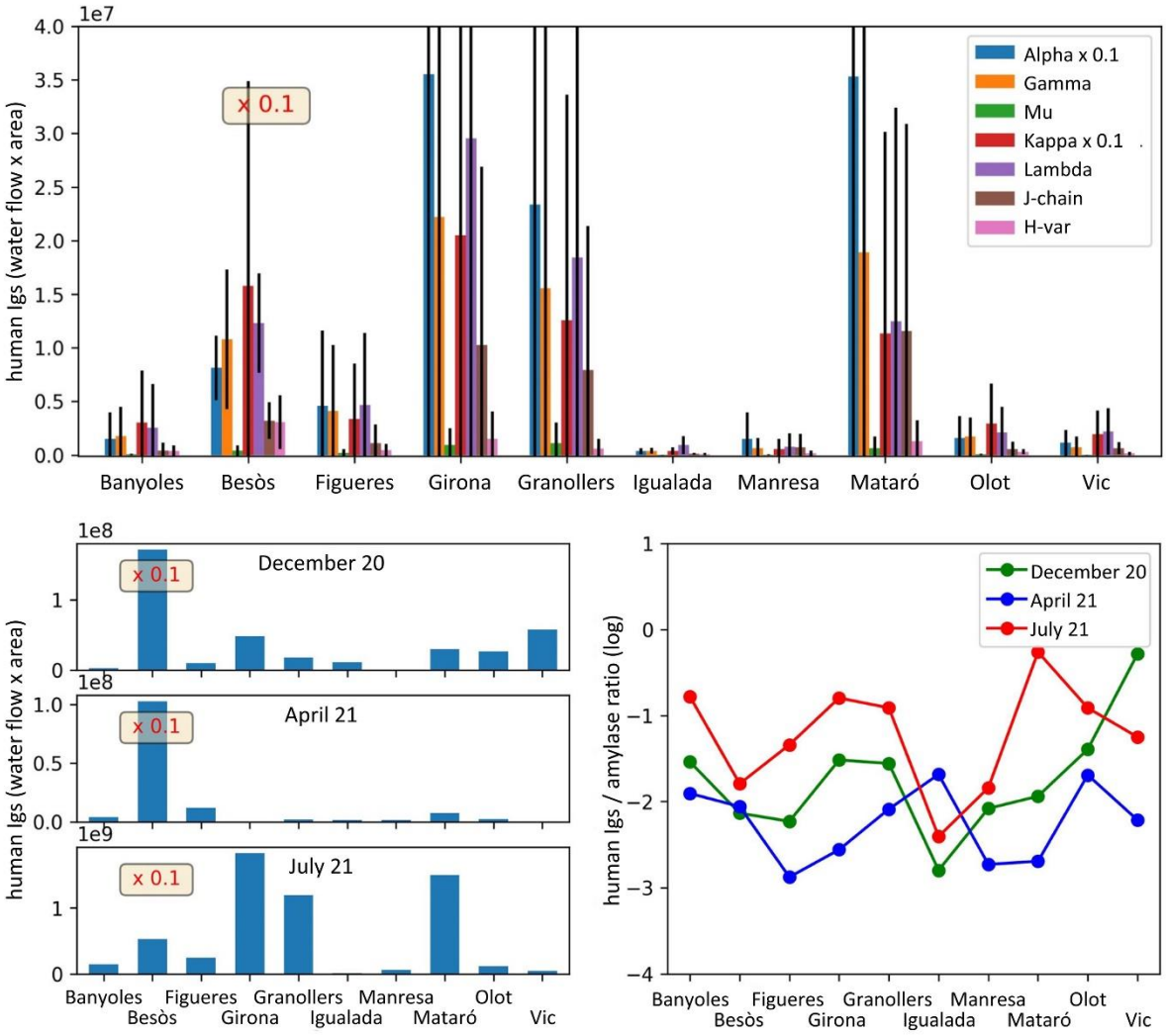


Figure 26. Distribution of human Igs in wastewater from the different municipalities (top), total human Igs abundance per site and campaign (bottom left), and human Ig/amylase ratios at the different sites and campaigns (bottom, right).

To optimize the classification, we used a protein set derived from the genes represented by the proteins with higher loads in discriminant factor analysis. This set was composed of 24 proteins expressed by ALB, SERPINB14, SERPINA1, SERPINA3-1, AMY1A, Amy2, TF, and Alpi genes in humans and other animals (Table 8). Analysis of this set produced a significantly more resolved separation of different clusters (Figure 27, bottom).

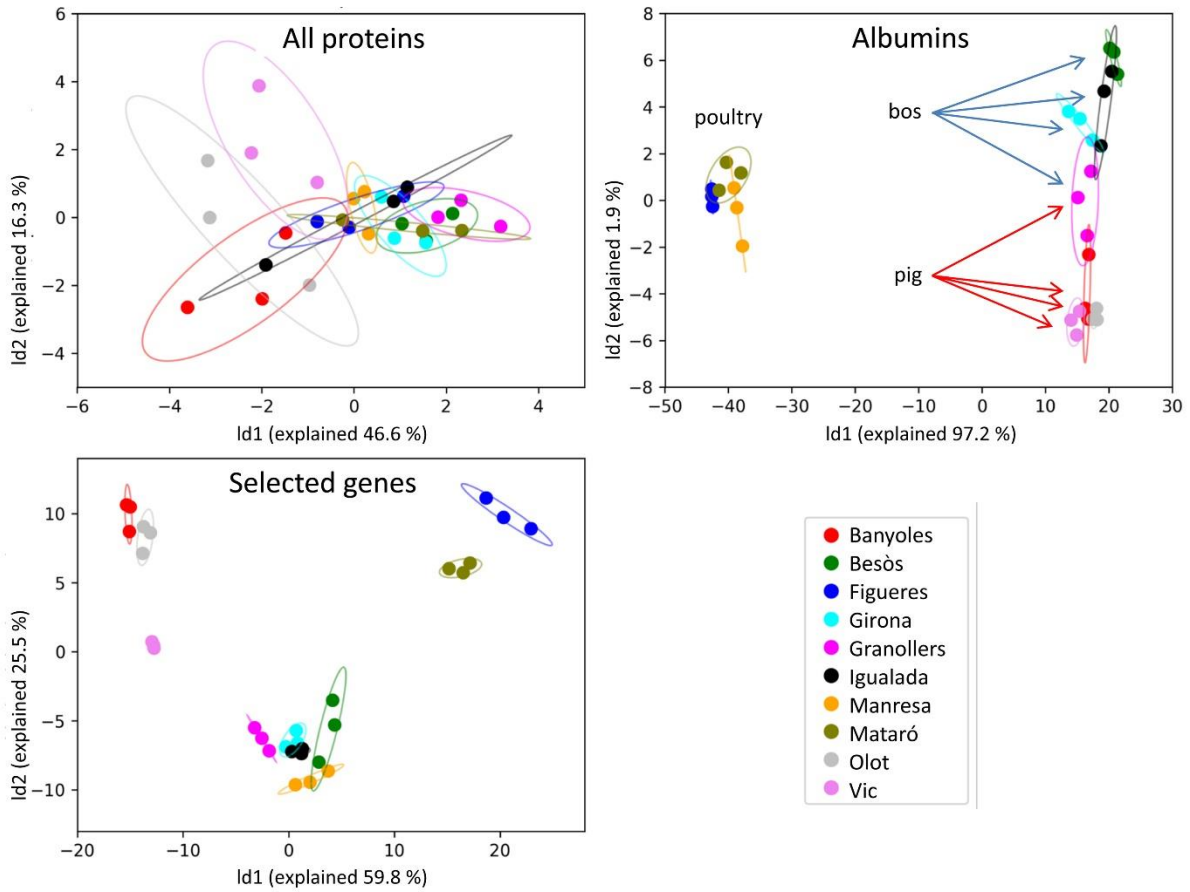


Figure 27. LDA of the full proteome profiles (top-left), the albumin profiles (top right), and of a pre-selected group of protein markers (bottom).

In summary, the present study greatly extends the knowledge we achieved in the comprehensive characterization of the wastewater proteome reported in our preliminary research (Carrascal et al., 2020). Here, we focused on the potential relevance of these protein profiles as new SCIM tools. To this end, analytical methods based on LC-HRMSMS shotgun proteomics were developed for both the dissolved and particulate materials contained in wastewater samples collected at the inlet of 10 WWTP serving municipalities in a broad range of population sizes.

Our data provide a comprehensive description of wastewater proteins, their distribution among different organisms, and a semiquantitative analysis of many of them. The data presented encompassed both prokaryotic (bacteria and, to a lesser extent, viruses), non-Chordata (plants), and Chordata eukaryotic organisms (including birds, mammals, and humans), which notably expands the scope of previous studies performed in wastewater and sludge, specifically focusing on the bacterial proteome (Westgate & Park, 2010; Zhang et al., 2019).

Table 8. Wastewater-origin discriminant proteins used for LDA-supervised classification.

Accession	Name	Species	Gene
P02769	Albumin	*Bos taurus*	ALB
P0DUB6	Alpha-amylase 1A	*Homo sapiens*	AMY1A
P08835	Albumin	*Sus scrofa*	ALB
P19121	Albumin	*Gallus*	ALB
P02768	Albumin	*Homo sapiens*	ALB
P01012	Ovalbumin	*Gallus gallus*	SERPINB14
P09571	Serotransferrin	*Sus scrofa*	TF
P00689	Pancreatic alpha-amylase	*Rattus norvegicus*	Amy2
P01009	Alpha-1-antitrypsin	*Homo sapiens*	SERPINA1
P02789	Ovotransferrin	*Gallus gallus*	TF
Q29443	Serotransferrin	*Bos taurus*	TF
P02787	Serotransferrin	*Homo sapiens*	TF
Q9TTE1	Serpin A3-1	*Bos taurus*	SERPINA3-1
P14639	Albumin	*Ovis aries*	ALB
P49064	Albumin	*Felis catus*	ALB
P49065	Albumin	*Oryctolagus cuniculus*	ALB
P50447	Alpha-1-antitrypsin	*Sus scrofa*	SERPINA1
P49822	Albumin	*Canis lupus familiaris*	ALB
P15693	Intestinal-type alkaline phosphatase 1	*Rattus norvegicus*	Alpi
P34955	Alpha-1-antiproteinase	*Bos taurus*	SERPINA1
P27425	Serotransferrin	*Equus caballus*	TF
P12725	Alpha-1-antiproteinase	*Ovis aries*	SERPINA1
A6YF56	Albumin	*Mesocricetus auratus*	ALB
Q5XLE4	Albumin	*Equus asinus*	ALB

We describe two main differential sources of proteins: excreta (urine and feces) from humans, and blood and other residues from livestock. The results highlighted significant differences between the proteomes in the soluble (filtered) phase and the particulate material, dominated by Chordata and bacterial proteins, respectively. Our findings also provide new insights into

the wastewater proteome that allow pointing out the possible practical use of some potential bioindicators in relation to wastewater-based environmental monitoring and WWTP management. Some relevant examples include amylases for mammalian population monitoring (applicable, for instance, to rodent pest surveys) and albumins as indicators of the cattle processing industry. Finally, in our previous work, we noted the presence in wastewater of endogenous human molecules, which are known disease biomarkers. Although we did not focus on human epidemiology, this study provides useful additional information on the presence of these and other endogenous human molecules of possible interest for WBE. The requirements that a protein must meet to be used as biomarker in WBE have been discussed in depth elsewhere (Rice & Kasprzyk-Hordern, 2019) including a well-defined disease−biomarker correlation, their excretion in high amounts and their stability both in vivo and in the wastewater media. Currently, the number of potential candidates is still small and none has been demonstrated yet (Rice & Kasprzyk-Hordern, 2019). Some limitations of the proteomics approach such as the need for specialized equipment and trained personnel have likely contributed to the situation. However, we can hope that once a potential biomarker is deemed worthy of further investigation, methods other than MS can be used for further validation and large-scale application.

Collectively, our prospect of the wastewater proteome is far from complete and raises new, unexpected scientific questions about the observed protein profiles. This is a consequence of our still limited knowledge about the numerous factors involved in protein dynamics along their route from the emission site to the sampling site as well as the actual emission rates of these proteins over time. Still, the enormous potential of proteins as health and environmental biomarkers compels an exhaustive characterization of possible confounding factors in order to develop accurate, robust applications for these molecules.

In conclusion, we have demonstrated for the first time the feasibility of wastewater proteome mining using modern proteomic technologies and have provided a protein database of value for future SCIM studies. We have shown that proteins in wastewater carry unique and specific information about their origin and we anticipate that these characteristics will open new avenues for the future development of new applications for environmental surveillance and monitoring.

4.3 FROM SOURCE TO STREAM: EVALUATING WASTEWATER TREATMENT PLANT PERFORMANCE

Objective 3: Assessment of the efficiency of wastewater treatment by monitoring samples at WWTP influent, effluent and receiving waters.

Treatment plants are built according to the population size to which they serve to enhance the efficiency in the removal of the contaminants. The efficiency of the removal of contaminants is controlled by classical parameters such as the chemical and biological oxygen demand (COD and BOD, respectively), total suspended solids, total nitrogen and phosphorus. Also, many studies about the behavior of the small molecules (such as illicit drugs, pharmaceutics, personal care products, alcohol or tobacco) during the different steps undergone along the treatment process are available. However, there are no studies about how these treatments affect big biomolecules like proteins. Although some confirm the presence of proteins after some of the treatments or in the effluent, these proteins can be either from the sewages themselves or from the sludge used for those treatments. In this chapter the proteins present in the entry, exit and receiving bodies of the treatment plants are studied. Three sites with different population sizes and industrial activities were chosen to broaden the scope of the protein types. Only two of the wastewater treatment plants (WWTP) end up in rivers, while the third one goes to the sea where the sampling is more complex, leaving this site out for that part of the study. This is a collaborative project in which the samples were also analyzed for the characterization of small molecules by the group of Water Quality from the Catalan Institute for Water Research (ICRA Girona). Some of their results are included for the interpretation of the proteomics results.

ABSTRACT

Contaminants in wastewater are removed through multiple stages at the treatment plants. Afterwards, the processed water is discharged into the environment. Numerous studies have assessed the removal efficiency of these systems for small molecules, like pharmaceuticals, endocrine disruptive compounds and other substances related to human activity and health. However, the efficiency of the protein removal remains unknown. In this study, we analyzed the influent, effluent and receiving waters of three WWTPs, each serving different populations and industrial activities. Both proteins and a series of small molecules were characterized in the samples. The protein composition in the influent was consistent with previous findings from our research group. Following treatment, most of the proteins were effectively removed; however, the WWTP in Girona showed the lowest percentage of protein removal. Overall, treatment efficiency is higher in spring and lower in winter. The protein profile in the receiving waters was similar to the effluent one. Regarding the small molecules, 94 compounds were analyzed with removal efficiency varying across substances. In a first attempt in integrating proteomic and chemical data, we observed similar trends in the occurrence of antibiotics and albumins, allowing us to distinguish antibiotics used for non-human purposes like the treatment of livestock. In conclusion, proteins are effectively removed during wastewater

treatment, and the remaining ones are not hazardous for the environment. Moreover, proteomics can complement the information from other omics by providing insight into the origin of the small molecules.

1. INTRODUCTION

Every WWTP adapt the sequence of treatments in order to remove as many contaminants as possible before discarding the treated waters into (most of the times) the environment, which would at the same time meet the corresponding governmental guidelines. This adaptation will depend on the specific wastewater composition for that catchment area. The first stage of sewage treatment is known as primary treatment or physical treatment, and its objective is to eliminate larger suspended particles from the water through methods such as screening, grit chambers, air flotation, and filtration. The next phase, secondary sewage treatment or biological treatment, involves microbial oxidation, decomposition, and transformation processes to convert organic matter (and some inorganic substances) into microbial biomass. Thus, pollutants in the sewage can be degraded, transformed, and purified by creating a controlled environment and facilitating microbial metabolism. The tertiary treatment stage, also known as deep treatment, has gained widespread use in recent years. It incorporates technologies such as conventional treatment, membrane bio-reactor (MBR) technology, and liquid membrane (LM) treatment technology (Zhao et al., 2023).

Despite these treatments, industries and agriculture runoffs, even after processing, are often discharged into rivers or other water bodies. Numerous studies have detected pesticides, surfactants, estrogens, pharmaceuticals, personal care products and even abuse drugs in these waters. These substances have been linked to alterations in species composition, abundance or biomass, and endocrine disruption measured by alterations in enzymatic activity or specific protein production (González et al., 2012). This means that the removal of contaminants by WWTPs is in some cases not complete; as a consequence, contaminants can enter into the environment via sewage effluents and become a potential risk to the receiving waters and in the production of drinking water.

Sometimes, contaminant concentrations detected downstream to the discharge point are similar or slightly higher to those found upstream. Skees et al. (2018) reported the presence of illicit drugs, benzoylecgonine (a metabolite of cocaine), prescribed neuropsychiatric drugs, sedatives-hypnotics anxiolytics, antidepressants, and opioids in the influent of the WWTP. Additionally, they also detected in the effluent illicit drugs, sedatives-hypnotics anxiolytics,

antidepressants, and prescribed opioids (Skees et al., 2018). When they collected samples from the Bee Creek, the immediate receiving water body, they found several neuropsychiatric and illicit drugs that were going to end up in the Clarks River. Similar results were obtained for the Llobregat River in Spain by González et al. (2012). These compounds can accumulate in aquatic organisms and have unpredictable effects on human health through bioaccumulation in the food chain. They also pose toxicological risks to both the environment and organisms. These substances are biologically active, and their long-term effects are not fully known (Zhao et al., 2023).

Zhao et al. (2023) have recently assessed the efficiency of each step of the treatment in five different WWTP in China for the removal of illicit drugs. Their results showed that the anaerobic-oxic (A/O) combined with MBR treatment process was the most efficient, followed by oxidation ditch and anaerobic-anoxic-oxic (A2/O) processes. The A2/O combined with MBR was found to be the least effective (Zhao et al., 2023). However, this comprehensive study has not been conducted yet for proteins. It is known that proteins represent a large portion of organic nitrogen and carbon in wastewater treatment effluents. Westgate & Park (2010) showed that some proteins persisted through secondary treatment, while others were produced during biological treatment. This leads to proteins comprising a significant fraction of effluent organic-N (and carbon). Previously, Park et al. (2008a, 2008b) found human-derived proteins, such as human pancreatic elastase, in the extracellular matrix of activated sludge flocs and from anaerobically digested sludge product. The presence of sewage-derived polypeptides in settled sludge was an indication that these proteins are resistant to degradation in secondary aerobic treatment, and raised the question of whether these protein enzymes may also have been present in the dissolved phase of activated sludge that was discharged into receiving waters. In 2020, Carrascal et al. (2020) used polymer probes immersed in the influent, anoxic reactor and effluent waters of a Spanish WWTP to study the proteins in each step. They identified a total of 690 proteins from bacteria, plants and animals, including humans. Bacterial proteins pointed at 175 genera distributed in 22 bacterial classes, but human was the species contributing the greatest number of identified proteins. Some of these human proteins were already known disease biomarkers: S100A8, uromodulin (also proposed by Rice & Kasprzyk-Hordern (2019)) and defensins. These proteins were found in influent and were removed throughout the treatment. Therefore, the authors concluded that proteins were efficiently removed at the effluent (Carrascal et al., 2020).

Microbial extracellular proteins also play key roles in wastewater treatment. In general, extracellular proteins consist of enzymes and structural proteins, whose major functions in wastewater treatment include the formations of microbial aggregate, pollutant migration and

transformation, and resistance to toxic substances. They also contribute to pollutant removal in wastewater treatment via adsorption and catalysis, strongly binding heavy metals, organic matter, and nanoparticles through hydroxyl, carboxyl, phosphate and amide groups. Besides, extracellular enzymes supply an external digestion system, playing an important role in the degradation of the organic small molecules, organic colloidal fraction and particulate biomass (Zhang et al., 2019). These proteins can, however, become exogenous refractory proteins, i.e. these proteins are hard to biodegrade and may persist in the aquatic environment. High molecular weight proteins are especially refractory and can interact with microbial cells, altering physiological processes, enhancing biofilm formation, modifying microstructure, disrupting quorum sensing, and triggering excessive polysaccharide production. These effects can negatively impact downstream ecosystems (Cui et al., 2019).

In this study, we unravel the soluble proteins present in the effluent compared with the inlet of three WWTP with different sizes of populations and industrial activities with the aim of determining the efficiency of the treatment plants in the removal of proteins. These results are compared with metabolomic data, specifically with compounds related to human activities in order to find possible tendencies and proteins that could be used as biomarkers together with the small molecules. Next, the upstream and downstream composition of two of the WWTPs, which end up in rivers, were determined in order to elucidate the origin of the proteins present.

2. MATERIAL AND METHODS

2.1. Sample collection

An automatic water sampler was used to collect twenty-four-hour composite wastewater samples at the inlet and outlet of 3 wastewater treatment plants (WWTPs) according to their hydraulic retention time: Besòs (16 hours), Girona (24 hours) and Vic (48 hours). Three collection campaigns were conducted in spring, summer, and winter to evaluate possible seasonal variations. Additionally, twenty-four-hour composite samples were collected at the rivers Ter (Girona) and Riera de Rimentol (Vic) before and after the corresponding wastewater treatment plants (WWTPs), called upstream and downstream, respectively. The collection was made on May 31st 2023 in Girona and June 21st 2023 in Vic. The samples were then transferred to the laboratory at 4 °C.

2.2. Sample preparation

2.2.1 Proteomics

Samples were prepared following a modified method from Sánchez-Jiménez et al. (2023) depending on the type of sample. All the samples were ultracentrifuged immediately after arrival at the laboratory. For this, up to 90 ml of 24-h composite wastewater sample was ultracentrifuged at 24000 × g (4 °C, 60 min), and the supernatant collected in a new tube. Influent and effluent samples were kept at -40 ºC until further preparation. On the other hand, samples from the rivers and their corresponding effluents (called river samples from now on) were frozen at -80 ºC and lyophilized in the freeze dryer Modulyo.

For the analysis, 10 ml (for influent) or 50 ml (for effluent) were concentrated using a 10 kDa cutoff device (Amicon®, NMWL 10 kDa), with a filter that was previously passivated to minimize protein adsorption as described in Sánchez-Jiménez et al. (2023). At the same time, for the analysis of the river samples, the lyophilized were reconstituted in 15 ml of MilliQ water, filtered through a 0.2 µm polyethersulfone membrane and concentrated using a 10 kDa cutoff device (Amicon®, NMWL 10 kDa), with a previously passivated filter. All the samples were concentrated to approximately 400 µl, then evaporated to dryness using a SpeedVac.

Proteins in the samples were cleaned and concentrated in the heads of sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) gels (5% stacking and 12% resolving) at 50 V for around 60 min. Bovine serum albumin (BSA) was used as a reference marker. After electrophoresis, the gels were stained with Coomassie Blue. The bands of concentrated proteins were excised and digested with trypsin using an automatic device (DigestPro MS, Intavis). The process is detailed in Sánchez-Jiménez et al. (2023).

2.2.1 Metabolomics

The group of Water Quality from the Catalan Institute for Water Research (ICRA Girona) processed at the same time the samples for metabolites using two previously developed and validated methods (Gros et al., 2012; Gros et al., 2013; Jakimska et al., 2013). Briefly, the samples were filtered through 1 µm glass fiber filters (GFF, Whatman) and 0.45 µm polyvinylidene fluoride (PVDF, Merck Millipore) filters. For the analysis of wastewater influents and effluents, 25 ml and 50 ml were measured, respectively. Before extraction, isotopically labeled standards were added as internal standards to each sample at a concentration of 25 ng/ml (Sigma-Aldrich and Toronto Research Chemicals). Samples were pre-concentrated using solid-phase extraction (SPE) with Oasis HLB (60 mg, 3 ml, Waters Corporation)

cartridges for pharmaceutically active compounds (PhACs) and antibiotics analysis (Gros et al., 2012; Gros et al., 2013) and Strata-X 33 µm (200 mg, 6 ml, Waters Corporation) cartridges for pesticides and endocrine disrupting compounds (EDCs) (Jakimska et al., 2013).

2.3. LC-HRMS/MS analysis and data treatment

2.3.1 Proteomics

The LC-HRMS/MS system consisted of a Waters NanoAcquity UPLC (a binary pump, a thermostatic micro injector, and a trap valve) coupled to an Orbitrap Exploris 480 Mass Spectrometer (ThermoFisher) equipped with a nanoESI ion source.

For the analysis, the tryptic digests were evaporated until dry and re-dissolved in 50 µl (inlet and outlet) or 11 µl (river samples) of 0.5% TFA 5% methanol with gentle agitation in a Thermomixer (5 min, at 22 °C, 500 rpm). In the case of the inlet and outlet samples, 5% of each sample was injected into the LC-MS system, while the 50% of the river samples were injected.

Separation was performed on a 10-cm long, 100 µm i.d. C18 column (Waters) preceded by a C18 trap column (Waters). Separation was done at 0.5 µl/min using a 120-min gradient for inlets samples, and 30-min gradient for outlets and river samples. The gradient runs from 2 to 35% solvent B (solvent A: 0.1% formic acid, solvent B: acetonitrile 0.1% formic acid) with a previous 3-min trapping at 2% solvent B.

The Orbitrap Exploris 480 was operated in positive ion mode with a spray voltage of 1.8 kV. Spectrometric analysis was performed in data-dependent mode, acquiring a full scan followed by 20 (inlets samples) or 15 (outlets and river samples) MS/MS scans of the 20 or 15 most intense signals detected in the MS scan. Full MS (range 400-1600) spectra were acquired in the Orbitrap with different resolutions to enhance the number of identifications: 60000 (inlets samples) and 120000 (outlets and river samples). All MS/MS spectra were acquired in the Orbitrap with a resolution of 15000.

MS/MS spectra were searched using SEQUEST (Proteome Discoverer v3.0, ThermoFisher) with the following parameters: peptide mass tolerance, 20 ppm; fragment tolerance, 0.02 Da; enzyme, trypsin, and allowance of up to two missed cleavages; dynamic modification, methionine oxidation (+16 Da), methionine loss (-131 Da), acetyl N-terminal (+42 Da) and methionine loss plus acetyl in N-terminal (-89 Da); and fixed modification, cysteine

carbamidomethylation (+57 Da). Searches were performed using the UniProt database (rev. 12-2023 for inlets and outlets samples, and rev. 06-2024 for river samples). Final results were filtered using 1% False-Discovery-Rate (Nesvizhskii, 2010).

Overall descriptions of the soluble wastewater proteome from the influent, the effluent and the receiving bodies were obtained from the protein identification output of a Protein Discoverer Multiconsensus analysis for each one, including all protein identifications from the different sites and campaigns. Estimation of the relative abundance of proteins was based on normalized spectral counts (NSCs). NSCs correspond to the total peptide sequence matches (PSM) obtained using Protein Discoverer and normalized to the mass of the protein to consider that the number of tryptic peptides produced by a protein increases with its size, and thus also the total PSMs measured (Carrascal et al., 2023). For the comparatives with the small molecules the abundance (normalized) calculated by the software was used, as it is more similar to the data obtained from metabolomics.

2.3.2 Metabolomics

The group of Water Quality from the Catalan Institute for Water Research (ICRA Girona) was also responsible for the LC-HRMS/MS analysis and treatment of the data for the metabolites. Briefly, micropollutant detection and quantification was carried out using ultra-high-performance liquid chromatography (UHPLC), using a Waters Acquity Binary Solvent Manager system, coupled to a quadrupole linear ion trap tandem mass spectrometer (5500 QTRAP, AB Sciex). Chromatographic separation under positive electrospray ionization was achieved using an Acquity HSS T3 column, while for negative electrospray ionization mode, an Acquity BEH C18 column was used. For quantitative and detection purposes, two selected reaction monitoring (SRM) transitions were monitored for each compound. Data acquisition and processing were performed using SCIEX 3.1.5 software.

3. RESULTS

3.1 Influent and effluent proteome

Samples from the influent and effluent of 3 WWTPs in three different campaigns and 3 consecutive days each time were collected. After concentration, clean-up in SDS-PAGE gels and gel digestion, peptide extracts were analyzed by LC-MS/MS and the raw data searched using Proteome Discoverer 3.0. This approach allowed the identification of 10946 peptides (1% FDR, >1 PSM) that pointed to 2085 proteins (1% FDR, >1 peptide) (see supplementary material 4.3_Proteome.xlsx). The number of identified proteins in each sample is in Table 9.

Table 9. Number of identified proteins in influent (In) and effluent (Out) with >1 PSM and >1 peptide.

WWTP		Besòs		Girona		Vic	
Campaign	Day	In	Out	In	Out	In	Out
Spring	1	491	-	611	-	764	-
	2	532	-	559	-	779	-
	3	446	-	386	-	842	-
Summer	1	687	15	783	-	818	-
	2	635	13	872	26	921	-
	3	486	-	682	45	1010	-
Winter	1	130	14	942	189	1146	36
	2	267	15	1232	22	1191	-
	3	169	14	1214	108	1095	39

In most of the campaigns, proteins are removed completely after treatment, at least below the limit of detection of the mass spectrometer (see Table 9 and Figure 28). The sample with the least percentage of proteins removed is Girona on the day 1 of the winter campaign, only 80% of the proteins were cleaned in the process. In general, it is seen that the treatment process is more efficient in spring and less efficient in winter.

The most abundant proteins in the influent were, as expected, human and animal amylases followed by albumins (Table 10) based on the number of normalized spectral counts (NSCs). On the other hand, the most abundant proteins in the effluent are keratins, followed by some chaperonins and enzymes (Table 11). Notice that the most abundant protein in the effluent has more than 25 times less NSCs than the most abundant one in the influent.

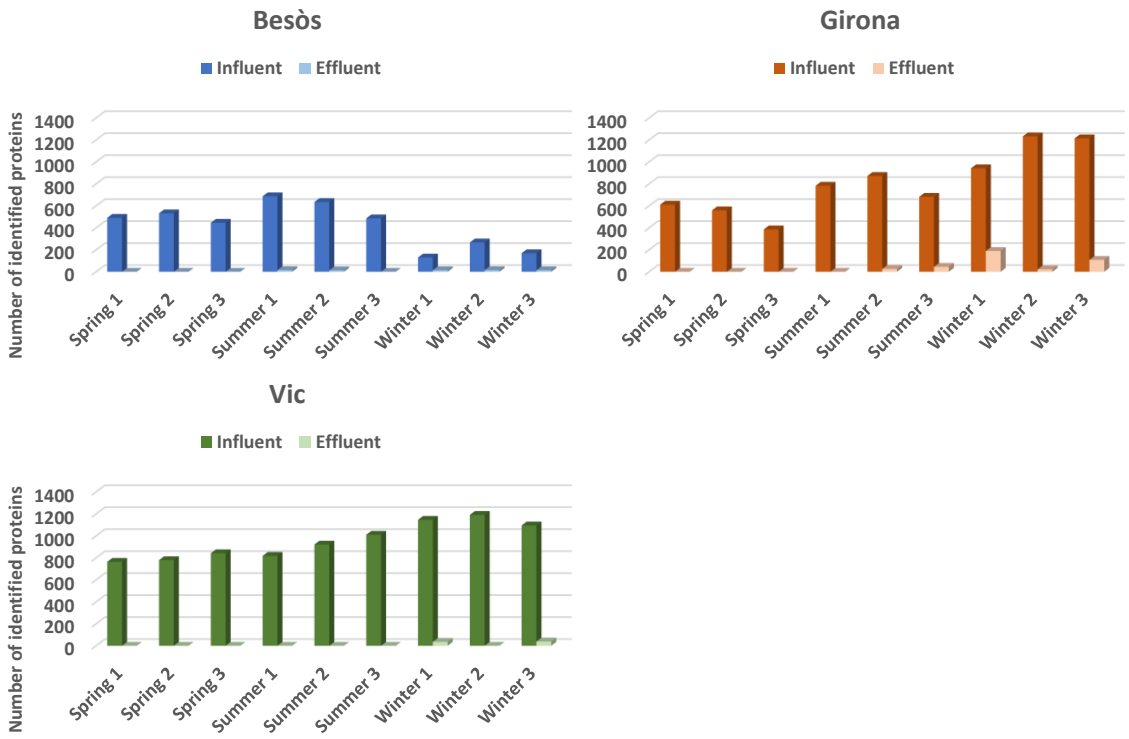


Figure 28. Number of identified proteins in each WWTP (Besòs, blue; Girona, orange; Vic, green) in the influent (dark color) and the effluent (light color) in each campaign and day.

3.2 Influent and effluent metabolome

In total 94 compounds were analyzed (see supplementary material 4.3_Metabolome.xlsx), among them there were: 52 pharmaceuticals, 25 antibiotics, 4 pesticides, and 13 endocrine disruptive compounds (EDCs). Each superclass is divided in different classes, the number of compounds per class are specified in Table 12.

These compounds were targeted according to their significance on human health and habits, and their environmental importance. Therefore, not all the compounds are going to be discussed here, only the ones related to human health. These compounds' classes are marked with an asterisk in Table 12.

Among the analgesics and anti-inflammatories compounds (see Figure 29A, left), the most abundant is acetaminophen in spring and summer. However, in winter its quantity is reduced and it is substituted by ibuprofen and naproxen. In general, it seems to be a tendency where these compounds are higher in Besòs. Although in less quantities (sometimes an order of magnitude lower), there are other important classes, such as diuretics, antihypertensives and lipid regulators (Figure 29B, left). Several compounds were analyzed for each class, but there is one from each that stands out: hydrochlorothiazide (a diuretic; higher in summer than spring

and winter), valsartan (an antihypertensive; it is more abundant in Besòs than in Girona and Vic) and gemfibrozil (a lipid regulator; it does not have a clear pattern along the seasons or sites). In total, 13 different psychiatric drugs were analyzed (see Figure 29C, left). Their quantities were not as high as other compounds, but some of the most abundant were O-desmethylvenlafaxine and venlafaxine, which were present in the three campaigns, mainly in summer.

Table 10. The 20 most abundant proteins in the influent samples. STRCA: Ostrich; PONAB: Orangutan.

Access	Protein name	Entry name[1]		Coverage (%)	# Peptides	# Protein unique peptides	# NSCs
		Gene	Species				
P04746	Pancreatic alpha-amylase	AMY2A	HUMAN	93	64	10	22374
P0DUB6	Alpha-amylase 1A	AMY1A	HUMAN	93	65	12	20719
P19961	Alpha-amylase 2B	AMY2B	HUMAN	93	62	2	19945
P01012	Ovalbumin	SERPINB 14	CHICK	89	35	26	12974
Q1KYT0	Beta-enolase	ENO3	PIG	91	75	9	11718
Q3ZC09	Beta-enolase	ENO3	BOVIN	88	68	3	10679
P08835	Albumin	ALB	PIG	92	123	95	10582
P02769	Albumin	ALB	BOVIN	90	108	59	9626
P83053	Pancreatic alpha-amylase	AMYP	STRCA	35	21	3	9485
P25704	Beta-enolase	ENO3	RABIT	79	64	1	8885
P01009	Alpha-1-antitrypsin	SERPINA 1	HUMAN	63	41	20	8413
P02768	Albumin	ALB	HUMAN	89	93	17	8115

P13929	Beta-enolase	ENO3	HUMAN	82	58	4	8032
P01834	Immunoglobulin kappa constant	IGKC	HUMAN	93	15	4	7983
P00687	Alpha-amylase 1	Amy1	MOUSE	40	25	8	7694
P00690	Pancreatic alpha-amylase	AMY2	PIG	61	31	16	7313
P00689	Pancreatic alpha-amylase	Amy2	RAT	36	19	5	7238
Q5E956	Triosephosphate isomerase	TPI1	BOVIN	98	46	9	7146
Q5NVH5	Albumin	ALB	PONAB	80	77	3	6360
Q29371	Triosephosphate isomerase	TPI1	PIG	75	35	6	6000

[1] UniProtKB/Swiss-Prot entry name. The two terms of the entry name (gene_species) have been separated for convenience.

For antibiotics (Figure 30), 10 different classes encompassing 25 compounds were studied. The most abundant antibiotic was doxycycline (a tetracycline), followed by sulfapyridine (from sulfonamide class). Interestingly, doxycycline was an order of magnitude higher in Vic in the spring season compared with the rest of samples, while sulfapyridine was more abundant in Besòs during spring. Moreover, the tetracycline was present in high quantities in Besòs in summer and ciprofloxacin (a fluoroquinolone) was mostly used in winter in Besòs.

Among the endocrine disruptive compounds (Figure 31), we focused on hormones, caffeine and the corrosion inhibitor 1H-benzotriazole due to their high abundance. Caffeine was the most abundant compound of this superclass, followed very close by 1H-benzotriazole. Both of them have the same pattern, being much higher in spring than summer and winter. Caffeine seems to tend to increase over the three days.

Table 11. The 20 most abundant proteins in the effluent samples. ALBFT, VARPS, LEPCP, POLSJ, DELAS: Bacteria; CANLF: Dog.

Access	Protein name	Entry name[1]		Coverage (%)	# Peptides	# Protein unique peptides	# NSCs
		Gene	Species				
P00761*	Trypsin	TRYP	PIG	35	7	5	832
P04264	Keratin, type II cytoskeletal 1	KRT1	HUMAN	54	32	3	277
P13645	Keratin, type I cytoskeletal 10	KRT10	HUMAN	39	23	8	224
P09093	Chymotrypsin-like elastase family member 3A	CELA3A	HUMAN	54	11	8	214
P35908	Keratin, type II cytoskeletal 2 epidermal	KRT2	HUMAN	48	26	16	203
P35527	Keratin, type I cytoskeletal 9	KRT9	HUMAN	49	22	21	158
P02769	Albumin	ALB	BOVIN	59	42	22	129
Q21ZD1	Chaperonin GroEL 1	groEL1	ALBFT	57	30	7	118
P07478	Trypsin-2	PRSS2	HUMAN	12	2	1	109
C5CPP8	Chaperonin GroEL	groEL	VARPS	41	24	0	106
P04259	Keratin, type II cytoskeletal 6B	KRT6B	HUMAN	30	14	0	105
B1XXY9	Chaperonin GroEL	groEL	LEPCP	46	22	1	103
Q6EIY9	Keratin, type II cytoskeletal 1	KRT1	CANLF	13	10	0	102
A1WL05	Chaperonin GroEL	groEL	VEREI	38	23	0	96
Q12FH7	Chaperonin GroEL	groEL	POLSJ	51	24	1	94

P04746	Pancreatic alpha-amylase	AMY2A	HUMAN	43	18	0	92
P02538	Keratin, type II cytoskeletal 6A	KRT6A	HUMAN	30	13	1	90
P02533	Keratin, type I cytoskeletal 14	KRT14	HUMAN	36	15	2	89
P08779	Keratin, type I cytoskeletal 16	KRT16	HUMAN	45	19	8	88
A9BXL3	Chaperonin GroEL	groEL	DELAS	25	15	1	84

[1] UniProtKB/Swiss-Prot entry name. The two terms of the entry name (gene_species) have been separated for convenience. * Enzyme used for in-gel digestion.

Table 12. Number of compounds in each class per superclass.

Superclass	Class	Nº	Class	Nº
Pharma-ceuticals	Analgesics/anti-inflammatories*	13	Lipid regulators*	4
	Anthelmintics	2	Prostatic hyperplasia treatment	1
	Antihypertensives*	3	Psychiatric drugs*	13
	Antiplatelet agents	1	Synthetic glucocorticoids	1
	Ca^{+} channel blockers	2	Asthma treatment	1
	Diuretics*	2	X-ray contrast agents	1
	Histamine receptor antagonists*	3	β-blocking agents*	5
Antibiotics	Cephalosporins*	2	Nitroimidazoles	3
	Diaminopyrimidines*	1	Penicillins*	2
	Fluoroquinolones*	4	Pleuromutilins*	1
	Lincosamides	2	Sulfonamides*	3
	Macrolide antibiotics*	4	Tetracyclines*	3
Pesticides	Phenyl ureas	2	Triazines	2
Endocrine disruptive compounds	Corrosion inhibitors*	2	Plasticizers	1
	Flame retardants	3	Progestin medications	1
	Hormones*	2	Stimulants*	1
	Parabens	3		

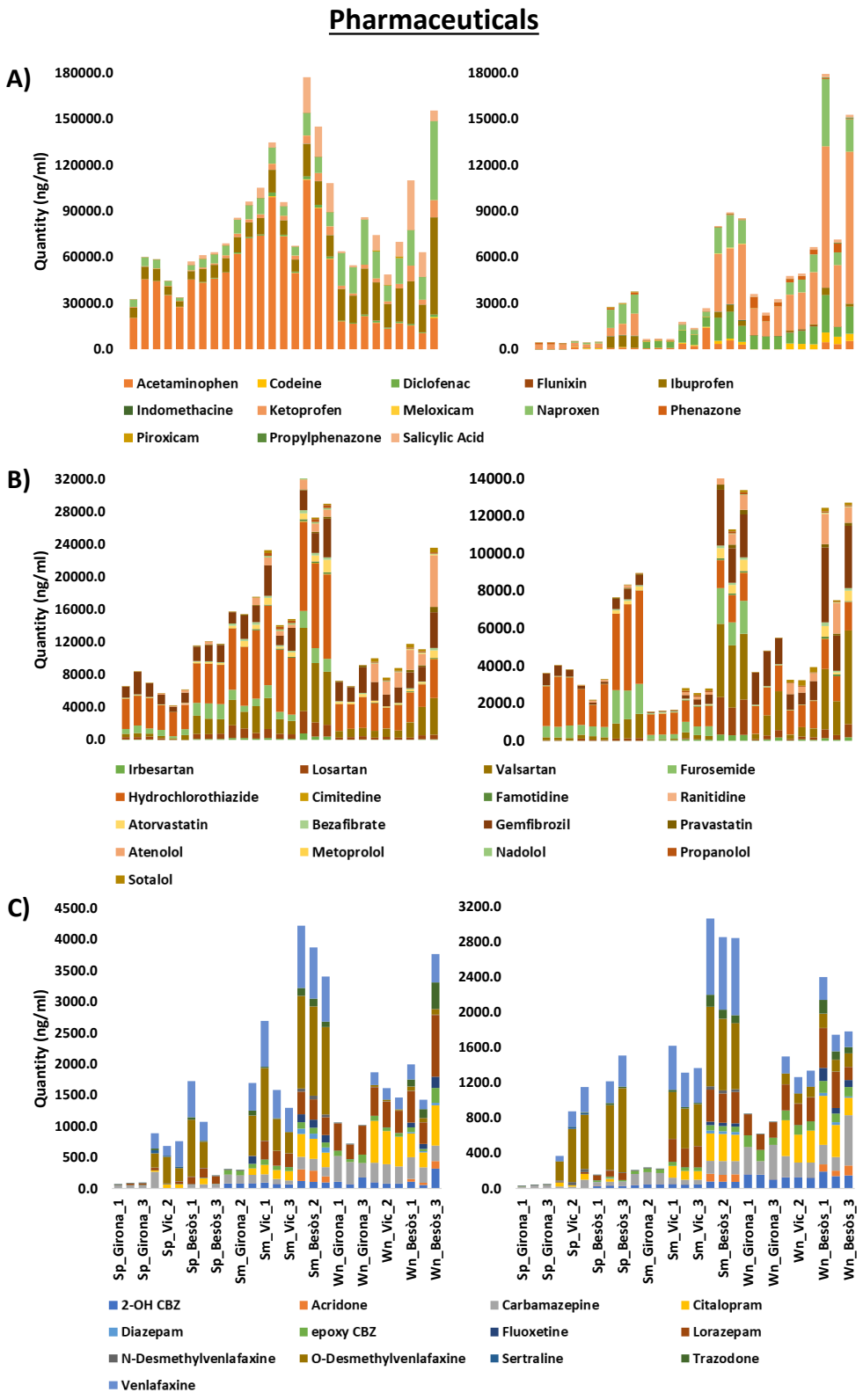


Figure 29. Quantity of the selected pharmaceuticals present in the influent (left column) and in the effluent (right column) in every sample (Sp = spring, Sm = summer, Wn = winter). Notice that the Y-axis range is different in each graphic.

Related to the efficiency of removal of the compounds it varies in each case (see Figures 29, 30 and 31, right). The analgesics and anti-inflammatories compounds in the effluent follow in general the same pattern than in the influent; an exception is the ketoprofen, whose removal efficiency is lower (Figure 29A, right). This is the same hydrochlorothiazide, valsartan and gemfibrozil, which are the most abundant both in the effluent and the effluent (Figure 29B, right). In psychiatric drugs, venlafaxine has a very low removal efficiency (Figure 29C, right). The antibiotics removal is good in Girona and Vic, but very poor in Besòs, regardless of the

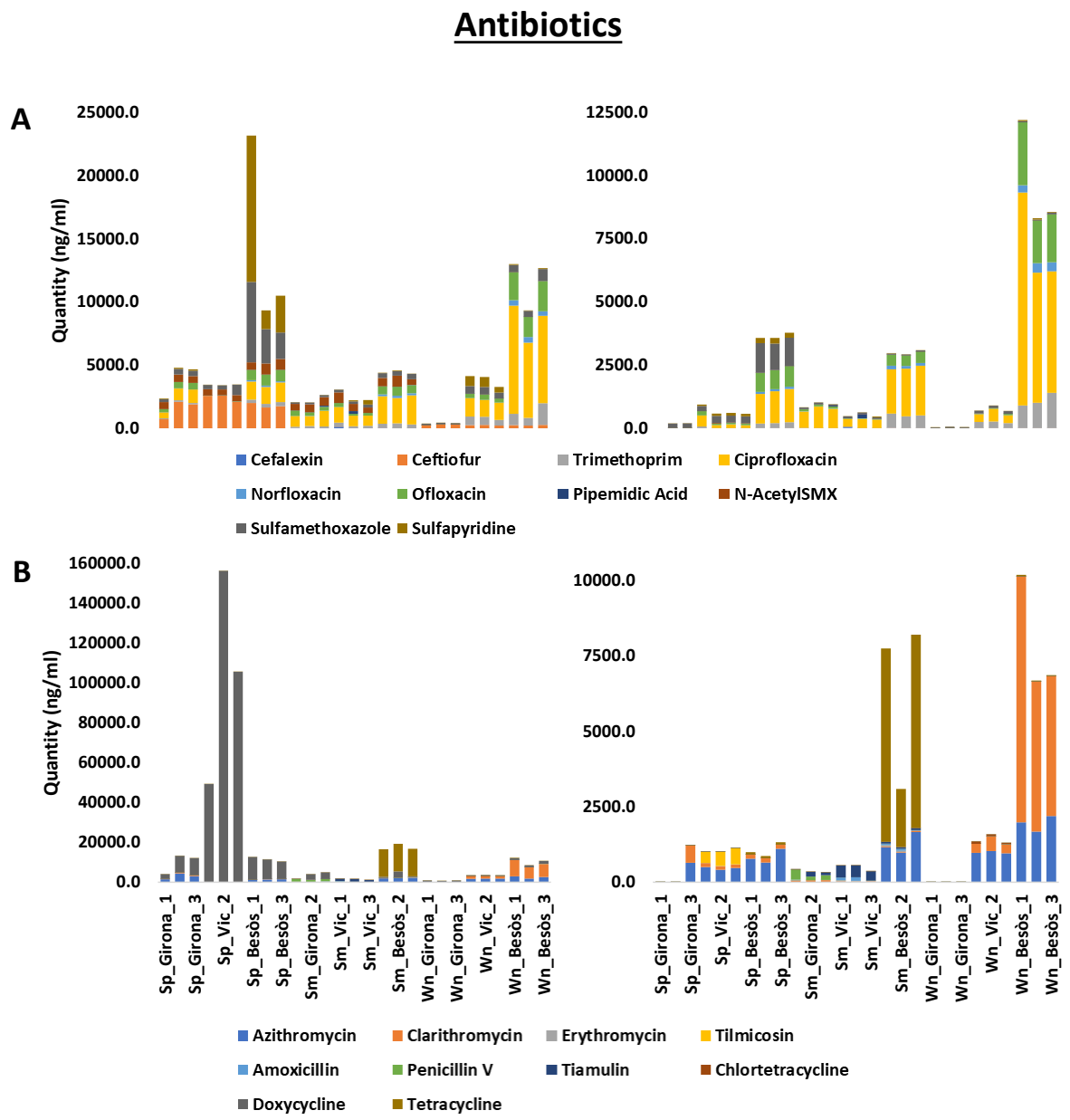


Figure 30. Quantity of the selected antibiotics present in the influent (left column) and in the effluent (right column) in every sample (Sp = spring, Sm = summer, Wn = winter). Notice that the Y-axis range is different in each graphic.

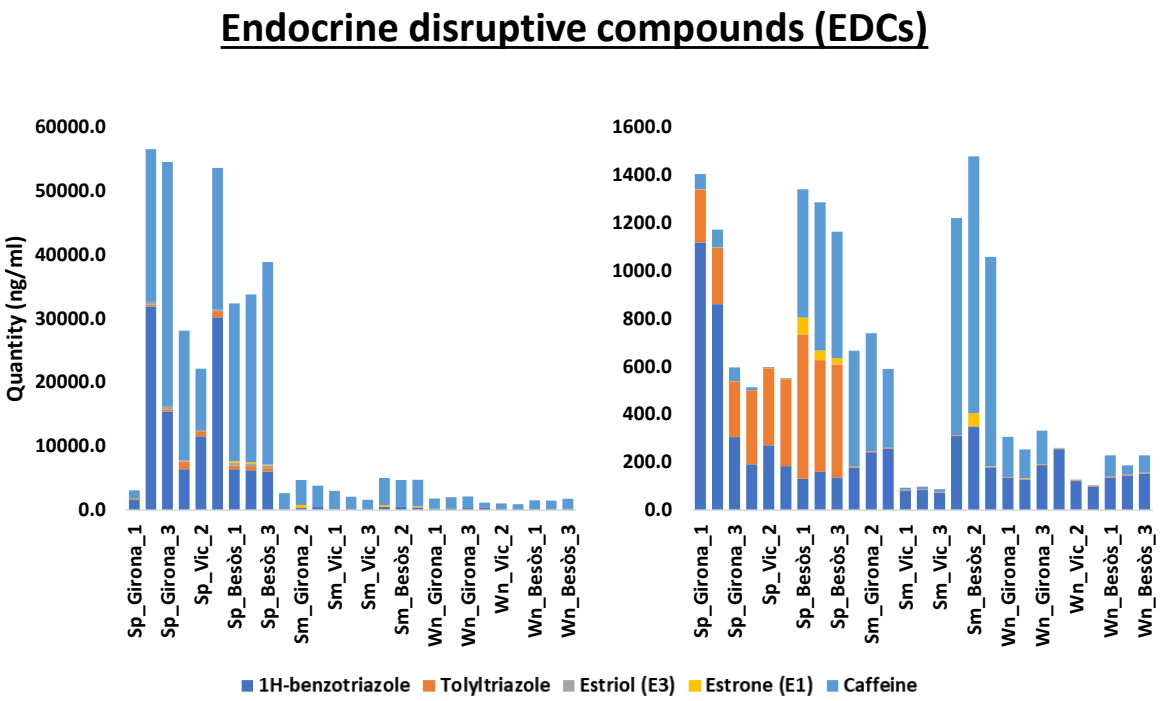


Figure 31. Quantity of the selected EDCs present in the influent (left column) and in the effluent (right column) in every sample (Sp = spring, Sm = summer, Wn = winter). Notice that the Y-axis range is different in each graphic.

season (Figure 30, right). The efficiency of removal of the EDCs seems to be very good, even in spring, being the effluent quantities very similar in all the sites and seasons (Figure 31, right).

3.3 Small molecules-proteins connectivity

Albumins and amylases are among the most abundant proteins detected in wastewater as it has already been described (Carrascal et al., 2023) and is confirmed in this study. Moreover, these proteins are particularly informative due to their origins: amylases come from feces and were proposed as mammalian population indicators, while albumins come from the blood and were suggested as livestock industry markers. So far, human habits and health biomarkers used in wastewater are small molecules due to the more advanced development of chemical analysis in this field. Here we explored the possible trends of livestock albumins and human amylases in relation to antibiotics and caffeine, which are important indicators. However, this research represents an initial inquiry, and the evidence at this stage is inconclusive. It is worth it to notice that small molecules injections run into issues, which led to very low reproducibility. Furthermore, protein quantification is not absolute as not internal standards were added, it is

a semi-quantification based on normalized abundance from the search engine. Experiments aimed at the confirmation or discard of these observations are in process.

The Figure 32A shows all the identified albumins from human and livestock origin found in the influent samples. Pig and cow albumins are the most abundant and have different patterns, being the pig albumin higher in Vic samples, while the cow albumin is such in Girona. Notably, Girona has other albumins with the same tendency: rabbit, donkey, cat, sheep and rat (Figure 32B). On the other hand, chicken albumin has a similar pattern as pig albumin (Figure 32C). The other types of albumin do not have significant differences among the sites and seasons.

The antibiotics show distribution patterns similar to albumins in the case of the livestock and to the human amylases when associated with human activity. The clearer example is doxycycline, which is by far the most abundant antibiotic detected. However, its prevalence is mostly limited to the site Vic in the spring season, following the same tendency as pig and chicken albumins (Figures 33A and 33C). Other antibiotics appear to be primarily associated with human use, as indicated by their higher concentrations in Besòs samples, and follow similar tendencies than the human alpha-amylase 1A, but their abundances vary depending on the seasons (see Figures 33A and 33D). Some of the compounds that stand out are: azithromycin, ciprofloxacin and ofloxacin in the three seasons (spring, summer and winter), sulfamethoxazole in spring and winter, and clarithromycin, trimethoprim and norfloxacin only in winter. Some of these antibiotics are also high in Girona, such as azithromycin in spring, and to a much less degree in Vic, for example, ciprofloxacin in summer, and sulfamethoxazole, azithromycin, ofloxacin and clarithromycin in winter.

Caffeine is used as a population marker (Rico et al., 2017) and have recently been used for normalization in wastewater (Oloye et al., 2023). We proposed amylases enzymes as a population marker alternative (Carrascal et al., 2023). For this reason, a similar trend was expected when comparing the quantity of caffeine and the abundance of human amylase enzymes (Figure 34). However, this was not the case, only with the alpha-amylase 1A there was a visual similar trend with caffeine, but the correlation was very poor, except in summer whose correlation was 0.7148 (Figure 34B to D). This could be, as commented before, due to a problem in the analysis of the caffeine and in the case of the proteins due to the use of the relative abundance and not an absolute concentration, but the summer results are promising.

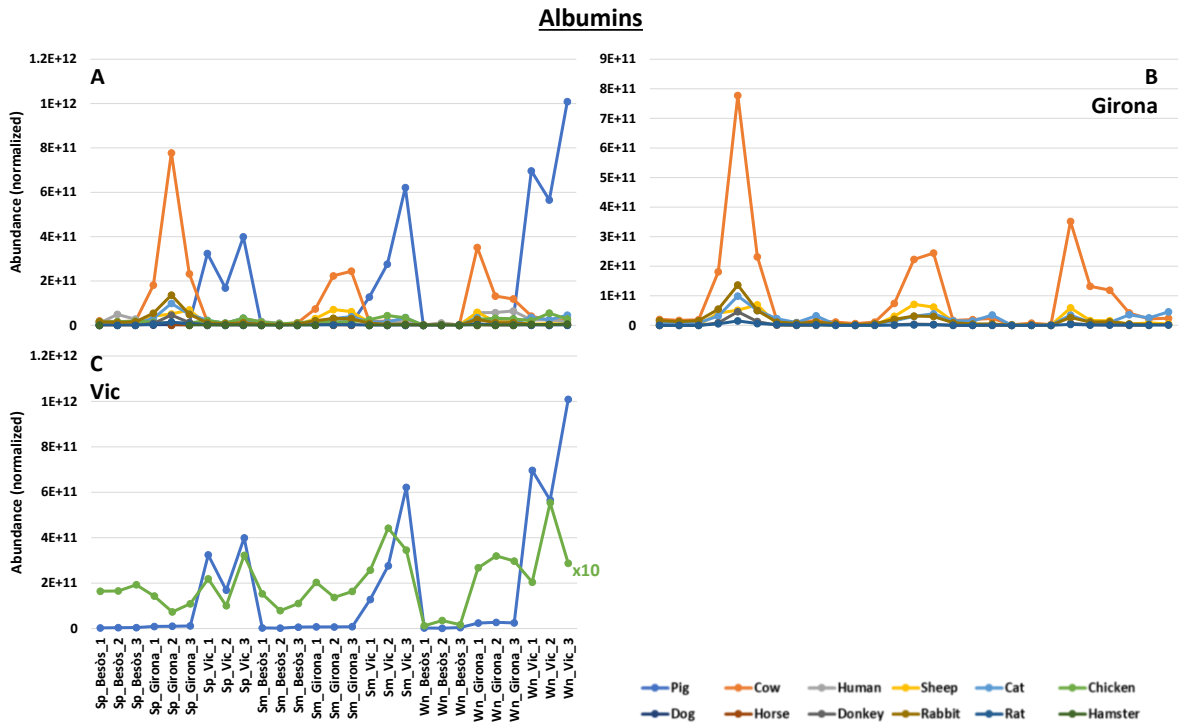


Figure 32. Albumin abundances. A) All the identified albumin from human and livestock; B) Albumins that are more abundant in Girona; C) Albumins which are higher in Vic. Notice that the Y-axis range is different in each graphic. (Sp = spring, Sm = summer, Wn = winter)

3.4 Receiving waters proteome

Samples were collected from the upstream, effluent and downstream of 2 WWTPs. After ultracentrifugation, concentration, clean-up in SDS-PAGE gels and gel digestion, peptide extracts were analyzed by LC-MS/MS. The raw data was searched using Proteome Discoverer 3.0. This allowed the identification of 729 peptides (1% FDR, >1 PSM) that accounted for 109 proteins (1% FDR, >1 peptide) (see supplementary material 4.3_Proteome.xlsx).

The most abundant proteins based on the number of NSCs are very different from the ones found in the influent, but similar to the ones in the effluent. The first one is the enzyme used in the in-gel digestion (trypsin from *Sus scrofa*) and then followed by different types of human keratins.

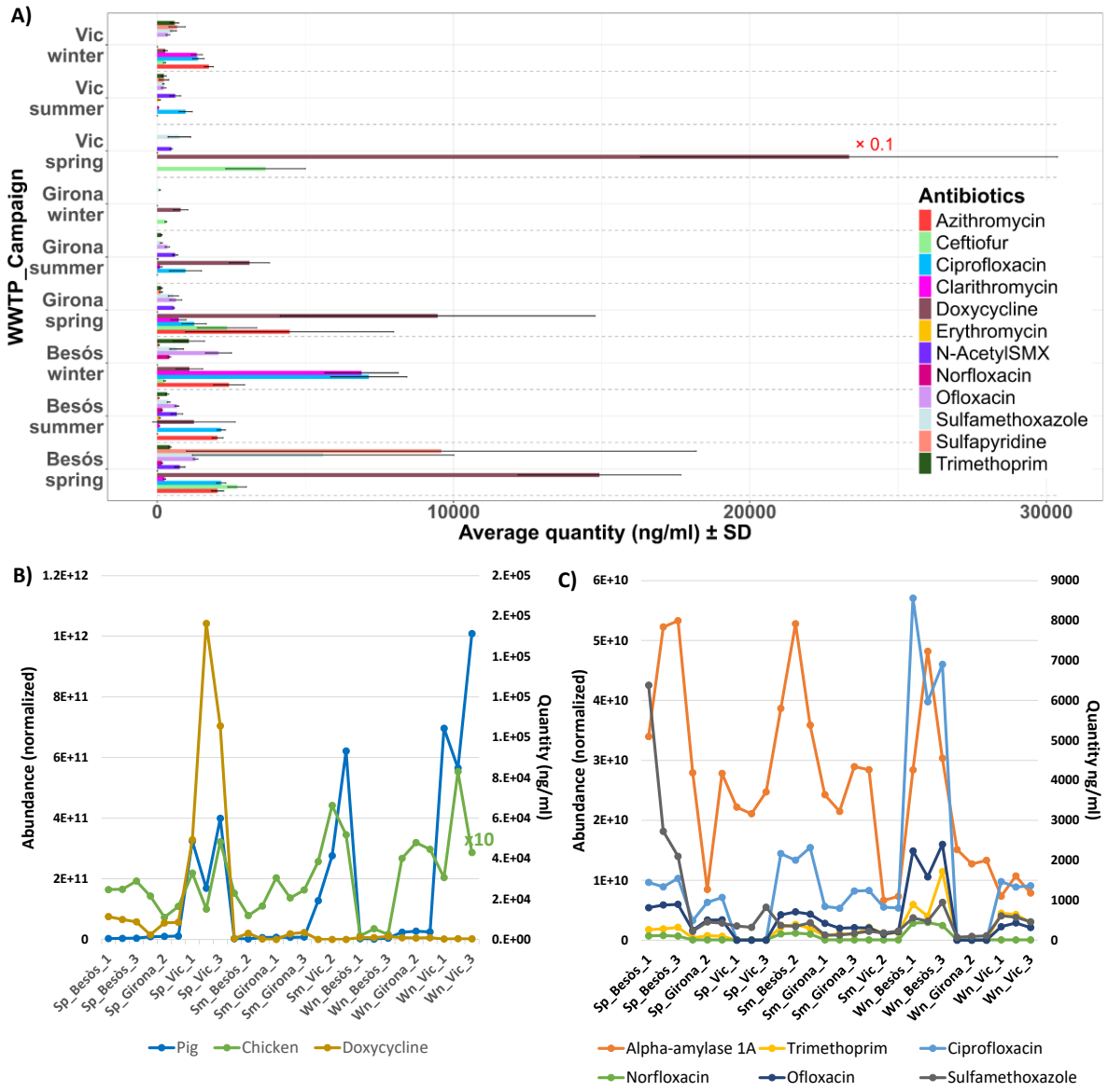


Figure 33. A) Average and standard deviation of each antibiotic calculated from the technical and biological replicates in the influent samples. B) Abundance of pig and chicken albumins, and doxycycline in every sample. C) Abundance of human alpha-amylase 1A, trimethoprim, ciprofloxacin, norfloxacin, ofloxacin, and sulfamethoxazole in every sample. (Sp = spring, Sm = summer, Wn = winter)

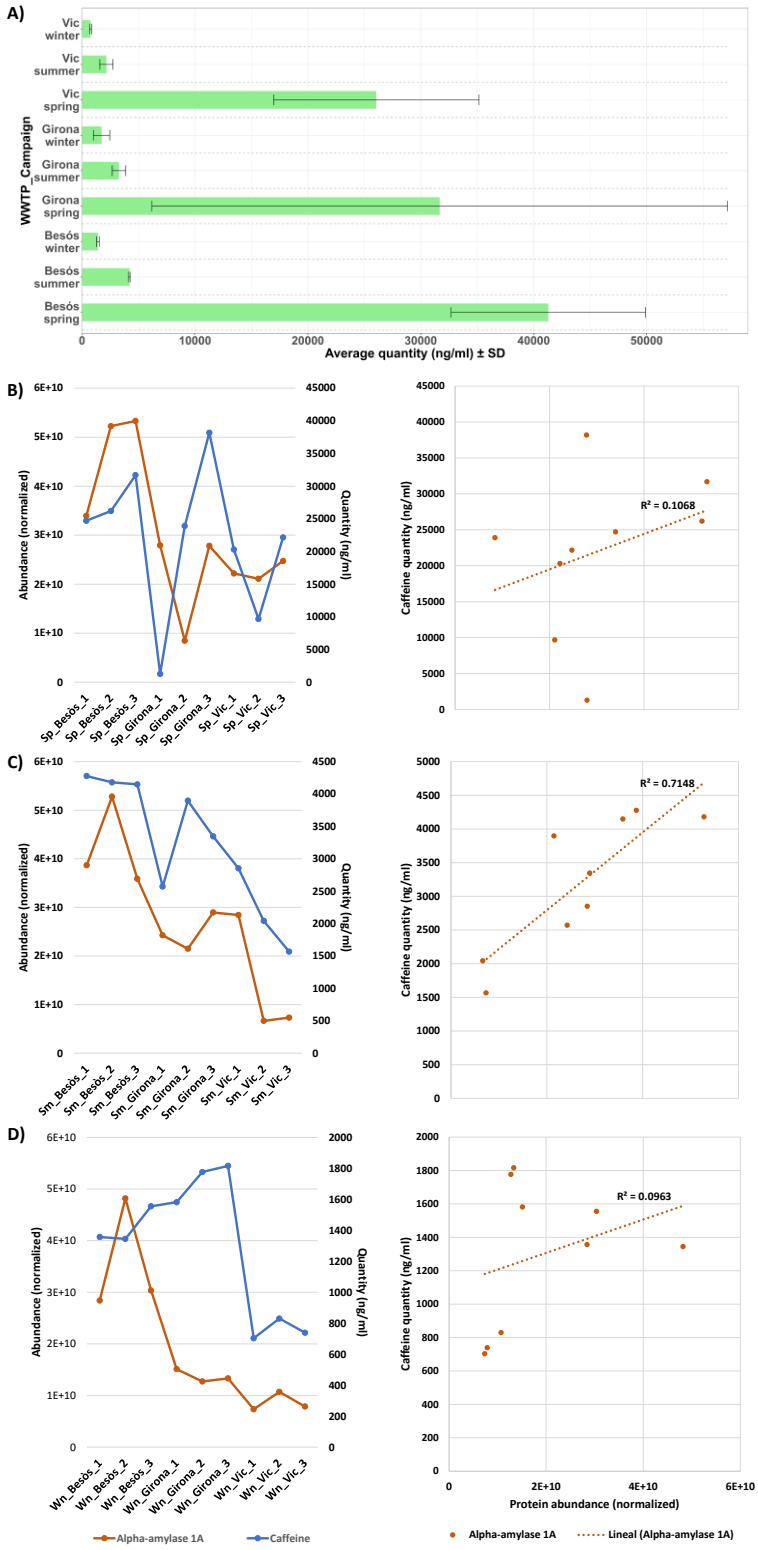


Figure 34. A) Average and standard deviation of the caffeine calculated from the technical and biological replicates in the influent samples. B) Abundance of caffeine and alpha-amylase 1A in spring (left) and their lineal correlation (right). C) Abundance of caffeine and alpha-amylase 1A in summer (left) and their lineal correlation (right). D) Abundance of caffeine and alpha-amylase 1A in winter (left) and their lineal correlation (right). (Sp = spring, Sm = summer, Wn = winter)

4. DISCUSSION

In this study, we characterized the protein composition of influent and effluent samples from three WWTPs considering the hydraulic retention time of each plant. This ensures that the water collected in the outlet is the same as the one collected in the inlet, which allows the comparison of their composition before and after the treatments. In addition, the upstream and the downstream of the rivers Ter and Riera de Rimentol, in which the Girona and Vic treatment plants respectively unload the treated water, was also studied. To the best of our knowledge this is the first study investigating the proteins that remain in the treated water and those present in the receiving waters where they end up. More importantly, it is the first research directed to study all the proteins present in this media using a complete protein database instead of a reduced one (bacteria or human database) (Park et al., 2008a; Cui et al., 2019). In parallel, our collaborators in the group of Water Quality from the Catalan Institute for Water Research (ICRA Girona) characterized, in the same samples, a series of small molecules compounds which are used in the study of human health and habits, and the environment status.

The WWTPs were selected to represent different anthropogenic influences: Besòs is the representation of a highly populated city as it serves half of Barcelona city, Vic is an industrial city and Girona could be considered an intermediate city with both a high human population and some industrial activity. The three WWTP have pre-, primary and secondary treatments, as well as decantation and thickening steps. The secondary biological treatment includes phosphorus and nitrogen removal. Girona also has tertiary treatments such as microfiltration and chlorination. Therefore, differences in treatment technologies and hydraulic retention leads to variability in treatment efficiency.

Among the three, the Vic WWTP has the longest retention time and, as we show here, the highest protein removal efficiency. In this site there are the maximum number of samples (7 out of 9) with 100% of protein removal, while in Besòs and Girona less than half of the samples (4 out of 9) have full removal. Here protein removal is understood as the absence of identified proteins or the presence of only one peptide or one PSM pointing to them. Since this is the first research to study the proteome in the effluent of treatment plants, direct comparisons with available bibliography are not possible. Nevertheless, this topic has been widely studied in the small molecules field.

Small molecules have different behaviors due to their distinct nature, making it more difficult to draw general conclusions about removal efficiency. In some compounds it depends on the

quantity, if this is very high, the treatment process is saturated and the efficiency is lower. Some similar studies have been made in different wastewater treatment plants in Spain, like the ones carried out by Santos et al. (2009) and Gros et al. (2010). The authors of these researches monitored different types of pharmaceuticals in the influent and the effluent of several WWTP in Sevilla city and the Ebro river basin, respectively. Both studies conclude that most of the pharmaceuticals were found both in the inlet and the outlet of the treatment plants with different removal efficiencies, even though the wastewater treatments were similar. These findings align with our observations: in most of the cases they are found before and after the treatment, although in much less quantity in the effluent. Notably, phenazone and 2-OH CBZ are more abundant in the effluent than in the influent, suggesting these compounds are products of another compound's degradation. Gros et al. (2010) proposed that the higher the hydraulic retention time the better compound removal, which is in accordance with our observations in proteins, but depends very much on the compound and its quantity in the case of the small molecules. However, for small molecules, it seems Besòs is usually the one with the lowest removal efficiency and it is the WWTP with the shortest retention time. It is important to note that this study also focused on antibiotics, pesticides and EDCs apart from pharmaceuticals.

Interestingly, but not surprising, the season in which the samples are collected seems to also influence the removal efficiency of the proteins. Among the spring, summer and winter samples analyzed, winter showed the lowest removal efficiency with only one sample reaching the 100% of removal, while spring is the season with the maximum efficiency as no proteins are identified in any of the samples. This lack of identification could also reflect limitations in detection sensitivity. However, the instruments used for this study are one of the most advanced currently available and well-suited to the sample complexity. More plausibly, this seasonal variation could result from changes in compound concentration and the influence of parameters like rain or temperature that can alter the operation of the treatment plants. Santos et al. (2009) reported similar seasonal trends. A recent study investigated the effects of the wet and dry months in the removal efficiency in a cohort of compounds. The lowest average removal efficiency was determined from the samples taken after heavier than usual rainfall which led to higher flow rates observed in the treatment plant (Inarmal & Moodley, 2024). This matches our own experience, as the samples in winter were collected days after a strong rainfall. However, for the small molecules studied here we have not observed a strong influence from the collection season, being the removal efficiency more depending on the site and the compounds concentrations. In general, the concentrations in the effluent are higher in winter and lower in spring, but so are the concentrations in the influent.

The correlation between the small molecules and the proteins in wastewater has rarely been studied. As a result of the elucidation of the influent proteome, our group hypothesized that proteins could be also used as biomarkers (Carrascal et al., 2020; Carrascal et al., 2023) expanding on the theoretical framework proposed by Rice & Kasprzyk-Hordern (2019). In this study, the same samples were analyzed both for small molecules and for proteins, which allowed a first step of the verification. However, the correlations were not as clear as expected. For example, caffeine is commonly used as a population biomarker and has recently been used to normalize SARS-CoV-2 concentrations (Hsu et al., 2022). Caffeine is excreted via urine and joins the amylases excreted via feces by the same individual in the sewage. However, we did not find a good correlation between caffeine and human amylases. This could be due to the difference in excretion pathways as well as to the lack of a good quantification both for the caffeine due to the lack of reproducibility and for the proteins since no internal standards were used. However, the results obtained from the second campaign (summer) are very promising. Therefore, quantification must be improved in both cases in order to find the expected good correlation and more research is needed to truly conclude the feasibility of human amylases as population markers.

Another interesting proposal from Carrascal et al. (2023) was the use of livestock albumins as markers for meat industry activity. Livestock are often treated with antibiotics and sometimes these will not be metabolized before the animal's life ends, therefore the compound would remain in the blood and end up in the sewage. This means that antibiotics and albumins could follow similar patterns and would be important since the proteomics field can point to the animal of origin. In this study, a possible link was observed between the antibiotic doxycycline, and pig and chicken albumins. Notably, tiamulin is used exclusively in veterinary medicine, specifically for swine (pig) and poultry (chickens) infections (Laber & Schütze, 1977), so it would be expected to have a pattern close to the pig and chicken albumins (Figure 32C). In the case of humans, antibiotics would be metabolized and excreted via urine, thus the correlation should be observed with the human amylases instead of the albumin. Further targeted experimental designs are needed to fully explore these potential correlations.

The upstream and the downstream part of the river of two of the studied WWTPs were also studied using different starting volumes, sample preparation and MS methods to broaden the number of identifications, as it was expected to be a very low quantity of proteins (only one of the workflows is shown here as the results were very similar). This is also the first study to focus on the whole range of proteins present in this kind of samples. Previous studies have mostly focused on the bacterial contamination of rivers that are the source for drinking and

domestic water. This bacterial contamination could endanger the quality drinking water supply for the public by reducing the efficiency of the treatment plants (Thilakarathna et al., 2025).

We can conclude that, even though WWTPs are not specifically designed to remove proteins, our results demonstrate that they do so effectively. Besides, the river streams before and after the studied treatment plants are cleaned at protein level, meaning there is not significant contamination from sewage pipes, industrial activities or surface runoff. It is also safe to say that the few proteins found are not a hazard for humans or other organisms.

5. CONCLUSION

In this study, we analyzed the influent and the effluent of two WWTPs, and the receiving waters of two of them in order to characterize the proteins and small molecules present. This is the first time that the efficiency of the treatment plants in the removal of proteins is determined. The protein profiles identified in the influent were consistent with those reported in previous work confirming the presence of abundant proteins such as human amylases and livestock albumins. However, the protein composition in the treated and receiving waters were quite different, demonstrating that wastewater treatment processes are effective in reducing the majority of proteins. Nevertheless, a small number of proteins persist, including human amylases, livestock albumin, some bacteria proteins and human keratins. This could be due either to the high concentration in the entry or their resistance to the degradation, as treatment plants were not built to enhance the removal of proteins. It is important to remark that these proteins do not suppose a hazard for humans when they are released into the environment. A key strength of this study lies in the combined analysis of proteins and small molecules together, which has allowed us to identify the source of the chemical compounds in the wastewater. For example, results suggest some antibiotics are more used in animals (doxycycline in pigs and chickens) as their concentration is higher in places with an important livestock component, while others antibiotics are more specific in zones with high human population (like azithromycin, ciprofloxacin or tetracycline among others). Thus, proteomics can be used to complement the information from other omics regarding environmental surveillance.

4.4 NON-TARGET PROFILING OF THE WASTEWATER METABOLOME USING A SUITE OF HRMS TOOLS: A STUDY ACROSS DIVERSE TREATMENT PLANTS

Objective 4: Expanding the Omics toolbox: Complementary metabolomic profiling of wastewater influent.

There are different approaches for the study of the metabolites present in wastewater. The most common and easy is the targeted one for metabolites that are already known. In recent years, a new approach has emerged called suspect screening, in which an inclusion list is used to prioritize some metabolites. However, there are not many studies using non-targeted approaches as the identification of new compounds can be very complex and time-consuming. In this study, non-targeted methods are used to enhance the identification of metabolites in 5 WWTP (Besòs, Girona, Olot, Vic and Banyoles) already profiled by proteomics. Moreover, the metabolite information can complement the one given by the proteins. This chapter was accomplished through an internship in the University of California Davis with a grant from the Spanish National Research Council (CSIC).

ABSTRACT

Targeted metabolomics is commonly used to study wastewater aiming at the qualitative and quantitative characterization of drugs and pharmaceutical products. These metabolites constitute the metabolome which provides information about the human health of the population served by the sewage system under survey. The study of the metabolome has become more popular with the development of mass spectrometry (MS). In this study, we analyzed the influent water from 5 WWTPs in winter, spring and summer. We employed gas chromatography (GC), reverse-phase liquid chromatography (RPLC) and hydrophilic liquid chromatography (HILIC), all coupled to mass spectrometry (MS). Moreover, we used an untargeted approach. The main objective was to identify as many compounds as possible and elucidate their probable origins. We identified a total of 828 among the three platforms. We achieved extensive metabolome coverage with minimal overlap among the platforms used, reflecting they were highly complementary. The main groups of compounds based on their chemical composition leading sample clustering are organic oxygen compounds, organic acids, organoheterocyclic compounds, benzenoids and lipids, suggesting they are either the ones most persistent (i.e., not easily degradable) or the most abundant in the source. It is not surprising that amino acids were identified in all the samples as they are the constituents of the proteins which have already been identified in influent waters. Monosaccharides are also commonly detected in wastewater either from the human excreta or the waste food. Interestingly, very long-chain fatty acyls (≥22 carbon atoms) are more annotated in Besòs, Girona and Olot, locations with higher human population and less industrial activity; these fatty acyls could originate not only from humans, but also from the pharmaceutical, chemical and cosmetic products they use. The same occurs with organoheterocyclic compounds, which are utilized in the development of drugs, pharmaceutical products and agrochemicals and were less significant in Vic and Banyoles, likely due to the higher industrial activities of these sites.

This difference is also pointed out by the presence of benzene class compounds, which were only detected in Besòs, Girona and Olot. Determining the precise origin of compounds present in wastewater remains challenging, as they cannot be assigned a definitive source as proteomics does when assigning species. This study marks a starting point, but further research is needed to deepen our understanding.

1. INTRODUCTION

Metabolomics is defined as the qualitative and quantitative analysis of the metabolites in biological samples such as cells, tissues, organisms, and biofluids. These metabolites constitute the metabolome that can be found in a specific context and sample. It is believed that metabolite levels reflect the cellular state. In this sense, the metabolome characterization would be vital to understanding the phenotype of the organisms. In environmental science studies, metabolomics is used together with other "-omic" technologies to characterize the impact of different stressors on ecosystems and human health (Bedia, 2022; Soga, 2007).

The study of the metabolome has become more popular with the development of mass spectrometry (MS). This technique allows different approximations depending on the study's objective. These approaches include targeted, untargeted or suspect screening methods. Targeted methods are more sensitive and reproducible but its coverage is limited. Untargeted methods allow for broader compound identification, but such identification of unknown compounds is challenging due to limited analytical standards and spectral data in public databases. A recent intermediate approach between targeted and untargeted is suspect screening, that prioritizes relevant compounds for case study. It uses a list of prioritized suspects from in-house libraries or databases to be evaluated and identified at higher or lower confirmation level. This method decreases the rate of false positives minimizing the need for extensive manual annotation (Bedia, 2022). All these MS-based approaches are usually coupled to chromatography techniques, such as gas chromatography (GC) and liquid chromatography (LC). Depending on the types of compounds to study, one or a combination of several can be used (Tolstikov et al., 2007). GC is mainly used for the analysis of volatile compounds, but it needs a two-step chemical derivatization for the substitution of carbonyl moieties through methoxyamination and the incorporation of a per-silylation (Erban et al., 2007). On the other hand, LC separations are better for labile, high molecular weight and nonvolatile polar compounds in their natural form. Reversed-phase (RP) chromatography is used for hydrophobic, capillary electrophoresis for hydrophilic and charged small molecules, and hydrophilic (HILIC) separation for hydrophilic and neutral compounds (Tolstikov et al., 2007).

In this study, we aimed to comprehensively profile the metabolome entering WWTPs and tracing the origin. For this purpose, we analyzed the influent water from five wastewater treatment plants (WWTPs) that serve human populations of different sizes and industrial activities, collected at three different times of the year (winter, spring and summer). Furthermore, we used different platforms available at the time (GC-MS, RPLC-MS and HILIC-MS) to maximize the coverage of compounds identified. We also used an untargeted approach to widen the types of compounds that could be identified. For the classification we focused on a purely structure-based chemical taxonomy. To our knowledge, this is the first attempt to characterize the wastewater metabolome using a non-target approach and a chemical classification in order to differentiate among sites and seasonal variations.

2. MATERIALS AND METHODS

2.1 Sample collection

Twenty-four-hour composite wastewater samples were collected at the inlet of 5 wastewater treatment plants (WWTPs) located in the Barcelona and Girona provinces in Catalonia (Spain) (Figure 35), through three collection campaigns: winter (14th of December 2020), spring (19th of April 2021) and summer (26th of July 2021), one sample per site and campaign (15 samples in total). The collection was carried out by an automatic water sampler and the samples were transferred to the laboratory at 4 ºC. These samples were also used for a previous proteomics analysis (Carrascal et al., 2023).

2.2 Sample preparation

Up to 100 ml of each twenty-four-hour composite wastewater sample was centrifuged at 4000 x g (10 ºC, 20 min) and the supernatant was filtered through 0.2 µm filters (VWR, North America, USA). In order to reduce the sample volume, the filtered samples (25 ml aliquots) were lyophilized using a freeze-dryer, then reconstituted in 1 ml of 50% methanol and finally evaporated to dryness using a SpeedVac (SPD130DLX Vacuum Concentrator, Thermo).

For the analysis, one of the 25 ml aliquot of each sample and campaign was thawed for 10 min and reconstituted in 1.5 ml of 50% methanol. Next, each sample was divided into 5 ml and 1 ml aliquots (300 µl and 60 µl, respectively) per quadruplicate. Then, all the samples were dried down in the centrivap (Labconco) and stored at -80 °C until extraction.

Figure 35. Location of the wastewater treatment plants.

The compounds were extracted using the methyl-tert-butyl protocol modified from Matyash et al. (2008). Briefly, 975 µl of ice-cold 3:10 MeOH/MTBE with QC mix (see Supplementary material: 4.4_Supplementary_Tables.docx (Table S1)) were added to each sample. After vortexing, shaking and sonication, 188 µl of LC/MS grade water were added. After another round of vortexing and sonication, samples were centrifuged at 14000 x g for 2 min and 3 phases were obtained: upper organic phase, bottom aqueous phase and precipitated pellet. The upper organic phase of each sample was transferred to 2 separate tubes (350 µl per tube) for RPLC-MS analysis. The bottom aqueous phase of each sample was transferred to 2 separate tubes (110 µl per tube) for GC-MS and HILIC-MS analysis. The extracted samples (RPLC-MS, GC-MS and HILIC-MS) were dried down by centrivap.

2.3 Mass spectrometry analysis

2.3.1 Reversed-phase liquid chromatography coupled to mass spectrometry (RPLC-MS)

Dried samples were reconstituted in 110 µl of 90:10 MeOH:Toluene with 50 ng/ml CUDA (12-(cyclohexylcarbamoylamino)-dodecanoic acid). Then, they were separated using a Waters Acquity Premier BEH C18 VanGuard FIT (2.1 x 50 mm; 1.7µm) coupled to a Waters Acquity Premier BEH C18 VanGuard FIT Cartridge (2.1 x 5 mm; 1.7µm). The column was maintained at 65 ◦C with a 0.8 ml/min flow rate. The positive ionization mobile phases consisted of (A) 60:40 (v/v) ACN:H_2O with 10 mM ammonium formate and 0.1 % formic acid, and (B) 90:10 (v/v) IPA/ACN with 10 mM ammonium formate and 0.1 % formic acid. For negative mode, 10 mM ammonium acetate was used as the only modifier. The injection volumes were 3 µl for positive mode and 10 µl for negative of each sample. The Agilent 6546 Q-TOF mass spectrometer (MS) coupled to an Agilent 1290 Infinity ultra-high performance liquid chromatography (UHPLC) was operated in both positive and negative electrospray ionization (ESI) modes.

2.3.2 Gas chromatography coupled to mass spectrometry (GC-MS)

Previous to injection, samples were derivatized as follows. First, they were reconstituted in 10 µl of MeOx (methoxyamine hydrochloride) and shake for 1.5 h at 30 °C and 750 rpm. Then, 90 µl of MSTFA (N-methyl-N-(trimethylsilyl)-trifluoroactamide) with a mix of 13 FAMEs (fatty acid methyl esters) internal standards (see Supplementary material: 4.4_Supplementary_Tables.docx (Table S3)) were added and samples shaken again for 30 min at 37 °C and 750 rpm. Finally, they were separated using a Restek RTX-5Sil MS column (30 m length, 0.25 mm i.d., and 0.25 µm 95 % dimethyl 5 % diphenyl polysiloxane film) with a 10 m guard column. A total of 0.5 µl of each sample was injected and acquired on the Leco Pegasus BT TOF-MS coupled to an Agilent 7890 B gas chromatograph with Agilent 7693 Autosampler.

2.3.3 Hydrophilic interaction liquid chromatography coupled to mass spectrometry (HILIC-MS)

Dried samples were reconstituted in 200 µl of 80:20 (v/v) ACN:H_2O with 42 internal standards (see Supplementary material: 4.4_Supplementary_Tables.docx (Table S4)). Then, they were separated on a Waters Acquity Premier BEH Amide VanGuard FIT (2.1 x 50 mm; 1.7 µm) coupled to a Waters Acquity Premier BEH Amide VanGuard FIT Cartridge (2.1 x 5 mm; 1.7 µm). The column was maintained at 45 ºC with a 0.8 ml/min flow rate. The mobile phase

consisted of (A) H_2O with 10 mM ammonium formate and 0.125 % formic acid, and (B) 95:5 (v/v) ACN/H_2O with 10 mM ammonium formate and 0.125 % formic acid. The injection volume was 5 µl (+/-) per each sample. The SCIEX 6600 TripleTOF MS coupled to an Agilent 1290 Infinity UHPLC was operated in both positive and negative ESI modes.

2.4 Compound identification

Internal standards were used for the correction of the retention times (see Supplementary material: 4.4_Supplementary_Tables.docx (Tables S2 and S5)). GC-MS raw files were converted into Abf format using the Reifycs Abf Converter (https://www.reifycs.com/abfconverter/). The acquired data from all the platforms and modes was analyzed separately by MS-DIAL software version 4.9 (Tsugawa et al., 2015). Detailed parameter settings for each case are listed in Supplementary material (4.4_MS-Dial_Parameters.xlsx). Metabolite annotations were done using the in-house developed mass-to-charge-retention time (m/z-RT) libraries from Fiehn's lab. MS/MS spectral matching was performed using freely available MS/MS libraries obtained from the Mass Bank of North America (MoNA) (www.massbank.us) and the NIST20 MS/MS library.

MS-DIAL annotations were curated using modified MSI (Metabolomics Standards Initiative) levels (Sumner et al., 2007). Detailed description of the used levels is in Supplementary material (4.4_MSI_Levels.docx). Results were exported replacing zero values with 1/10 of minimum peak height over all samples. Annotations from both modes of RPLC-MS and HILIC-MS were filtered by Fold 2 >5 and sample maximum >1000; then, the filtered annotations was run through MSFlo (https://msflo.fiehnlab.ucdavis.edu/#/) (DeFelice et al., 2017) in order to identify ion-adducts, duplicate peaks and isotopic features. Annotations from GC-MS were filtered by Fold 2 >3, total score >70 and S/N average >3. The parameters used for running MSFlo can be found in Supplementary material (4.4_MS-Flo_Parameters.xlsx).

2.5 Compound classification

Annotated compounds were classified using their International Chemical Identifiers (InChIKeys). First, all the InChIKeys were collected using CTS (Chemical translation Service) batch conversion (https://cts.fiehnlab.ucdavis.edu/batch) (Wohlgemuth et al., 2010) and PubChem (https://pubchem.ncbi.nlm.nih.gov/). Then, all the InChiKeys were run through ClassyFire Batch (https://cfb.fiehnlab.ucdavis.edu/#/) (Djoumbou Feunang et al., 2016) or Ref-Met (https://www.metabolomicsworkbench.org/databases/refmet/name_to_refmet_form.php) to find the classification.

2.6 Data treatment

Datasets (GC-MS, HILIC-MS, and RPLC-MS) were preprocessed and visualized using the R programming language (*R: The R Project for Statistical Computing*, n.d.) with the "Tidyverse" collection of packages, in particular "dplyr", "tidyr", "stringr", "purrr" and "ggplot2" (https://www.tidyverse.org/). This analysis was done in collaboration with Gianluca Arauz from the Institute for Research in Biomedicine (IRB Barcelona). Since RPLC-MS and HILIC-MS metabolites come from positive and negative modes, duplicated metabolites were removed, retaining the mode with the highest intensity or more compound information. All the GC-MS metabolites were kept. The quantification values for all three metabolomics datasets were transformed by adding one unit and taking the base-2 logarithm. This one-unit addition ensured a positive quantification matrix after the $\log_2$ transformation. Finally, quantile normalization (Bolstad et al., 2003) was applied to the quantification values for all three metabolomics datasets.

The strategy followed for the subsequent differential abundance analysis (DAA) at the metabolite level was analogous to that typically used in other omics disciplines such as genomics or proteomics (Rosati et al., 2024). All possible pairwise comparisons between the 5 WWTPs were defined and investigated (10 in total). The major steps of the DAA were carried out by leveraging the “limma” R package (Ritchie et al., 2015). Specifically, a linear model with the independent variables “WWTP” (Banyoles, Besòs, Girona, Olot, Vic) and “Campaign” (Camp1, Camp2, Camp3) as fixed effects was devised for all three metabolomics datasets:

$$I = \beta_0 + \beta_1 \cdot WWTP + \beta_2 \cdot Campaign + \varepsilon$$

where I represents the metabolite intensity after data preprocessing, the β_i are the regression coefficients for the metabolite, and the ε is the error term. After getting the linear model, the function “eBayes” from the “limma” R package was used to get compute t-statistics, F-statistic, and log-odds of differential abundance by empirical Bayes moderation of the standard errors towards a global value (Smyth, 2004). The resulting p-values were FDR-corrected for multiple hypothesis testing using the Benjamini & Hochberg procedure (Benjamini & Hochberg, 1995). Standard cutoffs for significance were applied by requesting |FC| > 1.5 to the absolute value of the fold change and $p_{adj} < 0.05$ to the adjusted p-value, and the significant metabolites for each of the 10 WWTP pairwise comparisons were defined.

After the DAA (see Supplementary material 4.4_DifferentialAbundanceAnalysis.xlsx), for each platform, all metabolites being significant in any of the 10 WWTPs pairwise comparisons were further analyzed and visualized by means of a hierarchical clustering heatmap. The corresponding three intensity submatrices with the significant metabolites were corrected by

subtracting the variability contribution associated with the “Campaign” variable and z-score standardized before their visualization.

In parallel to the DAA strategy just described, all metabolites quantified by each platform were investigated by means of a Principal Component Analysis (PCA) using their corresponding quantification matrices, both, before and after subtracting the variability contribution associated with the “Campaign” variable.

3. RESULTS

3.1 Compound annotations

After the search with MS-DIAL, 23824 features with 314 references were matched for positive RPLC-MS, 7555 features with 209 references for negative RPLC-MS, 2102 features with 484 references for GC-MS, 6679 features with 503 references for positive HILIC-MS and 3025 features with 382 references for negative HILIC-MS. After the manual curation and MSI level classification, the final number of annotated compounds based on their MSI level are summarized in Table 13. Finally, after filtering, merging of positive and negative modes for RPLC-MS and HILIC-MS, and removing duplicates, we had 142 compounds for RPLC-MS, 254 for GC-MS and 498 for HILIC-MS. The complete list with the annotated compounds with their corresponding classification and relative intensities can be found in Supplementary material (4.4_Annotated_CompoundsList.xlsx). There is nearly non-overlap among the platforms used, reflecting they were highly complementary (Figure 36A).

Table 13. Number of annotated compounds in each MSI level (described in Supplementary material: 4.4_MSI_Levels.docx) per platform and mode.

Platform	Mode	Total	Internal standards	Level 1	Level 2	Level 3	Level 4
RPLC-MS	Positive	248	15	8	-	201	24
	Negative	153	13	7	2	108	23
GC-MS	Positive	374	13	39	170	75	77
HILIC-MS	Positive	396	34	30	116	138	78
	Negative	355	28	53	62	149	63

3.2 Compound classification

Regarding the types of compounds identified on each platform, we could see that the extraction method and the chromatography technique used were very specific. In RPLC-MS (Figure 36B), most compounds (92%) were lipids or lipid-like molecules, with only 10 compounds classified as something else. These lipids were divided into 5 groups: fatty acyls (40%), sphingolipids (31%), glycerolipids (18%), glycerophospholipids (8%), and steroids (3%). In GC-MS and HILIC-MS, less than 20% of the identified compounds were lipids (18% and 11%, respectively). Although the lipid classes were the same as in RPLC-MS, some prenol lipids were also identified in both GC-MS and HILIC-MS.

In GC-MS (Figure 36C), there was not a predominant type of compound. The 2 most abundant groups were organic acids and derivatives (26%) and organic oxygen compounds (27%), followed by the lipids mentioned above. Other compounds belong to benzenoids, organoheterocyclic compounds, organic nitrogen compounds, phenylpropanoids and polyketides, and nucleosides, nucleotides and analogues.

In HILIC-MS (Figure 36D), nearly half of the identified compounds were organic acids and derivatives (45%). However, there were a few types of compounds that had not been found in the two previous platforms. These groups included alkaloids and derivatives, lignans, neolignans and related compounds, and organosulfur compounds. Even though the other groups were the same as in GC-MS, specific compounds within them differed. As shown in Figure 36A, only 55 compounds are shared between GC-MS and HILIC-MS.

3.3 Principal Component Analysis (PCA)

The Principal Component Analysis (PCA) of the compounds exhibited first of all a great technical reproducibility in all the campaigns and sites regardless of the platform used. When looking at those platforms independently GC-MS (Figures 37A) showed a trend of site-based grouping. In general, the campaigns clustered near each other, except for Banyoles, where the three campaigns appeared distinct, and Vic, where campaign 3 was more similar to Banyoles. Girona and Olot grouped together, while Besòs remained slightly apart. Overall, campaign 3 differed from the other two. When the campaign-associated variability was removed (Figure 38A), the grouping trend became clearer.

The HILIC-MS showed a similar distribution of the PCA (Figures 37B and 38B) to the GC-MS compounds (Figures 37A and 38A). Although in this case, Besòs seemed to group closer to Girona and Olot. On the contrary, RPLC-MS (Figures 37C and 38C) showed a different

grouping pattern, with Besòs and Girona standing out, while Banyoles, Olot, and Vic clustered closely together. Additionally, the campaign-associated variability was lower than that observed in GC-MS and HILIC-MS.

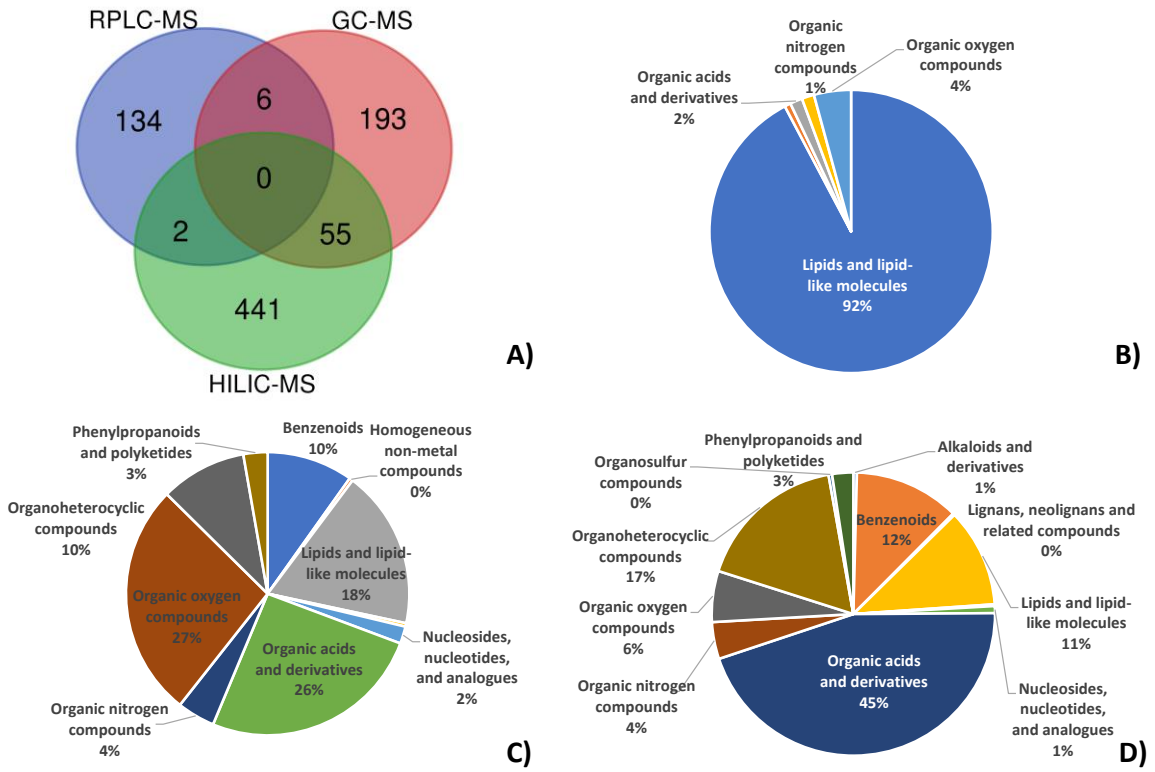


Figure 36. Venn's diagram of the identified compounds in each platform (A) and the corresponding classification of the annotations in RPLC-MS (B), GC-MS (C) and HILIC-MS (D).

3.4 Differential abundance and hierarchical clustering analyses

After filtering by |FC| > 1.5 and p_{adj} < 0.05 the number of significant metabolites for each platform was 122 for RPLC-MS, 241 for GC-MS and 485 for HILIC-MS. These metabolites were clustered in 6 groups in each platform in the hierarchical clustering heatmap (see Supplementary material 4.4_Heatmaps.pdf). Some tendencies per platform and some common patterns were observed. In general, Besòs and Vic seem to be the most opposed sites and they have different upregulated clusters that characterize them in all the platforms. Interestingly, the campaign 3 of Vic is grouped closer to Banyoles than to the other two campaigns of Vic in both GC-MS and HILIC-MS, as it was already observed in the PCAs. The composition of the upregulated cluster in each of the sites and platform was studied in order to elucidate the kind of compounds driving them.

The main groups of compounds leading the clustering of the samples were organic oxygen compounds, organic acids, organoheterocyclic compounds, benzenoids and lipids (Table 14). Regarding lipids, fatty acyls were found to be present in all the clusters across all the samples and platforms. In general, they were long-chain and very long-chain fatty acyls, hydroxy fatty acyls and acyl carnitines. The other two classes of lipids were sphingolipids and glycerolipids; these compounds were present in high proportion in all the WWTPs, except Girona where glycerolipids were absent. The glycerolipids were mostly triacylglycerols. As for the sphingolipids, they included ceramides (Besòs, Girona and Olot), long-chain ceramides (Besòs, Banyoles, Girona and Olot) and neutral glycosphingolipids (Besòs and Vic).

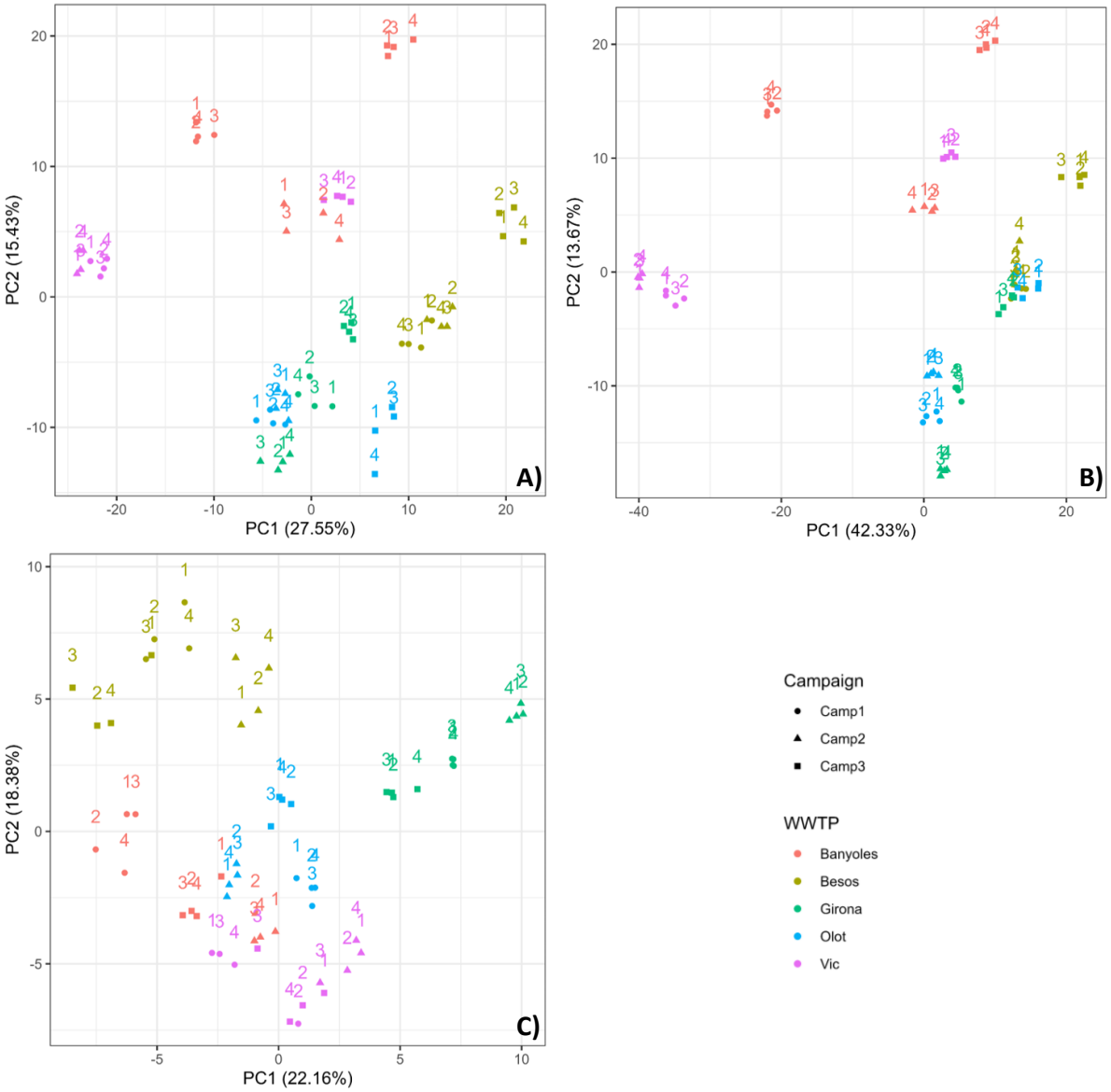


Figure 37. Principal Component Analysis (PCA) of the compounds from GC-MS (A), HILIC-MS (B) and RPLC-MS (C) without removing the campaign-associated variability.

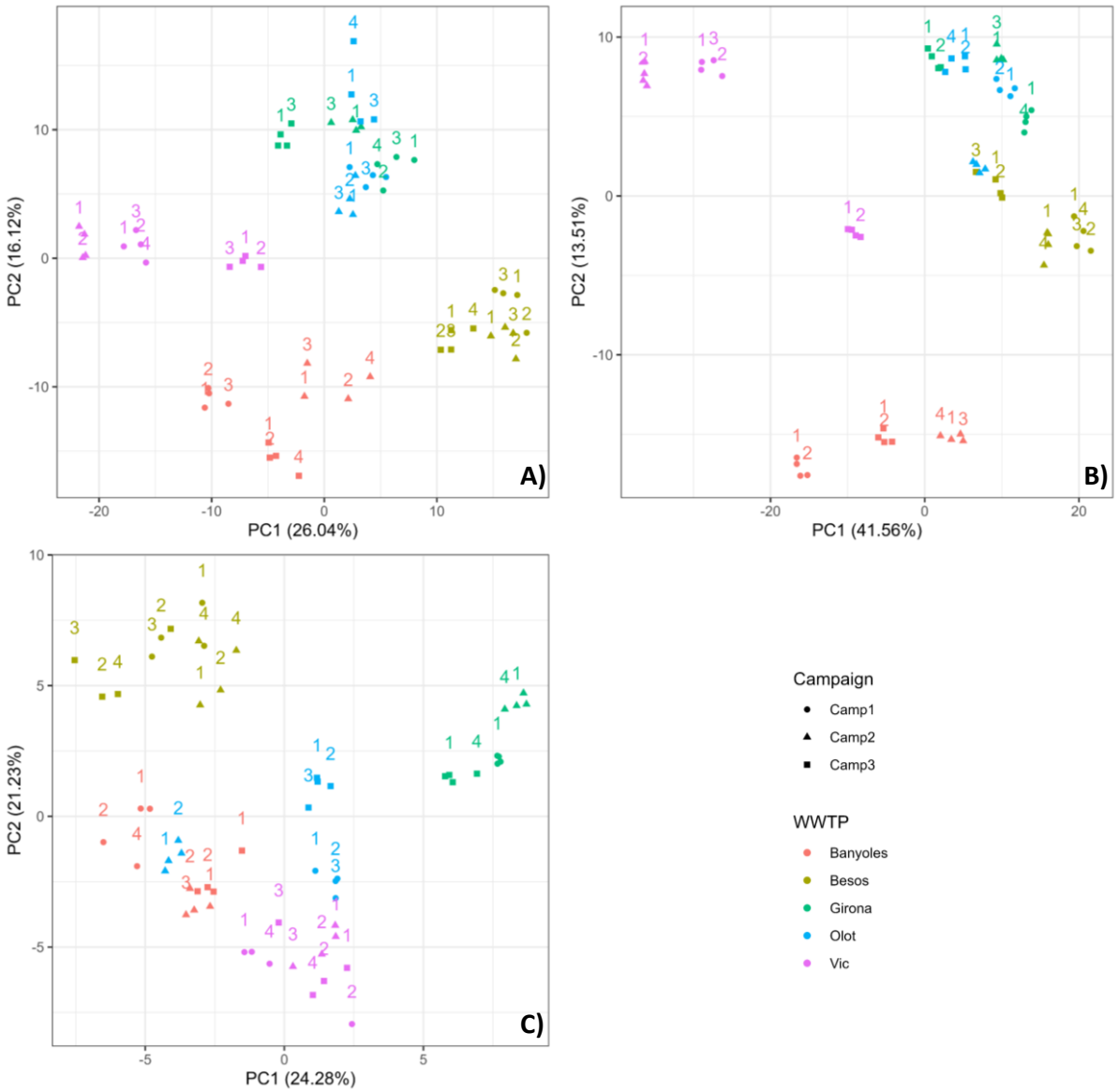


Figure 38. Principal Component Analysis (PCA) of the compounds from GC-MS (A), HILIC-MS (B) and RPLC-MS (C) removing the campaign-associated variability.

It appeared that the combination of lipids could differentiate each WWTP. Sometimes these lipids were only present in specific campaigns: long-chain fatty acyls in Olot's campaign 1, sphingolipids in Olot's campaign 3 and triacylgycerols in Vic's campaign 1. Benzenoids were found in Besòs, Girona and Olot, but not in Banyoles and Vic. Organoheterocyclic compounds were found in all the sites, except Vic. Organic acids were primarily amino acids and peptides (dipeptides) and were present in all the sites and campaigns. Organic oxygen compounds were monosaccharides, glycosyl compounds and sugars. Monosaccharides were found in all the sites, glycosyl compounds in Besòs, Girona and Olot, and sugars in Besòs, Girona, Olot and Vic (mainly alcohols). Banyoles was the only one with only one type of organic oxygen compounds.

Table 14. Types of compounds present in each treatment plant.

Superclass	Class	Parent level 1	Banyoles	Besòs	Girona	Olot	Vic
Lipids	Fatty acyls	Long-chain	✓	✓	✓	✓ (camp 1)	✓
		Very long-chain		✓	✓	✓	
		Hydroxy	✓	✓		✓	✓
		Acyl carnitines			✓	✓	
	Sphingolipids	Ceramides		✓	✓	✓ (camp 3)	
		Long-chain ceramides	✓	✓	✓	✓	
		Neutral sphingolipids		✓			✓
	Glycerolipids	Triacylglycerols	✓	✓		✓	✓ (camp 1)
Benzenoids	Benzene	-		✓	✓	✓	
Organo-heterocyclic	-	-	✓	✓	✓	✓	
Organic acids	-	Amino acids and dipeptides	✓	✓	✓	✓	✓
Organic oxygen	-	Monosaccharides	✓	✓	✓	✓	✓
		Glycosyl compounds		✓	✓	✓	
		Sugar acids and alcohols		✓	✓	✓	✓ (alcohol)

4. DISCUSSION

Although the small molecules present in wastewater have been studied for years, these studies usually relied on a small group of compounds. Here, we used five WWTPs, three campaigns, three different platforms, and an untargeted approach to carry out the first attempt to create a comprehensive profile of the metabolome found in the wastewater. The samples were collected in different seasons to broaden the types of compounds that could be present. Moreover, we sampled in sites where the human population and the industrial activity are different. For example, Besòs is the site with the largest human population (around one and a half million people) and no industrial activity, while Banyoles and Vic are known for their significant industrial activity, with much smaller human populations.

We used the already stablished methods in Fiehn's laboratory at the West Coast Metabolomics Center (University of California Davis). These methods are routinely used for this group in different kinds of matrix, such as tissues and serum (Vaniya et al., 2024; Questa et al., 2024). We identified a total of 828 unique compounds. It is difficult to justify if this is a high or a low number of identifications since the total number of metabolites present in wastewater is unknown and very dependent on the source. This number could likely be increased using other technologies, such as nuclear magnetic resonance (NMR) spectroscopy or inductively coupled plasma mass spectrometry (ICP-MS), by increasing the number of samples with sites with different characteristics, or by analyzing the particulate fraction (this study focused on the soluble one only). The choice of chromatography separation in all the techniques was made out of the availability at the time, but due to the complexity of the samples was the best option to avoid as much as possible the interference among different compounds, as it could occur when using direct injection.

The wastewater that enters a treatment plant is composed of the human excreta/biofluids (urine, feces, blood), the products and chemicals people use or consume, and the waste from the industrial activity surrounding them (agriculture, slaughterhouses). This makes determining the origin of the small molecules very challenging, which is joined by the vast variety of small molecules that exist. Because studies usually focus on specific compounds, these are chosen due the use they have in the human population, for example antibiotics, illicit drugs, flame retardants, dyes, etc. This kind of classification is very challenging to do when using a non-target approach. For this reason, we decided to use a chemical classification based on chemical structures and structural features (Djoumbou Feunang et al., 2016). For simplicity we only use superclasses, classes and parent level 1 (see Table 14). In this regard, we can say that benzenoids, lipids, organic acids, organic oxygen and organoheterocyclic compounds are the most common superclasses.

Amino acids were identified in all the samples. This is not surprising as they are the constituents of the proteins, which are essential biomolecules for the correct function of the organisms and are present in the wastewater. Our group unraveled for the first time this wastewater proteome profile of these same samples (Carrascal et al., 2023). We demonstrated that the main components of the wastewater are, at protein level, human urine and feces, and livestock blood. This conclusion aligns with the presence of amino acids in both our samples and as common annotation in the biofluids. In recent years, the Wishart group has focused on studying the human metabolome of each of the biofluids present in humans. For that, they created the Human Metabolome Database (Wishart et al., 2007; Wishart et al., 2009; Wishart et al., 2012; Wishart et al., 2017; Wishart et al., 2021) where they

combined both literature and experimental annotations with special attention at the normal and disease concentrations ranges. To date, they have studied the metabolome of urine (Bouatra et al., 2013), serum (Psychogios et al., 2011), feces (Karu et al., 2018), saliva (Dame et al., 2015), cerebrospinal fluid (CSF) (Wishart et al., 2008) and sweat (*Sweat Metabolome*, n.d.). Amino acids are present in all these biofluids as one of the most common compounds (except for feces).

Another type of compound found in all the samples is the monosaccharides (parent level of the organic oxygen superclass), which are the simplest sugars. These molecules are present in all the biofluids, except feces. The presence of this compound is expected since they are used to store and produce energy in the human body, as well as being part of food, like fruits, vegetables, or honey, both free and combined (Varney et al., 2017). Here, we identified some important monosaccharides, like 1,5-anhydroglucitol (used as a marker of short-term glycemic control (Dungan, 2008)), arabinose (present in plants and used as sweetener in the food industry (Mariette et al., 2021; Hu et al., 2018)), digitoxose (part of steroids, quinones, indoles, oligosaccharides and others, that exhibit antibacterial, anti-viral, antiarrhythmic, and antitumor activities (Li et al., 2024)), dihydroxyacetone (a color additive in sunless tanning products (Braunberger et al., 2018)), fructose (present in fruits, vegetables -including sugar cane-, and honey (Park & Yetley, 1993)), glucose (found in free state in fruits (Siddiqui et al., 2020)), ribose (constituent of numerous essential biomolecules, including RNA, nucleotides, and riboflavin (Tai et al., 2024)), sinigrin (present in broccoli, brussels sprouts, and mustard seeds, among others (Mazumder et al., 2016)), sorbose (found in some fruits and used in the vitamin C manufacturing (Zou et al., 2012)) or tagatose (used as a sweetener in beverages, yogurt, creams, and dietetic candy (Guerrero-Wyss et al., 2018)). Thus, either from the human excreta or the waste food, monosaccharides are a common component of wastewater. This kind of compound is also a good example of the benefits of combining different platforms, as they were mostly identified in just one of the used ones (GC-MS).

An important superclass of compounds in wastewater is lipids. The extraction method used in our study (MTBE method) was specifically developed to enhance their extraction (Matyash et al., 2008). Therefore, it is expected to annotate a wide range of them. Lipids are the second most abundant component when all the platforms are merged, being RPLC-MS the platform that provides more classes, aside from fatty acyls, and offers a different clustering of the samples compared to HILIC-MS and GC-MS. These lipids are very common in the different human biofluids, such as serum, urine or sweat, but they can also have other origins like pharmaceutical, chemical, and cosmetic products consumed by humans (Mohana et al., 2023).

Fatty acyls are identified by all the platforms and are present in all the samples. However, not all the sites have all the types of fatty acyls. Long-chain fatty acids (13 to 21 carbon atoms) are present in all the sites, while very long-chain fatty acids (≥22 carbon atoms) are more abundant in Besòs, Girona and Olot, being the sites with a higher ratio human population to industrial activity, which may suggest a more human origin of these very long-chain fatty acyls. Meat and poultry fats, dairy products, oils (fish, olive, sunflower, palm), nuts and seeds are rich in long-chain fatty acids (Abedi & Sahari, 2014), and all of them are consumed to some degree by humans. Moreover, they are the constituent of the cell membranes in the form of phospholipids and an energy storage in triglycerides (Nakamura et al., 2014). Some examples found in this study are: oleic acid (present in the olive oil, widely used in Spain (Yubero-Serrano et al., 2018)), eicosapentaenoic acid (found in salmon, sardines and tuna, which are source of omega-3 (Tomczyk et al., 2023)), palmitic and stearic acids (abundant in butter and cheese (El-Metwally et al., 2022)) and arachidonic acid (necessary for skeletal muscle growth and function (Korotkova & Lundberg, 2014)). On the other hand, we identified a wide range of these fatty acids from 22 to 32 carbons without double bonds. Very long-chain fatty acids, specifically the saturated ones, are part of the blood cells membrane. This type of fatty acid reflects the quantity and quality of dietary fat intake and influences endogenous lipid metabolism. Besides, some of these, such as behenic acid (FA 22:0) and lignoceric acid (FA 24:0) have been confirmed to have protective effects on cardiovascular disease, coronary heart disease, and all causes of death in the whole, hyperlipidemia, and hypertensive populations (Tao et al., 2023). In the brain, very long-chain saturated fatty acids are enriched in synaptic vesicles and mediate neuronal signaling by determining the rate of neurotransmitter release essential for normal neuronal function. One example identified here is FA 30:0, which are incorporated into complex sphingolipids and enriched in synaptic vesicles (Yeboah et al., 2021).

Another interesting type of lipids is triacylglycerols, which are abundant in all the samples, except in Girona. Triacylglycerols are known to be part of vegetable oils, animal fats and soaps (Mohana et al., 2023). Some of their functions are being the main energy reserves of the human body, and taking part in metabolic processes that determine the rate of fatty acid oxidation, the plasma levels of free fatty acids, the biosynthesis of other lipid molecules and the metabolic fate of lipoproteins (Karantonis et al., 2009). The low abundance observed in Girona does not correspond with these important functions and broad presence, which may be explained by the interference of other compounds.

Organoheterocyclic compounds are those containing rings of atoms, with at least one atom that is not carbon. These compounds are very important for the development of drugs,

pharmaceutical products, and agrochemicals. Some of their functions include the formation of the nucleobases present in genetic materials, treatment of cardiovascular, neurological or gastrointestinal disorders, insecticides, fungicides, dyes, and therapeutic potential (anti-inflammatory, anti-cancerous, etc.) (Tripathi et al., 2021). This explains the detection of these compounds as significant in all the sites and campaigns. However, these compounds are not such in Vic, moreover, Vic is the site with the fewest types of these compounds, followed by Banyoles. This could be explained by their industrial activities. While there is a human population in both sites, it is not as high as in other areas, with the industrial activity being more prominent. As a result, these sites may have a higher proportion of waste originating from industrial sources rather than human sources, though we have not identified any compound to verify this.

There are several classes of organoheterocyclic compounds that are upregulated. Regarding the pyridines, some interesting annotated compounds are those related to tobacco, as metabolites in the degradation of nicotine, such as 2-hydroxynicotinic acid (Tinschert et al., 1997), 3-trans-hydroxynorcotinine and 6-methylnicotinic acid (Jacob III et al., 2011). Moreover, other pyridine metabolites are indicators of vitamin, for example 4-pyridoxic acid is a product of vitamin B6 metabolism, which is excreted in urine (da Silva & Gregory, 2020). In the imidazopyrimidine class, we found stimulants like caffeine and theophylline (both found naturally in coffee, tea, chocolate, and other foods and beverages used daily by human (Bispo et al., 2002)); diseases' biomarkers as uric acid for hyperuricemia (leading to gout) (Heinig & Johnson, 2006); intermediaries of the purine metabolism, such as hypoxanthine acid and xanthine (Kimiyoshi et al., 1993); or the antiviral penciclovir (Schmid-Wendtner & Korting, 2004). Another abundant class is the indole one, where we identified a common naturally occurring plant hormone -3-indoleacetic acid- and many of its derivatives ((2-oxo-2,3-dihydro-1H-indol-3-yl)acetic acid, 5-hydroxy-3-indoleacetic acid, 5-methoxy-3-indoleacetic acid and indole-3-lactate, which participate in plants growth and development) (Tang et al., 2023); a human hormone that regulates the sleep-wake cycle -melatonin- (Claustrat & Leston, 2015); and an essential amino acid -tryptophan- (Richard et al., 2009). In the phenylpyrrole class, we found a common lipid-regulating drug, called atorvastatin (Kogawa et al., 2019)).

Nevertheless, we wondered if we could find the origin of compounds that are not present in one or more of the treatment plants as a way to differentiate those sites with different human populations and industrial activity. A clear example is the presence of compounds belonging to the benzenoid superclass, commonly the benzene class, which are more found in Besòs, Girona and Olot, and less in Banyoles and Vic. The benzene class gathers aromatic compounds containing one monocyclic ring system consisting of benzene; these compounds

are widely used in the production of pharmaceuticals, dyes and pigments, fragrances and flavorings, and plastics and polymers. This leads to the present of these compounds in many of the products and chemicals used by the population daily. At the same time, this is linked to the fact that the sites where they are present are the ones with a higher proportion of human population compared to the industrial activity in the area. This represents a more extreme case compared to the organoheterocyclic compounds. We have found several pharmaceuticals for the treatment of gastrointestinal disorders (5-aminosalicylic acid for the inflammation in ulcerative colitis (Hauso et al., 2015), trimebutine for irritable bowel syndrome (Yu et al., 2025)), mental diseases (amisulpride as an antipsychotic drug (Wang et al., 2025), bifemelane as an antidepressant agent (Fasipe, 2019)) or high blood pressure (losartan and telmisartan (Ogura & Shiraishi, 2025)). Other important identified compounds are 1-phenylethanol (food flavoring (Dong et al., 2017)), 4-aminophenol (discharged in a variety of industries, including petrochemistry, cosmetics, dyes, photography, agricultural chemicals, etc. and toxic for humans (Wang & Feng, 2023)), aniline (precursor of polyurethane-based polymers, used in accelerators, antioxidants, pesticides, dyes, and pharmaceuticals and constituent of tobacco smoke (Modick et al., 2015)), naproxen (a nonsteroidal anti-inflammatory drug (Haghgouei & Alizadeh, 2025)) or O-acetylsalicylic acid (commonly known as aspirin, widely used as analgesic, antipyretic, and anti-inflammatory agent (Kacso & Terézhalmy, 1994)). An important compound which is not from the benzene class but from the phenol ether class is venlafaxine, one of the most-prescribed antidepressants (Suwała et al., 2019).

Here, we attempted to investigate the small molecules present in the entry of wastewater treatment plants. On one hand, as we have shown here, different platforms are complementary and it would be necessary to increase the number of techniques used both for the extraction and analysis. On the other hand, more samples from other areas with different human populations and industrial activities would be needed to increment the types of compounds that could be identified. Lastly, determining the origin of compounds (human, animal, microorganism, industry, etc.) is not straightforward since it is not possible to give them a “name” as it can be done in proteomics. Thus, the only way to address this is by comparing as many different sites as possible to detect distinct profiles. This work is a start, but there is much more to be done.

5. CONCLUSION

In this study, we analyzed the influent water from five WWTPs in winter, spring and summer, using gas chromatography (GC), reverse-phase liquid chromatography (RPLC) and hydrophilic liquid chromatography (HILIC), all coupled to mass spectrometry (MS) in untargeted mode to enhance the number of compounds. We cannot determine whether the number of compounds we identified represents a low or a high proportion of the wastewater metabolome, as this is unknown. However, it is very likely this number can be increased using other extraction methods and platforms for the analysis, as the ones used here are already quite complementary. The main groups of identified compounds were organic oxygen compounds, organic acids, organoheterocyclic compounds, benzenoids and lipids. The origin of the compounds is very complex to elucidate, so we can only estimate it. For example, the most common organic acids are amino acids and dipeptides, and proteins have already been demonstrated to be present in a soluble form in the wastewater. On the other hand, the waste food likely contributes to the high presence of monosaccharides. Nevertheless, we have been able to observe some profiles that differentiate among sites with different population sizes and industrial activities. Very long-chain fatty acyls, organoheterocyclic compounds and benzene are more prevalent in sites with higher human populations. The profiling of wastewater at metabolome level is far from complete, but this study represents a promising start.

5. GENERAL DISCUSSION

This doctoral research has focused on the application of proteomics techniques to wastewater-based epidemiology (WBE), a field that investigates wastewater to gain insights into public health. Wastewater is a complex matrix composed of dissolved organic matter (DOM), natural organic matter (NOM), detergents, personal care products, pharmaceuticals, transformation products and other chemicals (Shon et al., 2006; Hertkorn et al., 2013). Wastewater has long been studied applying other "-omics" disciplines, such as metabolomics and genomics. Initially, the focus was on identifying small molecules like illicit drugs, pharmaceuticals, and other substances to assess collective human behavior and health. More recently, microbial content has been examined, including both beneficial microbes from wastewater treatment sludge (Park et al., 2008a; Zhang et al., 2019) and harmful pathogens such as viruses (Dutta et al., 2021; Mackuľak et al., 2021), which became especially relevant during the COVID-19 pandemic (Lara-Jacobo et al., 2022; O'Reilly et al., 2025).

The first objective of the project was: "*Development of novel non-target strategies for protein monitoring in different water matrices including urban sewage, wastewater treatment plants (WWTPs) effluents (treated water) and water at different stages of its treatment*". In order to achieve this, we selected three very different sampling sites: one with a high human population, another dominated by industrial activity and a third with a mix of moderate human and industrial influence. Previous protein studies in wastewater were mainly related to sludge-associated proteins, primarily of bacterial origin. However, no studies were made outside of these bacteria. The first study to consider non-bacteria proteins was carried out by Carrascal et al. (2020), identifying 690 proteins from bacteria, plants, animals and humans. Although most of them pointed to almost 200 bacterial genera, 57 proteins were from humans, being the highest number for any single species. Among the human proteins, there were biomarkers already described including uromodulin (Garimella & Sarnak, 2016), α-amylase (Mattes et al., 2014), and S100A8 (Wang et al., 2018).

As it arose during the development of the method (Sánchez-Jiménez et al., 2023), wastewater consists of both soluble substances and particulates, such as leaves, food waste, microorganisms, etc. This distinction led to the differentiation of two subproteomes: the soluble and the insoluble (also called particulate). Even though the work here has centered on the soluble fraction, some steps were made into the particulate one. Interestingly, the composition of both fractions is different and complementary: the soluble fraction contains predominantly eukaryotic proteins, while the particulate fraction is rich in bacterial proteins (Carrascal et al., 2023).

As this was the first method developed for the protein characterization in wastewater it prompts improvement. One major limitation is the time it takes to carry it out, around 4 to 5 days from the separation of the fractions to the final peptide extract, which could hinder its application on a daily basis. Currently, new workflows are being explored to reduce this time to ideally one day of sample preparation. For example, PreOmics iST kit (Kulak et al., 2014) enables robust and reproducible sample preparation in a few hours and can be combined with an SP3 Add-on (Hughes et al., 2018) that allows the purification of the proteins, which is an essential step in the case of wastewater samples due to all the types of molecules that can be present as well. The limitation found for this approach is the working range: 1-100 µg using a maximum of 50 µl, which is too concentrated for wastewater. It could be, however, a good option for the particulate fraction, which seems to have a higher protein concentration than the soluble fraction.

Protein identification is also impacted by mass spectrometry settings, particularly the use of data-dependent acquisition (DDA) and the availability of comprehensive reviewed databases. DDA may miss low-abundant peptides. This could be resolved with the use of the data-independent acquisition (DIA) approach, where all the peptides in the specified window ranges are fragmented (Doellinger et al., 2020; Fernández-Costa et al., 2020). However, DIA has seen limited use due to the data analysis bottleneck when handling large databases. Devianto et al. (2024) used DIA to identify biomarkers linked to the incidence of COVID-19 in wastewater. They identified 8866 in total from 19 samples, using 50 ml as starting volume. This demonstrates the feasibility of this approach (Devianto et al., 2024).

In relation with database limitations, the particulate fraction is the most strongly affected since bacteria and viruses' proteomes are poorly represented, particularly in curated database, leading to under-identification. On this line, Tugui et al. (2025) developed a method for the study of the metaproteomics in wastewater. Metaproteomics is a discipline that studies microbial mixtures to provide insights about their composition and functions. Furthermore, it allows the measurement of freely floating proteins, including those excreted by humans or released through industrial and agricultural activities (Wilmes & Bond, 2006; Kleiner et al., 2017). Tugui and colleagues introduces an efficient sample preparation procedure that extracts proteins from both soluble and particulate fractions using only 500 µl of wastewater. This workflow consists in the direct denaturalization and lysis of the sample, followed by reduction and alkylation, and cleaning and digestion using FASP filters. Similar protocols were tried during this PhD with limited success, new workflows are currently being evaluated. Additionally, they created a wastewater metaproteomics data processing pipeline that employs de novo sequencing to focus generic reference sequence databases in order to

maximize metaproteomic coverage (Tugui et al., 2025). However, this data processing also has limitations, as the identifications of the microbes can only be made at family or genus level and not species-level.

The development of the method for the study of wastewater proteins led to the second objective: "*Characterization of potential protein biomarker signatures for early epidemiological alerts and event follow up by correlating population and human health, habits, and activities with sewage protein profiles at different geographical sites*". A modified version of the original protocol was used to extract the protein profiles of 10 sites over 3 seasons, focusing exclusively on the soluble fraction. The analysis revealed two main sources of proteins in wastewater: excreta (urine and feces) from humans, and blood and other residues from livestock. This allows the study of both the human and industrial activities in different environments, like urban and rural areas. Overall, the most abundant protein detected was the human pancreatic α-amylase. However, this protein could be replaced for livestock albumins in sites with a high predominance of industrial activities, such as slaughterhouses.

Amylases are proposed as potential markers of human population and could serve as a potential tool to normalize the abundances of other biomarkers. This is due to their consistently high abundance in wastewater, their probable resistance to protease action (their main role in the intestine reflects their high stability against hydrolytic degradation) and the species-specific information carried in their sequences. As discussed earlier, one of the key challenges in WBE studies is the estimation of the human population served by a WWTP (Daughton, 2012; Hsu et al., 2022). Chemical oxygen demand (COD), biological oxygen demand (BOD), total nitrogen or phosphorus are the parameters frequently used for this purpose, however they are highly unspecific. Census data, while used, may be outdated and does not reflect population dynamics (commute or tourism) (Rico et al., 2017). In recent years, small molecules excreted in urine or feces, such as creatinine, cholesterol, 5-hydroxyindoleacetic acid or caffeine, have been evaluated and used as markers of population (Oloye et al., 2023). Nevertheless, human amylases have the advantage of being virtually free of the contribution from other non-controlled exogenous sources and thus to provide more accurate measurements.

The albumins identified in wastewater probably originates from industrial discharge of animal blood. Although albumins from different species have a high sequence homology, their unique peptides profiles generated by tryptic digestion allow species-level identification. We were able to identify albumin from human, livestock (cow, pig, sheep, rabbit), chicken, cat, dog, rat and mouse. Due to their abundance and species-specific peptide markers, albumins are proposed as indicators of livestock industry activity. Currently, indicators like BOD, COD, or TOC

parameters can be used for pinpointing occasional discharges by routine monitoring of the organic load content in wastewater; however, these methods do not provide information on the contributing molecules or their origin (Bustillo-Lecompte & Mehrvar, 2015). Mass spectrometry based targeted methods could be a powerful and nearly real-time monitoring tool not only for environmental studies assessing the status of a water body but also for regulatory agencies in the surveillance of discharges of animal residues (controlled and uncontrolled) in wastewater systems and receiving waters.

With the knowledge of the wastewater protein profile we moved to the third objective: "*Assessment of the efficiency of wastewater treatment by monitoring samples at WWTP influent, effluent and receiving waters*". All the proteins that enter the treatment plants are subject to the same processes as the small molecules. However, the efficiency of these procedures is not completely known for proteins. In this study we observed that when compared with the influent, the effluent could be considered cleaned from proteins. There were some remaining ones, such as keratins and amylases from humans, albumin from livestock and bacterial proteins. This aligns with the scarce studies available about proteins after the treatments. Park and colleagues were able to identified human-derived proteins in the extracellular matrix of activated sludge flocs and from anaerobically digested sludge product, meaning they were resistant to the secondary aerobic treatment, but they could not answer the question of whether these proteins were released into the receiving waters (Park et al., 2008b). Later, Westgate & Park (2010) confirmed these findings and added other proteins that were produced during the secondary biological treatment.

Even though we have identified the proteins leaving the treatment plants and their organism of origin, it is unknown if these proteins could affect the biota present in the receiving waters. A preliminary study showed that these proteins can bind to bacterial cells and induce changes in the physiological reactions and activity of bacteria, for example accelerating biofilm formation or modifying biofilm microstructure, which could negatively affect downstream environments (Cui et al., 2019). However, more studies are needed to elucidate the possible effects of proteins in the receiving waters.

As it has been already commented, one of the main advantages in the use of proteomics is the capacity to know the origin of the proteins through the variations in their sequences. However, this is not possible in the study of the small molecules using metabolomics techniques as they do not have different configurations depending on the origin. Furthermore, small molecules are mainly studied with targeted methods where you already have a list of the compounds you want to identify. This approach leaves many compounds that are not

identified and, thus, the wastewater metabolome is incomplete. Because of these, we addressed the fourth objective: "*Expanding the Omics toolbox: Complementary metabolomic profiling of wastewater influent*". To carry it out we decided to use an untargeted metabolomic approach to identify as many compounds as possible in samples where the protein profile was already known. Then we studied the superclasses in which the metabolites grouped in order to trace back to their probable origin.

The most common superclasses present in the wastewater were organic oxygen compounds, organic acids, organoheterocyclic compounds, benzenoids and lipids. All of these superclasses can be found in many substances, like food waste, personal care products, biofluids, etc. Amino acid and dipeptides were the most abundant organic acids and their origin the easiest to trace as they have to belong to the proteins already identified. As described previously, the protein origin is mainly human urine and feces, and livestock blood in the studied samples. For the human side, this was already confirmed in the Wishart group's studies, which outline the composition of different human biofluids, such as urine (Bouatra et al., 2013), serum (Psychogios et al., 2011), feces (Karu et al., 2018), saliva (Dame et al., 2015) or cerebrospinal fluid (CSF) (Wishart et al., 2008). Monosaccharides and fatty acyls from organic oxygen compounds and lipids superclasses, respectively, were present in all the studied sites. This suggests a human origin both from the humans themselves or for the products consumed or used by them (food as described by Varney et al., (2017), or pharmaceutical, chemical and cosmetic products (Zhu et al., 2017; Mohana et al., 2023)).

Nevertheless, more interesting were the metabolites that were not present in all the sites. This is the case of very long-chain fatty acyls, which were only identified in sites with low or none industrial activity. This differentiation of places with, and without or low industrial activity was reinforced by benzenoids and organoheterocyclic compounds superclasses, as these types of metabolites come from products used in day-to-day life by the human populations. Some of these products are pharmaceuticals (Tripathi et al., 2021), plastics, dyes, detergents or cosmetics (Thanekar et al., 2021).

It is safe to say that the study of the proteins in wastewater complement the information found through other -omics by clarifying molecular origins. A good example would be antibiotics. These compounds can be used both for human and animals, and end up in the wastewater and sometimes in the receiving waters if the WWTP efficiency is low. For example, kanamycin is a widely used antibiotic in human and veterinary medicine, as well as in food production and livestock breeding, making elucidating its source very challenging (Zheng et al., 2025). In this case, with the use of proteomics and metabolomics, we could obtain more information about

the provenance of the compounds. Although, more research is needed to confirm this, we found in our work a promising example. Doxycycline is an antibiotic used both in humans and animals. However, we found that their concentration was higher in the site with high predominance of livestock than in the site with a majority of human population, where other antibiotics presented the highest concentrations (azithromycin, ciprofloxacin or tetracycline among others).

Limitations and future work

This work is the first step into the analysis of proteins in wastewater and their utility as biomarkers to complement metabolomics and genomics studies. However, there are a number of limitations that have to be overcome before wastewater proteomics becomes a stablished field. Some of them have been already discussed along the work, such as the time needed for sample preparation, loss of protein identification due to the use of a DDA approach and a wide database, or effects of non-cleaned proteins in receiving waters. However, there are more pressing obstacles that have to be cared of, like the stability of the proteins and their limit of detection/quantification.

There is a lack of studies about how stable proteins are in wastewater, let alone about their real absolute concentrations in this matrix, even though they have been proposed as potential biomarkers for several authors now (Rice & Kasprzyk-Hordern, 2019; Amin et al., 2023; Carrascal et al., 2023; Armenta-Castro et al., 2024). Amin et al. (2023) resumed in detail some of the problems that have yet to be addressed: low concentrations leading to challenges in detection and confident quantification; matrix effect from urine, feces, and other biofluids or matters; susceptibility to degradation or transformation within the sewage system due to soluble bacterial enzymes or biofilms in the inner walls of the pipes; and impact of pH, hydrophobicity, temperatures of other parameters. These obstacles make a lot of room for future studies.

Another important line of research to follow is the analysis of the particulate fraction of wastewater. We found the proteins in this fraction to be mostly bacterial, but some biomarkers could also be found. This leads to the hypothesis that some proteins can be attached to the microorganisms present in wastewater, either because they are excreted that way or because they get attached during the period in the sewage system or the storage after collection if not done properly. As in the soluble fraction, sample preparation should be optimized and shortened if possible. Moreover, the use of a complete database is more restrictive in the

analysis of this fraction due to the lack of microorganisms' reviewed proteins. Metaproteomics approaches can help in this case, as discussed above.

Regarding our findings, they offer novel insights into wastewater proteomics, enabling the proposal of specific bioindicators useful for practical applications in WBE monitoring. Applying this methodology to assess wastewater treatment performance becomes crucial in determining the origin of protein contaminants and, consequently, formulating effective measures to reduce their presence. In the future we aim to go deeper into these relevant findings, trying to establish new tools for determining protein markers of the human population that can be used for population data normalization (not only theoretically but experimentally) and for rodent pest detection

About the monitoring of rodent populations in urban areas, protein-related WBE enables the detection of rodent pests in cities and can be a tool with potential for their population control. Rat feces, like human feces, contain proteins that are secreted in the pancreas so, detection of these enzymes in wastewater indicates the presence of live animals, and relative quantification to rodent amylase could allow us to monitor the increase or decrease of rodent feces in these samples. Nowadays, there is no standardized method to determine the numbers of rodents in a city or to estimate the population density, although some groups have tried to understand their population dynamics (Himsworth et al., 2014; Murray et al., 2020). Currently, various strategies are used for the surveillance of these pests, generally based on the count of animals and their extrapolation to the total population (Jurišić et al., 2022). In contrast, we can propose detecting and quantifying rats according to rat feces located in sewage water using some appropriate protein biomarkers.

Data derived from this job point to the possibility of monitoring the general health status of a population using protein health-related biomarkers. More than 400 human proteins have been detected with our methodology and many authors have reported the interest in using protein markers to know the health status of the population emphasizing the need to obtain synergistic information with classic epidemiological studies (Kasprzyk-Hordern et al., 2022; Robins et al., 2023). Recently, Devianto and Sano (2023) published an in-silico study where they describe potential health-related proteins with relatively high concentrations in urine and stool as candidates for protein biomarkers in wastewater. However, no experimental studies have been performed until now directed to the real detection and quantification of these biomarkers in wastewater. Together with the data in the literature and in databases, our experimental results have allowed us to select a list of protein biomarkers related to different human pathologies to be monitored (Table 15).

Table 15. Potential WBE human health-related biomarkers detected in urine and feces (Devianto & Sano, 2023) and detected in this study. Proteins are sorted according to the intensity detected in our experiments.

Protein name	Accession	Source	Related disease
Alpha-1-antitrypsin	P01009	Urine, feces	Environmental enteropathy, bladder cancer
Zinc-alpha-2-glycoprotein (ZAG2)	P25311	Urine	Bladder cancer
Uromodulin	P07911	Urine	Kidney disease
E-cadherin	P12830	Urine	Type 2 diabetes mellitus diabetic nephropathy
Hemopexin	P02790	Urine	Lupus nephritis
Cell adhesion molecule CEACAM-1	P13688	Urine	Bladder cancer
Complement C3	P01024	Urine	Focal segmental glomerulosclerosis
Haptoglobin	P00738	Feces	Gastrointestinal diseases
Neutrophil gelatinase-associated lipocalin (NGAL)	P80188	Feces	Gastrointestinal diseases
Ceruloplasmin	P00450	Urine	Lupus nephritis
Alpha-1B-glycoprotein (A1BG)	P04217	Urine	Immunoglobulin A nephropathy, Henoch-Schöenlein purpura nephritis
Non-secretory ribonuclease	P10153	Urine, feces	Eosinophilic esophagitis
Calprotectin (S100A8, S100A9)	P06702, P05109	Urine, feces	Gastrointestinal diseases, ANCA-associated systemic vasculitis
Matrix metalloproteinase-9 (MMP-9)	P14780	Feces	Crohn’s disease, ulcerative colitis
Complement C5	P01031	Urine	Focal segmental glomerulosclerosis
Insulin-like growth factor-binding protein 7 (IGFBP-7)	Q16270	Urine	Acute kidney injury

6. CONCLUSIONS

1. The developed method is an effective tool for discovering community biomarkers, and represents the first step toward the creation of specific test devices for health and environmental monitoring. Furthermore, it is the first approach directed to the large-scale characterization of proteins in wastewater addressing critical challenges such as the heterogeneity and complexity of the matrix, and the interferences from other molecules.

2. Wastewater proteome is compartmentalized into the soluble and insoluble or particulate fractions. The particulate fraction is rich in bacteria-related proteins, while in the soluble fraction Eukaryotic proteins are the most abundant.

3. The soluble proteins in wastewater transport information on the human and industrial activities occurring in the urban and rural areas from which the influent originates. Overall, amylases are proposed as indicators of mammal population while albumins are associated with livestock-related industrial activity.

4. As wastewater passes through treatment plants, proteins are removed. However, certain proteins - either due to their abundance or resistance to degradation - persist. These include human keratins and amylases, livestock albumin and some bacteria proteins (such as chaperonins).

5. Molecular profiles vary by site, reflecting population size and industrial activity. Larger populations show more human amylases in proteins, and higher levels of long-chain fatty acyls, organoheterocyclic compounds, and benzenoids in metabolites.

6. Proteomics complements the information from other omics regarding environmental surveillance. In some occasions it improves the knowledge due to the capacity to identify the species origin of the contamination.

7. Wastewater proteome could be used to enhance environmental surveillance techniques. Some applications are pest control, population size estimation, illegal discharges, and habit- and health-related biomarkers.

7. SUPPLEMENTARY MATERIAL

The supplementary material can be downloaded from the following link:

- https://www.mediafire.com/folder/kco7pt9w3v9t4/Sanchez_Jimenez_Thesis_2025
- https://saco.csic.es/s/R9aqAPQFZmm8GYH/download

The list of the files found in these links is:

- **4.1_Particulate_Identifications.xlsx:** identified proteins and peptides in the particulate fraction.
- **4.1_PD30_SearchParameters.docx:** parameters used for the search with Proteome Discoverer 3.0.
- **4.1_Soluble_Identifications.xlsx:** identified proteins and peptides in the soluble fraction.

- **4.2_Sol+Part+Probes_Identifications.xlsx:** Proteins identified in the soluble, particulate and in the previously analyzed probes.
- **4.2_Soluble_Identifications.xlsx:** Peptides and proteins identified in the soluble wastewater proteome.
- **4.2_WWTP_Physicochemical.xlsx:** WWTP and physicochemical characteristics of the wastewater samples.

- **4.3_Metabolome.xlsx:** quantification of the small molecules in the influent and the effluent.
- **4.3_Proteome.xlsx:** identified proteins in the influent, effluent and receiving waters.

- **4.4_Annotated_Compounds.xlsx:** complete list of the annotated compounds with their classification, their relative intensities and other metadata.
- **4.4_DifferentialAbundanceAnalysis.xlsx:** output of the differential abundance analysis of each platform.
- **4.4_Heatmaps.pdf:** hierarchical clustering heatmap for each of the platforms.
- **4.4_MS-Dial_Parameters.xlsx:** MS-DIAL parameters for each platform (GC-MS, HILIC-MS in positive and negative, RPLC-MS in positive and negative).
- **4.4_MS-Flo_Parameters.xlsx:** MS-FLO parameters for HILIC-MS and RPLC-MS.
- **4.4_MSI_Levels.docx:** MSI levels used for the curation of MS-DIAL annotations for RPLC-MS (both positive and negative), GC-MC and HILIC-MS (both positive and negative).
- **4.4_Supplementary_Tables.docx:** supplemental tables.

- **Wastewater_Protocol.pdf:** article “Shotgun proteomics to characterize wastewater proteins” (chapter Results 4.1).
- **Wastewater_Proteome.pdf:** article “Sewage Protein Information Mining: Discovery of Large Biomolecules as Biomarkers of Population and Industrial Activities” (chapter Results 4.2).

8. BIBLIOGRAPHY

Abedi, E., & Sahari, M. A. (2014). Long-chain polyunsaturated fatty acid sources and evaluation of their nutritional and functional properties. *Food Science & Nutrition*, *2*(5), 443–463. https://doi.org/10.1002/fsn3.121

Agan, M. L., Taylor, W. R., Young, I., Willis, W. A., New, G. D., Lair, H., Murphy, A., Marinelli, A., Juel, M. A. I., Munir, M., Dornburg, A., Schlueter, J., & Gibas, C. (2022). *Wastewater as a back door to serology?* Cold Spring Harbor Laboratory. http://dx.doi.org/10.1101/2022.11.11.22282224

Ahmad, F. (2022). Protein stability [determination] problems. *Frontiers in Molecular Biosciences*, *9*. https://doi.org/10.3389/fmolb.2022.880358

Ahmed, F., Tscharke, B., O'Brien, J., Thompson, J., Samanipour, S., Choi, P., Li, J., Mueller, J. F., & Thomas, K. (2020). Wastewater-based estimation of the prevalence of gout in Australia. *Science of The Total Environment*, *715*, 136925. https://doi.org/10.1016/j.scitotenv.2020.136925

Alygizakis, N., Markou, A. N., Rousis, N. I., Galani, A., Avgeris, M., Adamopoulos, P. G., Scorilas, A., Lianidou, E. S., Paraskevis, D., Tsiodras, S., Tsakris, A., Dimopoulos, M.-A., & Thomaidis, N. S. (2021). Analytical methodologies for the detection of SARS-CoV-2 in wastewater: Protocols and future perspectives. *TrAC Trends in Analytical Chemistry*, *134*, 116125. https://doi.org/10.1016/j.trac.2020.116125

Amin, V., Bowes, D. A., & Halden, R. U. (2023). Systematic scoping review evaluating the potential of wastewater-based epidemiology for monitoring cardiovascular disease and cancer. *Science of The Total Environment*, *858*, 160103. https://doi.org/10.1016/j.scitotenv.2022.160103

Ansede, M. (2018, September 19). Pioneering study finds more than 200,000 rats in Barcelona's sewers. *Ediciones EL PAÍS S.L.* https://english.elpais.com/elpais/2018/09/19/inenglish/1537368873_338066.html

Armenta-Castro, A., Núñez-Soto, M. T., Rodriguez-Aguillón, K. O., Aguayo-Acosta, A., Oyervides-Muñoz, M. A., Snyder, S. A., Barceló, D., Saththasivam, J., Lawler, J., Sosa-Hernández, J. E., & Parra-Saldívar, R. (2024). Urine biomarkers for Alzheimer's disease: A new opportunity for wastewater-based epidemiology? *Environment International*, *184*, 108462. https://doi.org/10.1016/j.envint.2024.108462

Auerbach, J. (2014). Does New York City Really have as Many Rats as People? *Significance*, *11*(4), 22–27. https://doi.org/10.1111/j.1740-9713.2014.00764.x

Baker, D. R., Barron, L., & Kasprzyk-Hordern, B. (2014). Illicit and pharmaceutical drug consumption estimated via wastewater analysis. Part A: Chemical analysis and drug use estimates. *Science of The Total Environment*, *487*, 629–641. https://doi.org/10.1016/j.scitotenv.2013.11.107

Barceló, D. (2020). Wastewater-Based Epidemiology to monitor COVID-19 outbreak:

Present and future diagnostic methods to be in your radar. *Case Studies in Chemical and Environmental Engineering*, *2*, 100042. https://doi.org/10.1016/j.cscee.2020.100042

Bedia, C. (2022). Metabolomics in environmental toxicology: Applications and challenges. *Trends in Environmental Analytical Chemistry*, *34*, e00161. https://doi.org/10.1016/j.teac.2022.e00161

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bispo, M. S., Veloso, M. C. C., Pinheiro, H. L. C., De Oliveira, R. F. S., Reis, J. O. N., & De Andrade, J. B. (2002). Simultaneous determination of caffeine, theobromine, and theophylline by high-performance liquid chromatography. *Journal of Chromatographic Science*, *40*(1), 45–48. https://doi.org/10.1093/chromsci/40.1.45

Boleda, M. R., Galceran, M. T., & Ventura, F. (2009). Monitoring of opiates, cannabinoids and their metabolites in wastewater, surface water and finished water in Catalonia, Spain. *Water Research*, *43*(4), 1126–1136. https://doi.org/10.1016/j.watres.2008.11.056

Bolstad, B. M., Irizarry, R. A., Åstrand, M., & Speed, T. P. (2003). A comparison of normalization methods for high densityoligonucleotide array data based on variance and bias. *Bioinformatics*, *19*(2), 185–193. https://doi.org/10.1093/bioinformatics/19.2.185

Bouatra, S., Aziat, F., Mandal, R., Guo, A. C., Wilson, M. R., Knox, C., Bjorndahl, T. C., Krishnamurthy, R., Saleem, F., Liu, P., Dame, Z. T., Poelzer, J., Huynh, J., Yallou, F. S., Psychogios, N., Dong, E., Bogumil, R., Roehring, C., & Wishart, D. S. (2013). The human urine metabolome. *PLoS ONE*, *8*(9), e73076. https://doi.org/10.1371/journal.pone.0073076

Bowes, D. A., Driver, E. M., Choi, P. M., Barcelo, D., & Beamer, P. I. (2024). Wastewater-based epidemiology to assess environmentally influenced disease. *Journal of Exposure Science & Environmental Epidemiology*, *34*(3), 387–388. https://doi.org/10.1038/s41370-024-00683-w

Bradbury, A., & Plückthun, A. (2015). Reproducibility: Standardize antibodies used in research. *Nature*, *518*(7537), 27–29. https://doi.org/10.1038/518027a

Braunberger, T. L., Nahhas, A. F., Katz, L. M., Sadrieh, N., & Lim, H. W. (2018). Dihydroxyacetone: A Review. *Journal of Drugs in Dermatology*, *17*(4), 387–391.

Bringans, S. D., Ito, J., Stoll, T., Winfield, K., Phillips, M., Peters, K., Davis, W. A., Davis, T. M. E., & Lipscombe, R. J. (2017). Comprehensive mass spectrometry based

biomarker discovery and validation platform as applied to diabetic kidney disease. *EuPA Open Proteomics*, *14*, 1–10. https://doi.org/10.1016/j.euprot.2016.12.001

Brodbelt, J. S. (2022). Deciphering combinatorial post-translational modifications by top-down mass spectrometry. *Current Opinion in Chemical Biology*, *70*, 102180. https://doi.org/10.1016/j.cbpa.2022.102180

Burgard, D. A., Banta-Green, C., & Field, J. A. (2013). Working upstream: How far can you go with sewage-based drug epidemiology? *Environmental Science & Technology*, *48*(3), 1362–1368. https://doi.org/10.1021/es4044648

Bustillo-Lecompte, C. F., & Mehrvar, M. (2015). Slaughterhouse wastewater characteristics, treatment, and management in the meat processing industry: A review on trends and advances. *Journal of Environmental Management*, *161*, 287–302. https://doi.org/10.1016/j.jenvman.2015.07.008

Butreddy, A., Janga, K. Y., Ajjarapu, S., Sarabu, S., & Dudhipala, N. (2021). Instability of therapeutic proteins — An overview of stresses, stabilization mechanisms and analytical techniques involved in lyophilized proteins. *International Journal of Biological Macromolecules*, *167*, 309–325. https://doi.org/10.1016/j.ijbiomac.2020.11.188

Byers, K. A., Lee, M. J., Patrick, D. M., & Himsworth, C. G. (2019). Rats about town: A systematic review of rat movement in urban ecosystems. *Frontiers in Ecology and Evolution*, *7*. https://doi.org/10.3389/fevo.2019.00013

Candiano, G., Santucci, L., Petretto, A., Bruschi, M., Dimuccio, V., Urbani, A., Bagnasco, S., & Ghiggeri, G. M. (2010). 2D-electrophoresis and the urine proteome map: Where do we stand? *Journal of Proteomics*, *73*(5), 829–844. https://doi.org/10.1016/j.jprot.2009.12.003

Carrascal, M., Abian, J., Ginebreda, A., & Barceló, D. (2020). Discovery of large molecules as new biomarkers in wastewater using environmental proteomics and suitable polymer probes. *Science of The Total Environment*, *747*, 141145. https://doi.org/10.1016/j.scitotenv.2020.141145

Carrascal, M., Sánchez-Jiménez, E., Fang, J., Pérez-López, C., Ginebreda, A., Barceló, D., & Abian, J. (2023). Sewage protein information mining: Discovery of large biomolecules as biomarkers of population and industrial activities. *Environmental Science & Technology*, *57*(30), 10929–10939. https://doi.org/10.1021/acs.est.3c00535

Casanovas, A., Carrascal, M., Abián, J., López-Tejero, M. D., & Llobera, M. (2009). Discovery of lipoprotein lipase pI isoforms and contributions to their characterization. *Journal of Proteomics*, *72*(6), 1031–1039. https://doi.org/10.1016/j.jprot.2009.06.002

Casanovas, A., Pinto-Llorente, R., Carrascal, M., & Abian, J. (2017). Large-Scale filter-aided

sample preparation method for the analysis of the ubiquitinome. *Analytical Chemistry*, *89*(7), 3840–3846. https://doi.org/10.1021/acs.analchem.6b04804

Casas, V., Rodríguez-Asiain, A., Pinto-Llorente, R., Vadillo, S., Carrascal, M., & Abian, J. (2017). Brachyspira hyodysenteriae and B. pilosicoli Proteins Recognized by Sera of Challenged Pigs. *Frontiers in Microbiology*, *8*. https://doi.org/10.3389/fmicb.2017.00723

Casas, V., Vadillo, S., San Juan, C., Carrascal, M., & Abian, J. (2016). The Exposed Proteomes of Brachyspira hyodysenteriae and B. pilosicoli. *Frontiers in Microbiology*, *7*. https://doi.org/10.3389/fmicb.2016.01103

Castiglioni, S., Senta, I., Borsotti, A., Davoli, E., & Zuccato, E. (2014). A novel approach for monitoring tobacco use in local communities by wastewater analysis. *Tobacco Control*, *24*(1), 38–42. https://doi.org/10.1136/tobaccocontrol-2014-051553

Choi, P. M., Tscharke, B. J., Donner, E., O'Brien, J. W., Grant, S. C., Kaserzon, S. L., Mackie, R., O'Malley, E., Crosbie, N. D., Thomas, K. V., & Mueller, J. F. (2018). Wastewater-based epidemiology biomarkers: Past, present and future. *TrAC Trends in Analytical Chemistry*, *105*, 453–469. https://doi.org/10.1016/j.trac.2018.06.004

Cilento, E. M., Jin, L., Stewart, T., Shi, M., Sheng, L., & Zhang, J. (2019). Mass spectrometry: A platform for biomarker discovery and validation for Alzheimer's and Parkinson's diseases. *Journal of Neurochemistry*, *151*(4), 397–416. https://doi.org/10.1111/jnc.14635

Claustrat, B., & Leston, J. (2015). Melatonin: Physiological effects in humans. *Neurochirurgie*, *61*(2–3), 77–84. https://doi.org/10.1016/j.neuchi.2015.03.002

Cui, X., Chen, C., Liu, Y., Zhou, D., & Liu, M. (2019). Exogenous refractory protein enhances biofilm formation by altering the quorum sensing system: A potential hazard of soluble microbial proteins from WWTP effluent. *Science of The Total Environment*, *667*, 384–389. https://doi.org/10.1016/j.scitotenv.2019.02.370

Cupp-Sutton, K. A., & Wu, S. (2020). High-throughput quantitative top-down proteomics. *Molecular Omics*, *16*(2), 91–99. https://doi.org/10.1039/c9mo00154a

da Silva, V. R., & Gregory, J. F., III. (2020). Vitamin B6. In *Present Knowledge in Nutrition* (pp. 225–237). Elsevier. https://doi.org/10.1016/b978-0-323-66162-1.00013-5

Dame, Z. T., Aziat, F., Mandal, R., Krishnamurthy, R., Bouatra, S., Borzouie, S., Guo, A. C., Sajed, T., Deng, L., Lin, H., Liu, P., Dong, E., & Wishart, D. S. (2015). The human saliva metabolome. *Metabolomics*, *11*(6), 1864–1883. https://doi.org/10.1007/s11306-015-0840-5

Daughton, C. G. (2001). Illicit drugs in municipal sewage. In *ACS Symposium Series* (pp. 348–364). American Chemical Society. https://doi.org/10.1021/bk-2001-0791.ch020

Daughton, C. G. (2012). Using biomarkers in sewage to monitor community-wide human

health: Isoprostanes as conceptual prototype. *Science of The Total Environment*, *424*, 16–38. https://doi.org/10.1016/j.scitotenv.2012.02.038

Daughton, C. G. (2018). Monitoring wastewater for assessing community health: Sewage Chemical-Information Mining (SCIM). *Science of The Total Environment*, *619–620*, 748–764. https://doi.org/10.1016/j.scitotenv.2017.11.102

Daughton, C. G., & Ternes, T. A. (1999). Pharmaceuticals and personal care products in the environment: Agents of subtle change? *Environmental Health Perspectives*, *107*, 907. https://doi.org/10.2307/3434573

DeFelice, B. C., Mehta, S. S., Samra, S., Čajka, T., Wancewicz, B., Fahrmann, J. F., & Fiehn, O. (2017). Mass spectral feature list optimizer (MS-FLO): A tool to minimize false positive peak reports in untargeted liquid chromatography–mass spectroscopy (LC-MS) data processing. *Analytical Chemistry*, *89*(6), 3250–3255. https://doi.org/10.1021/acs.analchem.6b04372

Devault, D. A., & Karolak, S. (2020). Wastewater-based epidemiology approach to assess population exposure to pesticides: A review of a pesticide pharmacokinetic dataset. *Environmental Science and Pollution Research*, *27*(5), 4695–4702. https://doi.org/10.1007/s11356-019-07521-9

Devianto, L. A., Amarasiri, M., Wang, L., Iizuka, T., & Sano, D. (2024). Identification of protein biomarkers in wastewater linked to the incidence of COVID-19. *Science of The Total Environment*, *951*, 175649. https://doi.org/10.1016/j.scitotenv.2024.175649

Devianto, L. A., & Sano, D. (2023). Systematic review and meta-analysis of human health-related protein markers for realizing real-time wastewater-based epidemiology. *Science of The Total Environment*, *897*, 165304. https://doi.org/10.1016/j.scitotenv.2023.165304

Directive - 91/271 - EN - EUR-Lex. (n.d.). https://eur-lex.europa.eu/eli/dir/1991/271/oj/eng

Directive - 2013/39 - EN - EUR-Lex. (n.d.). https://eur-lex.europa.eu/eli/dir/2013/39/oj/eng

Djoumbou Feunang, Y., Eisner, R., Knox, C., Chepelev, L., Hastings, J., Owen, G., Fahy, E., Steinbeck, C., Subramanian, S., Bolton, E., Greiner, R., & Wishart, D. S. (2016). ClassyFire: Automated chemical classification with a comprehensive, computable taxonomy. *Journal of Cheminformatics*, *8*(1). https://doi.org/10.1186/s13321-016-0174-y

Doellinger, J., Blumenscheit, C., Schneider, A., & Lasch, P. (2020). Isolation window optimization of data-independent acquisition using predicted libraries for deep and accurate proteome profiling. *Analytical Chemistry*, *92*(18), 12185–12192. https://doi.org/10.1021/acs.analchem.0c00994

Dong, F., Zhou, Y., Zeng, L., Watanabe, N., Su, X., & Yang, Z. (2017). Optimization of the Production of 1-Phenylethanol Using Enzymes from Flowers of Tea (Camellia

sinensis) Plants. *Molecules*, *22*(1), 131. https://doi.org/10.3390/molecules22010131

Dungan, K. M. (2008). 1,5-anhydroglucitol (GlycoMarkTM) as a marker of short-term glycemic control and glycemic excursions. *Expert Review of Molecular Diagnostics*, *8*(1), 9–19. https://doi.org/10.1586/14737159.8.1.9

Dutta, H., Kaushik, G., & Dutta, V. (2021). Wastewater-based epidemiology: A new frontier for tracking environmental persistence and community transmission of COVID-19. *Environmental Science and Pollution Research*, *29*(57), 85688–85699. https://doi.org/10.1007/s11356-021-17419-0

El-Metwally, R. I., El-Menawy, R. K., & Ismail, M. M. (2022). Correlation between free fatty acids content and textural properties of Gouda cheese supplemented with denatured whey protein paste. *Journal of Food Science and Technology*, *60*(2), 590–599. https://doi.org/10.1007/s13197-022-05643-6

Erban, A., Schauer, N., Fernie, A. R., & Kopka, J. (2007). Nonsupervised construction and application of mass spectral and retention time index libraries from time-of-flight gas chromatography-mass spectrometry metabolite profiles. In *Metabolomics* (pp. 19–38). Humana Press. http://dx.doi.org/10.1385/1-59745-244-0:19

Fasipe, O. J. (2019). The emergence of new antidepressants for clinical use: Agomelatine paradox versus other novel agents. *IBRO Reports*, *6*, 95–110. https://doi.org/10.1016/j.ibror.2019.01.001

Feng, A. Y. T., & Himsworth, C. G. (2013). The secret life of the city rat: A review of the ecology of urban Norway and black rats (Rattus norvegicus and Rattus rattus). *Urban Ecosystems*, *17*(1), 149–162. https://doi.org/10.1007/s11252-013-0305-4

Feng, S., Ji, H.-L., Wang, H., Zhang, B., Sterzenbach, R., Pan, C., & Guo, X. (2022). MetaLP: An integrative linear programming method for protein inference in metaproteomics. *PLOS Computational Biology*, *18*(10), e1010603. https://doi.org/10.1371/journal.pcbi.1010603

Fernández-Costa, C., Martínez-Bartolomé, S., McClatchy, D. B., Saviola, A. J., Yu, N.-K., & Yates, J. R., III. (2020). Impact of the identification strategy on the reproducibility of the DDA and DIA results. *Journal of Proteome Research*, *19*(8), 3153–3161. https://doi.org/10.1021/acs.jproteome.0c00153

Fioretti, J. M., Fumian, T. M., Rocha, M. S., dos Santos, I. de A. L., Carvalho-Costa, F. A., de Assis, M. R., Rodrigues, J. de S., Leite, J. P. G., & Miagostovich, M. P. (2017). Surveillance of noroviruses in Rio De Janeiro, Brazil: Occurrence of new GIV genotype in clinical and wastewater samples. *Food and Environmental Virology*, *10*(1), 1–6. https://doi.org/10.1007/s12560-017-9308-2

Foo, A. Y., & Rosalki, S. B. (1986). Measurement of plasma amylase activity. *Annals of Clinical Biochemistry: International Journal of Laboratory Medicine*, *23*(6), 624–637.

https://doi.org/10.1177/000456328602300602

Gao, J., O'Brien, J., Du, P., Li, X., Ort, C., Mueller, J. F., & Thai, P. K. (2016). Measuring selected PPCPs in wastewater to estimate the population in different cities in China. *Science of The Total Environment*, *568*, 164–170. https://doi.org/10.1016/j.scitotenv.2016.05.216

Gao, Y., & Yates, J. R., III. (2019). Protein analysis by shotgun proteomics. *Mass Spectrometry-Based Chemical Proteomics*, 1–38. https://doi.org/10.1002/9781118970195.ch1

Garimella, P. S., & Sarnak, M. J. (2016). Uromodulin in kidney health and disease. *Current Opinion in Nephrology and Hypertension*, 1. https://doi.org/10.1097/mnh.0000000000000299

Gonsior, M., Schmitt-Kopplin, P., Stavklint, H., Richardson, S. D., Hertkorn, N., & Bastviken, D. (2014). Changes in dissolved organic matter during the treatment processes of a drinking water plant in Sweden and formation of previously unknown disinfection byproducts. *Environmental Science & Technology*, *48*(21), 12714–12722. https://doi.org/10.1021/es504349p

González, S., López-Roldán, R., & Cortina, J.-L. (2012). Presence and biological effects of emerging contaminants in Llobregat River basin: A review. *Environmental Pollution*, *161*, 83–92. https://doi.org/10.1016/j.envpol.2011.10.002

González-Mariño, I., Rodil, R., Barrio, I., Cela, R., & Quintana, J. B. (2017). Wastewater-Based epidemiology as a new tool for estimating population exposure to phthalate plasticizers. *Environmental Science & Technology*, *51*(7), 3902–3910. https://doi.org/10.1021/acs.est.6b05612

Gracia-Lor, E., Rousis, N. I., Zuccato, E., Bade, R., Baz-Lomba, J. A., Castrignanò, E., Causanilles, A., Hernández, F., Kasprzyk-Hordern, B., Kinyua, J., McCall, A.-K., van Nuijs, A. L. N., Plósz, B. G., Ramin, P., Ryu, Y., Santos, M. M., Thomas, K., de Voogt, P., Yang, Z., & Castiglioni, S. (2017). Estimation of caffeine intake from analysis of caffeine metabolites in wastewater. *Science of The Total Environment*, *609*, 1582–1588. https://doi.org/10.1016/j.scitotenv.2017.07.258

Grilo, A. L., & Mantalaris, A. (2019). The increasingly human and profitable monoclonal antibody market. *Trends in Biotechnology*, *37*(1), 9–16. https://doi.org/10.1016/j.tibtech.2018.05.014

Gros, M., Petrović, M., Ginebreda, A., & Barceló, D. (2010). Removal of pharmaceuticals during wastewater treatment and environmental risk assessment using hazard indexes. *Environment International*, *36*(1), 15–26. https://doi.org/10.1016/j.envint.2009.09.002

Gros, M., Rodríguez-Mozaz, S., & Barceló, D. (2012). Fast and comprehensive multi-residue

analysis of a broad range of human and veterinary pharmaceuticals and some of their metabolites in surface and treated waters by ultra-high-performance liquid chromatography coupled to quadrupole-linear ion trap tandem mass spectrometry. *Journal of Chromatography A*, *1248*, 104–121. https://doi.org/10.1016/j.chroma.2012.05.084

Gros, M., Rodríguez-Mozaz, S., & Barceló, D. (2013). Rapid analysis of multiclass antibiotic residues and some of their metabolites in hospital, urban wastewater and river water by ultra-high-performance liquid chromatography coupled to quadrupole-linear ion trap tandem mass spectrometry. *Journal of Chromatography A*, *1292*, 173–188. https://doi.org/10.1016/j.chroma.2012.12.072

Guerrero-Latorre, L., Collado, N., Abasolo, N., Anzaldi, G., Bofill-Mas, S., Bosch, A., Bosch, L., Busquets, S., Caimari, A., Canela, N., Carcereny, A., Chacón, C., Ciruela, P., Corbella, I., Domingo, X., Escoté, X., Espiñeira, Y., Forés, E., Gandullo-Sarró, I., … Borrego, C. M. (2022). The Catalan Surveillance Network of SARS-CoV-2 in Sewage: Design, implementation, and performance. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-20957-3

Guerrero-Wyss, M., Durán Agüero, S., & Angarita Dávila, L. (2018). D-Tagatose is a promising sweetener to control glycaemia: A new functional food. *BioMed Research International*, *2018*, 1–7. https://doi.org/10.1155/2018/8718053

Haghgouei, H., & Alizadeh, N. (2025). Cellulose-based potentiometric sensor array for simultaneous determination of non-steroidal anti-inflammatory drugs in human serum and saliva samples. *International Journal of Biological Macromolecules*, 144026. https://doi.org/10.1016/j.ijbiomac.2025.144026

Hauso, Ø., Martinsen, T. C., & Waldum, H. (2015). 5-Aminosalicylic acid, a specific drug for ulcerative colitis. *Scandinavian Journal of Gastroenterology*, *50*(8), 933–941. https://doi.org/10.3109/00365521.2015.1018937

Hawkes, J. A., Dittmar, T., Patriarca, C., Tranvik, L., & Bergquist, J. (2016). Evaluation of the orbitrap mass spectrometer for the molecular fingerprinting analysis of natural dissolved organic matter. *Analytical Chemistry*, *88*(15), 7698–7704. https://doi.org/10.1021/acs.analchem.6b01624

Heinig, M., & Johnson, R. J. (2006). Role of uric acid in hypertension, renal disease, and metabolic syndrome. *Cleveland Clinic Journal of Medicine*, *73*(12), 1059–1064. https://doi.org/10.3949/ccjm.73.12.1059

Hertkorn, N., Harir, M., Koch, B. P., Michalke, B., & Schmitt-Kopplin, P. (2013). High-field NMR spectroscopy and FTICR mass spectrometry: Powerful discovery tools for the molecular level characterization of marine dissolved organic matter. *Biogeosciences*, *10*(3), 1583–1624. https://doi.org/10.5194/bg-10-1583-2013

Hignite, C., & Azarnoff, D. L. (1977). Drugs and drug metabolites as environmental contaminants: Chlorophenoxyisobutyrate and salicylic acid in sewage water effluent. *Life Sciences*, *20*(2), 337–341. https://doi.org/10.1016/0024-3205(77)90329-0

Himsworth, C. G., Jardine, C. M., Parsons, K. L., Feng, A. Y. T., & Patrick, D. M. (2014). The Characteristics of Wild Rat (Rattus spp.) Populations from an Inner-City Neighborhood with a Focus on Factors Critical to the Understanding of Rat-Associated Zoonoses. *PLoS ONE*, *9*(3), e91654. https://doi.org/10.1371/journal.pone.0091654

Hsu, S.-Y., Bayati, M., Li, C., Hsieh, H.-Y., Belenchia, A., Klutts, J., Zemmer, S. A., Reynolds, M., Semkiw, E., Johnson, H.-Y., Foley, T., Wieberg, C. G., Wenzel, J., Johnson, M. C., & Lin, C.-H. (2022). Biomarkers selection for population normalization in SARS-CoV-2 wastewater-based epidemiology. *Water Research*, *223*, 118985. https://doi.org/10.1016/j.watres.2022.118985

Hu, B., Li, H., Wang, Q., Tan, Y., Chen, R., Li, J., Ban, W., & Liang, L. (2018). Production and utilization of l-arabinose in china. *World Journal of Engineering and Technology*, *06*(03), 24–36. https://doi.org/10.4236/wjet.2018.63b004

Huang, T., Wang, J., Yu, W., & He, Z. (2012). Protein inference: A review. *Briefings in Bioinformatics*, *13*(5), 586–614. https://doi.org/10.1093/bib/bbs004

Hughes, C. S., Moggridge, S., Müller, T., Sorensen, P. H., Morin, G. B., & Krijgsveld, J. (2018). Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. *Nature Protocols*, *14*(1), 68–85. https://doi.org/10.1038/s41596-018-0082-x

Huizer, M., ter Laak, T. L., de Voogt, P., & van Wezel, A. P. (2021). Wastewater-based epidemiology for illicit drugs: A critical review on global data. *Water Research*, *207*, 117789. https://doi.org/10.1016/j.watres.2021.117789

Husøy, T., Andreassen, M., Hjertholm, H., Carlsen, M. H., Norberg, N., Sprong, C., Papadopoulou, E., Sakhi, A. K., Sabaredzovic, A., & Dirven, H. A. A. M. (2019). The Norwegian biomonitoring study from the EU project EuroMix: Levels of phenols and phthalates in 24-hour urine samples and exposure sources from food and personal care products. *Environment International*, *132*, 105103. https://doi.org/10.1016/j.envint.2019.105103

Inarmal, N., & Moodley, B. (2024). Removal efficiencies and environmental risk assessment of selected pharmaceuticals and metabolites at a wastewater treatment plant in Pietermaritzburg, South Africa. *Environmental Monitoring and Assessment*, *197*(1). https://doi.org/10.1007/s10661-024-13515-z

Jacob III, P., Yu, L., Duan, M., Ramos, L., Yturralde, O., & Benowitz, N. L. (2011). Determination of the nicotine metabolites cotinine and trans-3′-hydroxycotinine in

biologic fluids of smokers and non-smokers using liquid chromatography–tandem mass spectrometry: Biomarkers for tobacco smoke exposure and for phenotyping cytochrome P450 2A6 activity. *Journal of Chromatography B*, *879*(3–4), 267–276. https://doi.org/10.1016/j.jchromb.2010.12.012

Jakimska, A., Huerta, B., Bargańska, Ż., Kot-Wasik, A., Rodríguez-Mozaz, S., & Barceló, D. (2013). Development of a liquid chromatography–tandem mass spectrometry procedure for determination of endocrine disrupting compounds in fish from Mediterranean rivers. *Journal of Chromatography A*, *1306*, 44–58. https://doi.org/10.1016/j.chroma.2013.07.050

Jarnuczak, A. F., & Vizcaíno, J. A. (2017). Using the PRIDE database and proteomexchange for submitting and accessing public proteomics datasets. *Current Protocols in Bioinformatics*, *59*(1). https://doi.org/10.1002/cpbi.30

Jiang, Y., Rex, D. A. B., Schuster, D., Neely, B. A., Rosano, G. L., Volkmar, N., Momenzadeh, A., Peters-Clarke, T. M., Egbert, S. B., Kreimer, S., Doud, E. H., Crook, O. M., Yadav, A. K., Vanuopadath, M., Hegeman, A. D., Mayta, M. L., Duboff, A. G., Riley, N. M., Moritz, R. L., & Meyer, J. G. (2024). Comprehensive overview of bottom-up proteomics using mass spectrometry. *ACS Measurement Science Au*, *4*(4), 338–417. https://doi.org/10.1021/acsmeasuresciau.3c00068

Jurišić, A., Ćupina, A. I., Kavran, M., Potkonjak, A., Ivanović, I., Bjelić-Čabrilo, O., Meseldžija, M., Dudić, M., Poljaković-Pajnik, L., & Vasić, V. (2022). Surveillance strategies of rodents in agroecosystems, forestry and urban environments. *Sustainability*, *14*(15), 9233. https://doi.org/10.3390/su14159233

Kacso, G., & Terézhalmy, G. T. (1994). ACETYLSALICYLIC ACID AND ACETAMINOPHEN. *Dental Clinics of North America*, *38*(4), 633–644. https://doi.org/10.1016/s0011-8532(22)00181-1

Karantonis, H., Nomikos, T., & Demopoulos, C. (2009). Triacylglycerol metabolism. *Current Drug Targets*, *10*(4), 302–319. https://doi.org/10.2174/138945009787846443

Karu, N., Deng, L., Slae, M., Guo, A. C., Sajed, T., Huynh, H., Wine, E., & Wishart, D. S. (2018). A review on human fecal metabolomics: Methods, applications and the human fecal metabolome database. *Analytica Chimica Acta*, *1030*, 1–24. https://doi.org/10.1016/j.aca.2018.05.031

Kasprzyk-Hordern, B., Béen, F., Bijlsma, L., Brack, W., Castiglioni, S., Covaci, A., Martincigh, B. S., Mueller, J. F., van Nuijs, A. L. N., Oluseyi, T., & Thomas, K. (2022). Wastewater-Based epidemiology for the assessment of population exposure to chemicals: The need for integration with human biomonitoring for global one health actions. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4305813

Kimiyoshi, I., Yoshihiro, A., Kumi, N., Shinsei, M., Tatsuo, H., Osamu, S., Nobuyoshi, S., &

Takeshi, N. (1993). Cloning of the cDNA encoding human xanthine dehydrogenase (oxidase): Structural analysis of the protein and chromosomal location of the gene. *Gene*, *133*(2), 279–284. https://doi.org/10.1016/0378-1119(93)90652-j

Kleiner, M., Thorson, E., Sharp, C. E., Dong, X., Liu, D., Li, C., & Strous, M. (2017). Assessing species biomass contributions in microbial communities via metaproteomics. *Nature Communications*, *8*(1). https://doi.org/10.1038/s41467-017-01544-x

Kogawa, A. C., Pires, A. E. D. T., & Salgado, H. R. N. (2019). Atorvastatin: A review of analytical methods for pharmaceutical quality control and monitoring. *Journal of AOAC INTERNATIONAL*, *102*(3), 801–809. https://doi.org/10.5740/jaoacint.18-0200

Kolpin, D. W., Furlong, E. T., Meyer, M. T., Thurman, E. M., Zaugg, S. D., Barber, L. B., & Buxton, H. T. (2002). Pharmaceuticals, hormones, and other organic wastewater contaminants in U.S. streams, 1999−2000: A national reconnaissance. *Environmental Science & Technology*, *36*(6), 1202–1211. https://doi.org/10.1021/es011055j

Korecka, M., & Shaw, L. M. (2021). Mass spectrometry-based methods for robust measurement of Alzheimer's disease biomarkers in biological fluids. *Journal of Neurochemistry*, *159*(2), 211–233. https://doi.org/10.1111/jnc.15465

Korotkova, M., & Lundberg, I. E. (2014). The skeletal muscle arachidonic acid cascade in health and inflammatory disease. *Nature Reviews Rheumatology*, *10*(5), 295–303. https://doi.org/10.1038/nrrheum.2014.2

Kuhn, R., Benndorf, D., Rapp, E., Reichl, U., Palese, L. L., & Pollice, A. (2011). Metaproteome analysis of sewage sludge from membrane bioreactors. *PROTEOMICS*, *11*(13), 2738–2744. https://doi.org/10.1002/pmic.201000590

Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N., & Mann, M. (2014). Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nature Methods*, *11*(3), 319–324. https://doi.org/10.1038/nmeth.2834

Laber, G., & Schütze, E. (1977). Blood level studies in chickens, turkey poults and swine with tiamulin, a new antibiotic. *The Journal of Antibiotics*, *30*(12), 1119–1122. https://doi.org/10.7164/antibiotics.30.1119

Lanzillotti, M., & Brodbelt, J. S. (2023). Comparison of Top-Down Protein Fragmentation Induced by 213 and 193 nm UVPD. *Journal of the American Society for Mass Spectrometry*, *34*(2), 279–285. https://doi.org/10.1021/jasms.2c00288

Lara-Jacobo, L. R., Islam, G., Desaulniers, J.-P., Kirkwood, A. E., & Simmons, D. B. D. (2022). Detection of sars-cov-2 proteins in wastewater samples by mass spectrometry. *Environmental Science & Technology*, *56*(8), 5062–5070. https://doi.org/10.1021/acs.est.1c04705

Lavonen, E. E., Kothawala, D. N., Tranvik, L. J., Gonsior, M., Schmitt-Kopplin, P., & Köhler, S. J. (2015). Tracking changes in the optical properties and molecular composition of dissolved organic matter during drinking water production. *Water Research*, *85*, 286–294. https://doi.org/10.1016/j.watres.2015.08.024

Li, K., Guo, Z., & Bai, L. (2024). Digitoxose as powerful glycosyls for building multifarious glycoconjugates of natural products and un-natural products. *Synthetic and Systems Biotechnology*, *9*(4), 701–712. https://doi.org/10.1016/j.synbio.2024.05.012

Lin, Y., Reino, C., Carrera, J., Pérez, J., & van Loosdrecht, M. C. M. (2018). Glycosylated amyloid-like proteins in the structural extracellular polymers of aerobic granular sludge enriched with ammonium-oxidizing bacteria. *MicrobiologyOpen*, *7*(6). https://doi.org/10.1002/mbo3.616

Lindblom, E. U., & Samuelsson, O. (2022). Comparison of guideline- and model-based WWTP design for uncertain influent conditions. *Water Science and Technology*, *87*(1), 218–227. https://doi.org/10.2166/wst.2022.426

Liu, R., Li, Q., & Smith, L. M. (2014). Detection of large ions in time-of-flight mass spectrometry: Effects of ion mass and acceleration voltage on microchannel plate detector response. *Journal of the American Society for Mass Spectrometry*, *25*(8), 1374–1383. https://doi.org/10.1007/s13361-014-0903-2

Lopardo, L., Petrie, B., Proctor, K., Youdan, J., Barden, R., & Kasprzyk-Hordern, B. (2019). Estimation of community-wide exposure to bisphenol A via water fingerprinting. *Environment International*, *125*, 1–8. https://doi.org/10.1016/j.envint.2018.12.048

Mackuľak, T., Birošová, L., Grabic, R., Škubák, J., & Bodík, I. (2015). National monitoring of nicotine use in Czech and Slovak Republic based on wastewater analysis. *Environmental Science and Pollution Research*, *22*(18), 14000–14006. https://doi.org/10.1007/s11356-015-4648-7

Mackuľak, T., Gál, M., Špalková, V., Fehér, M., Briestenská, K., Mikušová, M., Tomčíková, K., Tamáš, M., & Butor Škulcová, A. (2021). Wastewater-Based epidemiology as an early warning system for the spreading of sars-cov-2 and its mutations in the population. *International Journal of Environmental Research and Public Health*, *18*(11), 5629. https://doi.org/10.3390/ijerph18115629

Maizel, A. C., & Remucal, C. K. (2017). The effect of advanced secondary municipal wastewater treatment on the molecular composition of dissolved organic matter. *Water Research*, *122*, 42–52. https://doi.org/10.1016/j.watres.2017.05.055

Mao, K., Zhang, H., Pan, Y., & Yang, Z. (2021). Biosensors for wastewater-based epidemiology for monitoring public health. *Water Research*, *191*, 116787. https://doi.org/10.1016/j.watres.2020.116787

Mariette, A., Kang, H. S., Heazlewood, J. L., Persson, S., Ebert, B., & Lampugnani, E. R.

(2021). Not just a simple sugar: Arabinose metabolism and function in plants. *Plant and Cell Physiology*, *62*(12), 1791–1812. https://doi.org/10.1093/pcp/pcab087

Marimuthu, A., O'Meally, Robert. N., Chaerkady, R., Subbannayya, Y., Nanjappa, V., Kumar, P., Kelkar, D. S., Pinto, S. M., Sharma, R., Renuse, S., Goel, R., Christopher, R., Delanghe, B., Cole, Robert. N., Harsha, H. C., & Pandey, A. (2011). A comprehensive map of the human urinary proteome. *Journal of Proteome Research*, *10*(6), 2734–2743. https://doi.org/10.1021/pr2003038

Mastroianni, N., López-García, E., Postigo, C., Barceló, D., & López de Alda, M. (2017). Five-year monitoring of 19 illicit and legal substances of abuse at the inlet of a wastewater treatment plant in Barcelona (NE Spain) and estimation of drug consumption patterns and trends. *Science of The Total Environment*, *609*, 916–926. https://doi.org/10.1016/j.scitotenv.2017.07.126

Mattes, W., Yang, X., Orr, M. S., Richter, P., & Mendrick, D. L. (2014). Biomarkers of tobacco smoke exposure. In *Advances in Clinical Chemistry* (pp. 1–45). Elsevier. https://doi.org/10.1016/bs.acc.2014.09.001

Matyash, V., Liebisch, G., Kurzchalia, T. V., Shevchenko, A., & Schwudke, D. (2008). Lipid extraction by methyl-tert-butyl ether for high-throughput lipidomics. *Journal of Lipid Research*, *49*(5), 1137–1146. https://doi.org/10.1194/jlr.d700041-jlr200

Mazumder, A., Dwivedi, A., & Du Plessis, J. (2016). Sinigrin and its therapeutic benefits. *Molecules*, *21*(4), 416. https://doi.org/10.3390/molecules21040416

Melby, J. A., Roberts, D. S., Larson, E. J., Brown, K. A., Bayne, E. F., Jin, S., & Ge, Y. (2021). Novel strategies to address the challenges in top-down proteomics. *Journal of the American Society for Mass Spectrometry*, *32*(6), 1278–1294. https://doi.org/10.1021/jasms.1c00099

Modick, H., Weiss, T., Dierkes, G., Koslitz, S., Käfferlein, H. U., Brüning, T., & Koch, H. M. (2015). Human metabolism and excretion kinetics of aniline after a single oral dose. *Archives of Toxicology*, *90*(6), 1325–1333. https://doi.org/10.1007/s00204-015-1566-x

Mohana, A. A., Roddick, F., Maniam, S., Gao, L., & Pramanik, B. K. (2023). Component analysis of fat, oil and grease in wastewater: Challenges and opportunities. *Analytical Methods*, *15*(39), 5112–5128. https://doi.org/10.1039/d3ay01222k

Murray, M. H., Fidino, M., Fyffe, R., Byers, K. A., Pettengill, J. B., Sondgeroth, K. S., Killion, H., Magle, S. B., Rios, M. J., Ortinau, N., & Santymire, R. M. (2020). City sanitation and socioeconomics predict rat zoonotic infection across diverse neighbourhoods. *Zoonoses and Public Health*, *67*(6), 673–683. https://doi.org/10.1111/zph.12748

Nakamura, M. T., Yudell, B. E., & Loor, J. J. (2014). Regulation of energy metabolism by long-chain fatty acids. *Progress in Lipid Research*, *53*, 124–144.

https://doi.org/10.1016/j.plipres.2013.12.001

Nesatyy, V. J., & Suter, M. J.-F. (2007). Proteomics for the analysis of environmental stress responses in organisms. *Environmental Science & Technology*, *41*(20), 6891–6900. https://doi.org/10.1021/es070561r

Nesvizhskii, A. I. (2010). A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of Proteomics*, *73*(11), 2092–2123. https://doi.org/10.1016/j.jprot.2010.08.009

Nguyen, H. T., Thai, P. K., Kaserzon, S. L., O'Brien, J. W., Eaglesham, G., & Mueller, J. F. (2018). Assessment of drugs and personal care products biomarkers in the influent and effluent of two wastewater treatment plants in Ho Chi Minh City, Vietnam. *Science of The Total Environment*, *631–632*, 469–475. https://doi.org/10.1016/j.scitotenv.2018.02.309

Noor, Z., Ahn, S. B., Baker, M. S., Ranganathan, S., & Mohamedali, A. (2020). Mass spectrometry–based protein identification in proteomics—a review. *Briefings in Bioinformatics*, *22*(2), 1620–1638. https://doi.org/10.1093/bib/bbz163

O'Brien, J. W., Thai, P. K., Brandsma, S. H., Leonards, P. E. G., Ort, C., & Mueller, J. F. (2015). Wastewater analysis of Census day samples to investigate per capita input of organophosphorus flame retardants and plasticizers into wastewater. *Chemosphere*, *138*, 328–334. https://doi.org/10.1016/j.chemosphere.2015.06.014

Ogura, T., & Shiraishi, C. (2025). Comparison of adverse events among angiotensin receptor blockers in hypertension using the United States Food and Drug Administration Adverse Event Reporting System. *Cureus*. https://doi.org/10.7759/cureus.81912

Oloye, F. F., Xie, Y., Challis, J. K., Femi-Oloye, O. P., Brinkmann, M., McPhedran, K. N., Jones, P. D., Servos, M. R., & Giesy, J. P. (2023). Understanding common population markers for SARS-CoV-2 RNA normalization in wastewater – A review. *Chemosphere*, *333*, 138682. https://doi.org/10.1016/j.chemosphere.2023.138682

O'Reilly, K., Wade, M., Farkas, K., Amman, F., Lison, A., Munday, J., Bingham, J., Mthombothi, Z., Fang, Z., Brown, C., Kao, R., & Danon, L. (2025). Analysis insights to support the use of wastewater and environmental surveillance data for infectious diseases and pandemic preparedness. *Epidemics*, *51*, 100825. https://doi.org/10.1016/j.epidem.2025.100825

Park, C., Helm, R. F., & Novak, J. T. (2008a). Investigating the fate of activated sludge extracellular proteins in sludge digestion using sodium dodecyl sulfate polyacrylamide gel electrophoresis. *Water Environment Research*, *80*(12), 2219–2227. https://doi.org/10.2175/106143008x325791

Park, C., Novak, J. T., Helm, R. F., Ahn, Y.-O., & Esen, A. (2008b). Evaluation of the

extracellular proteins in full-scale activated sludges. *Water Research*, *42*(14), 3879–3889. https://doi.org/10.1016/j.watres.2008.05.014

Park, Y., & Yetley, E. (1993). Intakes and food sources of fructose in the United States. *The American Journal of Clinical Nutrition*, *58*(5), 737S-747S. https://doi.org/10.1093/ajcn/58.5.737s

Peng, H., Wong, L., & Goh, W. W. B. (2023). ProInfer: An interpretable protein inference tool leveraging on biological networks. *PLOS Computational Biology*, *19*(3), e1010961. https://doi.org/10.1371/journal.pcbi.1010961

Perez-Lopez, C., Ginebreda, A., Carrascal, M., Barcelò, D., Abian, J., & Tauler, R. (2021). Non-target protein analysis of samples from wastewater treatment plants using the regions of interest-multivariate curve resolution (ROIMCR) chemometrics method. *Journal of Environmental Chemical Engineering*, *9*(4), 105752. https://doi.org/10.1016/j.jece.2021.105752

Peters-Clarke, T. M., Coon, J. J., & Riley, N. M. (2024). Instrumentation at the leading edge of proteomics. *Analytical Chemistry*, *96*(20), 7976–8010. https://doi.org/10.1021/acs.analchem.3c04497

Phungsai, P., Kurisu, F., Kasuga, I., & Furumai, H. (2016). Molecular characterization of low molecular weight dissolved organic matter in water reclamation processes using Orbitrap mass spectrometry. *Water Research*, *100*, 526–536. https://doi.org/10.1016/j.watres.2016.05.047

Picó, Y., & Barceló, D. (2021). Mass spectrometry in wastewater-based epidemiology for the determination of small and large molecules as biomarkers of exposure: Toward a global view of environment and human health under the COVID-19 outbreak. *ACS Omega*, *6*(46), 30865–30872. https://doi.org/10.1021/acsomega.1c04362

Picó, Y., Ginebreda, A., Carrascal, M., Abian, J., & Barceló, D. (2024). Simultaneous determination of small molecules and proteins in wastewater-based epidemiology. *Frontiers in Analytical Science*, *4*. https://doi.org/10.3389/frans.2024.1367448

Psychogios, N., Hau, D. D., Peng, J., Guo, A. C., Mandal, R., Bouatra, S., Sinelnikov, I., Krishnamurthy, R., Eisner, R., Gautam, B., Young, N., Xia, J., Knox, C., Dong, E., Huang, P., Hollander, Z., Pedersen, T. L., Smith, S. R., Bamforth, F., … Wishart, D. S. (2011). The human serum metabolome. *PLoS ONE*, *6*(2), e16957. https://doi.org/10.1371/journal.pone.0016957

Questa, M., Weimer, B. C., Fiehn, O., Chow, B., Hill, S. L., Ackermann, M. R., Lidbury, J. A., Steiner, J. M., Suchodolski, J. S., & Marsilio, S. (2024). Unbiased serum metabolomic analysis in cats with naturally occurring chronic enteropathies before and after medical intervention. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-57004-2

R: The R Project for Statistical Computing. (n.d.). Retrieved March 12, 2025, from https://www.R-project.org/

Rice, J., & Kasprzyk-Hordern, B. (2019). A new paradigm in public health assessment: Water fingerprinting for protein markers of public health using mass spectrometry. *TrAC Trends in Analytical Chemistry*, *119*, 115621. https://doi.org/10.1016/j.trac.2019.115621

Richard, D. M., Dawes, M. A., Mathias, C. W., Acheson, A., Hill-Kapturczak, N., & Dougherty, D. M. (2009). L-Tryptophan: Basic metabolic functions, behavioral research and therapeutic indications. *International Journal of Tryptophan Research*, *2*. https://doi.org/10.4137/ijtr.s2129

Rico, M., Andrés-Costa, M. J., & Picó, Y. (2017). Estimating population size in wastewater-based epidemiology. Valencia metropolitan area as a case study. *Journal of Hazardous Materials*, *323*, 156–165. https://doi.org/10.1016/j.jhazmat.2016.05.079

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47–e47. https://doi.org/10.1093/nar/gkv007

Robins, K., Leonard, A. F. C., Farkas, K., Graham, D. W., Jones, D. L., Kasprzyk-Hordern, B., Bunce, J. T., Grimsley, J. M. S., Wade, M. J., Zealand, A. M., & McIntyre-Nolan, S. (2023). Research needs for optimising wastewater-based epidemiology monitoring for public health protection. In *Wastewater-based Epidemiology at the Frontier of Global Public Health*. IWA Publishing. https://doi.org/10.2166/9781789064780_ch11

Rosati, D., Palmieri, M., Brunelli, G., Morrione, A., Iannelli, F., Frullanti, E., & Giordano, A. (2024). Differential gene expression analysis pipelines and bioinformatic tools for the identification of specific biomarkers: A review. *Computational and Structural Biotechnology Journal*, *23*, 1154–1168. https://doi.org/10.1016/j.csbj.2024.02.018

Rousis, N. I., Gracia-Lor, E., Reid, M. J., Baz-Lomba, J. A., Ryu, Y., Zuccato, E., Thomas, K. V., & Castiglioni, S. (2020). Assessment of human exposure to selected pesticides in Norway by wastewater analysis. *Science of The Total Environment*, *723*, 138132. https://doi.org/10.1016/j.scitotenv.2020.138132

Rousis, N. I., Gracia-Lor, E., Zuccato, E., Bade, R., Baz-Lomba, J. A., Castrignanò, E., Causanilles, A., Covaci, A., de Voogt, P., Hernàndez, F., Kasprzyk-Hordern, B., Kinyua, J., McCall, A.-K., Plósz, B. Gy., Ramin, P., Ryu, Y., Thomas, K. V., van Nuijs, A., Yang, Z., & Castiglioni, S. (2017). Wastewater-based epidemiology to assess pan-European pesticide exposure. *Water Research*, *121*, 270–279. https://doi.org/10.1016/j.watres.2017.05.044

Ryu, Y., Barceló, D., Barron, L. P., Bijlsma, L., Castiglioni, S., de Voogt, P., Emke, E., Hernández, F., Lai, F. Y., Lopes, A., de Alda, M. L., Mastroianni, N., Munro, K.,

O'Brien, J., Ort, C., Plósz, B. G., Reid, M. J., Yargeau, V., & Thomas, K. V. (2016a). Comparative measurement and quantitative risk assessment of alcohol consumption through wastewater-based epidemiology: An international study in 20 cities. *Science of The Total Environment*, *565*, 977–983. https://doi.org/10.1016/j.scitotenv.2016.04.138

Ryu, Y., Gracia-Lor, E., Bade, R., Baz-Lomba, J. A., Bramness, J. G., Castiglioni, S., Castrignanò, E., Causanilles, A., Covaci, A., de Voogt, P., Hernandez, F., Kasprzyk-Hordern, B., Kinyua, J., McCall, A.-K., Ort, C., Plósz, B. G., Ramin, P., Rousis, N. I., Reid, M. J., & Thomas, K. V. (2016b). Increased levels of the oxidative stress biomarker 8-iso-prostaglandin F2α in wastewater associated with tobacco use. *Scientific Reports*, *6*(1). https://doi.org/10.1038/srep39055

Sánchez-Jiménez, E., Abian, J., Ginebreda, A., Barceló, D., & Carrascal, M. (2023). Shotgun proteomics to characterize wastewater proteins. *MethodsX*, *11*, 102403. https://doi.org/10.1016/j.mex.2023.102403

Sanchís, J., Gernjak, W., Munné, A., Catalán, N., Petrovic, M., & Farré, M. J. (2021). Fate of N-nitrosodimethylamine and its precursors during a wastewater reuse trial in the Llobregat River (Spain). *Journal of Hazardous Materials*, *407*, 124346. https://doi.org/10.1016/j.jhazmat.2020.124346

Santos, J. L., Aparicio, I., Callejón, M., & Alonso, E. (2009). Occurrence of pharmaceutically active compounds during 1-year period in wastewaters from four wastewater treatment plants in Seville (Spain). *Journal of Hazardous Materials*, *164*(2–3), 1509–1516. https://doi.org/10.1016/j.jhazmat.2008.09.073

Savin, M., Bierbaum, G., Mutters, N. T., Schmithausen, R. M., Kreyenschmidt, J., García-Meniño, I., Schmoger, S., Käsbohrer, A., & Hammerl, J. A. (2022). Genetic Characterization of Carbapenem-Resistant Klebsiella spp. from Municipal and Slaughterhouse Wastewater. *Antibiotics*, *11*(4), 435. https://doi.org/10.3390/antibiotics11040435

Schauer, N., Steinhauser, D., Strelkov, S., Schomburg, D., Allison, G., Moritz, T., Lundgren, K., Roessner-Tunali, U., Forbes, M. G., Willmitzer, L., Fernie, A. R., & Kopka, J. (2005). GC–MS libraries for the rapid identification of metabolites in complex biological samples. *FEBS Letters*, *579*(6), 1332–1337. https://doi.org/10.1016/j.febslet.2005.01.029

Schmid-Wendtner, M.-H., & Korting, H. C. (2004). Penciclovir Cream – Improved Topical Treatment for Herpes simplex Infections. *Skin Pharmacology and Physiology*, *17*(5), 214–218. https://doi.org/10.1159/000080214

Senta, I., Rodríguez-Mozaz, S., Corominas, L., & Petrovic, M. (2020). Wastewater-based epidemiology to assess human exposure to personal care and household products –

A review of biomarkers, analytical methods, and applications. *Trends in Environmental Analytical Chemistry*, *28*, e00103. https://doi.org/10.1016/j.teac.2020.e00103

Shah, K., & Maghsoudlou, P. (2016). Enzyme-linked immunosorbent assay (ELISA): The basics. *British Journal of Hospital Medicine*, *77*(7), C98–C101. https://doi.org/10.12968/hmed.2016.77.7.c98

Sheng, L.-H., Chen, H.-R., Huo, Y.-B., Wang, J., Zhang, Y., Yang, M., & Zhang, H.-X. (2014). Simultaneous determination of 24 antidepressant drugs and their metabolites in wastewater by ultra-high performance liquid chromatography–tandem mass spectrometry. *Molecules*, *19*(1), 1212–1222. https://doi.org/10.3390/molecules19011212

Sherman, B. T., Hao, M., Qiu, J., Jiao, X., Baseler, M. W., Lane, H. C., Imamichi, T., & Chang, W. (2022). DAVID: A web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Research*, *50*(W1), W216–W221. https://doi.org/10.1093/nar/gkac194

Shon, H. K., Vigneswaran, S., & Snyder, S. A. (2006). Effluent organic matter (efom) in wastewater: Constituents, effects, and treatment. *Critical Reviews in Environmental Science and Technology*, *36*(4), 327–374. https://doi.org/10.1080/10643380600580011

Siddiqui, H., Sami, F., & Hayat, S. (2020). Glucose: Sweet or bitter effects in plants-a review on current and future perspective. *Carbohydrate Research*, *487*, 107884. https://doi.org/10.1016/j.carres.2019.107884

Sims, N., & Kasprzyk-Hordern, B. (2020). Future perspectives of wastewater-based epidemiology: Monitoring infectious disease spread and resistance to the community level. *Environment International*, *139*, 105689. https://doi.org/10.1016/j.envint.2020.105689

Sinha, A., & Mann, M. (2020). A beginner's guide to mass spectrometry–based proteomics. *The Biochemist*, *42*(5), 64–69. https://doi.org/10.1042/bio20200057

Skees, A. J., Foppe, K. S., Loganathan, B., & Subedi, B. (2018). Contamination profiles, mass loadings, and sewage epidemiology of neuropsychiatric and illicit drugs in wastewater and river waters from a community in the Midwestern United States. *Science of The Total Environment*, *631–632*, 1457–1464. https://doi.org/10.1016/j.scitotenv.2018.03.060

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, *3*(1), 1–25. https://doi.org/10.2202/1544-6115.1027

Soga, T. (2007). Capillary electrophoresis-mass spectrometry for metabolomics. In

Metabolomics (pp. 129–138). Humana Press. http://dx.doi.org/10.1385/1-59745-244-0:129

Subedi, B., Balakrishna, K., Joshua, D. I., & Kannan, K. (2017). Mass loading and removal of pharmaceuticals and personal care products including psychoactives, antihypertensives, and antibiotics in two sewage treatment plants in southern India. *Chemosphere*, *167*, 429–437. https://doi.org/10.1016/j.chemosphere.2016.10.026

Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., Fan, T. W.-M., Fiehn, O., Goodacre, R., Griffin, J. L., Hankemeier, T., Hardy, N., Harnly, J., Higashi, R., Kopka, J., Lane, A. N., Lindon, J. C., Marriott, P., Nicholls, A. W., … Viant, M. R. (2007). Proposed minimum reporting standards for chemical analysis. *Metabolomics*, *3*(3), 211–221. https://doi.org/10.1007/s11306-007-0082-2

Suwała, J., Machowska, M., & Wiela-Hojeńska, A. (2019). Venlafaxine pharmacogenetics: a Comprehensive review. *Pharmacogenomics*, *20*(11), 829–845. https://doi.org/10.2217/pgs-2019-0031

Sweat metabolome. (n.d.). Retrieved July 6, 2025, from https://sweatmetabolome.ca/

Tai, Y., Zhang, Z., Liu, Z., Li, X., Yang, Z., Wang, Z., An, L., Ma, Q., & Su, Y. (2024). D-ribose metabolic disorder and diabetes mellitus. *Molecular Biology Reports*, *51*(1). https://doi.org/10.1007/s11033-023-09076-y

Tang, J., Li, Y., Zhang, L., Mu, J., Jiang, Y., Fu, H., Zhang, Y., Cui, H., Yu, X., & Ye, Z. (2023). Biosynthetic pathways and functions of indole-3-acetic acid in microorganisms. *Microorganisms*, *11*(8), 2077. https://doi.org/10.3390/microorganisms11082077

Tao, X., Liu, L., Ma, P., Hu, J., Ming, Z., Dang, K., Zhang, Y., & Li, Y. (2023). Association of Circulating Very Long-Chain Saturated fatty acids with cardiovascular mortality in NHANES 2003-2004, 2011-2012. *The Journal of Clinical Endocrinology & Metabolism*, *109*(2), e633–e645. https://doi.org/10.1210/clinem/dgad561

Thanekar, P., Gogate, P. R., Znak, Z., Sukhatskiy, Yu., & Mnykh, R. (2021). Degradation of benzene present in wastewater using hydrodynamic cavitation in combination with air. *Ultrasonics Sonochemistry*, *70*, 105296. https://doi.org/10.1016/j.ultsonch.2020.105296

Thilakarathna, P. T. A., Fareed, F., Athukorala, S. N. P., Jinadasa, R., Premachandra, T., Noordeen, F., Gamage, C. D., Makehelwala, M., Weragoda, S. K., Fernando, B. R., Zhang, Y., Wei, Y., Yang, M., & Karunaratne, S. H. P. P. (2025). Spatio-temporal variation of microbial indicators of river water and treatment efficiencies of drinking water treatment plants along the upper Mahaweli river segment of Sri Lanka. *Environmental Pollution*, *367*, 125628. https://doi.org/10.1016/j.envpol.2025.125628

Thomas, K. V., Bijlsma, L., Castiglioni, S., Covaci, A., Emke, E., Grabic, R., Hernández, F.,

Karolak, S., Kasprzyk-Hordern, B., Lindberg, R. H., Lopez de Alda, M., Meierjohann, A., Ort, C., Pico, Y., Quintana, J. B., Reid, M., Rieckermann, J., Terzic, S., van Nuijs, A. L. N., & de Voogt, P. (2012). Comparing illicit drug use in 19 European cities through sewage analysis. *Science of The Total Environment*, *432*, 432–439. https://doi.org/10.1016/j.scitotenv.2012.06.069

Tinschert, A., Kiener, A., Heinzmann, K., & Tschech, A. (1997). Isolation of new 6-methylnicotinic-acid-degrading bacteria, one of which catalyses the regioselective hydroxylation of nicotinic acid at position C2. *Archives of Microbiology*, *168*(5), 355–361. https://doi.org/10.1007/s002030050509

Tolstikov, V. V., Fiehn, O., & Tanaka, N. (2007). Application of liquid chromatography-mass spectrometry analysis in metabolomics: Reversed-Phase monolithic capillary chromatography and hydrophilic chromatography coupled to electrospray ionization-mass spectrometry. In *Metabolomics* (pp. 141–156). Humana Press. http://dx.doi.org/10.1385/1-59745-244-0:141

Tomczyk, M., Heileson, J. L., Babiarz, M., & Calder, P. C. (2023). Athletes can benefit from increased intake of EPA and dha—evaluating the evidence. *Nutrients*, *15*(23), 4925. https://doi.org/10.3390/nu15234925

Tripathi, G., Kumar, A., Rajkhowa, S., & Tiwari, V. K. (2021). Synthesis of biologically relevant heterocyclic skeletons under solvent-free condition. In *Green Synthetic Approaches for Biologically Relevant Heterocycles* (pp. 421–459). Elsevier. http://dx.doi.org/10.1016/b978-0-12-820586-0.00013-3

Tscharke, B. J., Chen, C., Gerber, J. P., & White, J. M. (2016). Temporal trends in drug use in Adelaide, South Australia by wastewater analysis. *Science of The Total Environment*, *565*, 384–391. https://doi.org/10.1016/j.scitotenv.2016.04.183

Tsugawa, H., Cajka, T., Kind, T., Ma, Y., Higgins, B., Ikeda, K., Kanazawa, M., VanderGheynst, J., Fiehn, O., & Arita, M. (2015). MS-DIAL: Data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nature Methods*, *12*(6), 523–526. https://doi.org/10.1038/nmeth.3393

Tugui, C. G., Cordesius, F., van Holthe, W., van Loosdrecht, M. C. M., & Pabst, M. (2025). *Wastewater metaproteomics: Tracking microbial and human protein biomarkers*. Cold Spring Harbor Laboratory. https://doi.org/10.1101/2025.02.08.637285

Uszkoreit, J., Marcus, K., & Eisenacher, M. (2024). A review of protein inference. In *Methods in Molecular Biology* (pp. 53–64). Springer US. https://doi.org/10.1007/978-1-0716-4152-1_4

van Lipzig, M. M. H., Commandeur, J. N., de Kanter, F. J. J., Damsten, M. C., Vermeulen, N. P. E., Maat, E., Groot, E. J., Brouwer, A., Kester, M. H. A., Visser, T. J., & Meerman, J. H. N. (2005). Bioactivation of dibrominated biphenyls by cytochrome P450 activity

to metabolites with estrogenic activity and estrogen sulfotransferase inhibition capacity. *Chemical Research in Toxicology*, *18*(11), 1691–1700. https://doi.org/10.1021/tx0501233

Vaniya, A., Karlstaedt, A., Gulkok, D., Thottakara, T., Liu, Y., Fan, S., Eades, H., Vakrou, S., Fukunaga, R., Vernon, H. J., Fiehn, O., & Abraham, M. R. (2024). Allele-specific dysregulation of lipid and energy metabolism in early-stage hypertrophic cardiomyopathy. *Journal of Molecular and Cellular Cardiology Plus*, *8*, 100073. https://doi.org/10.1016/j.jmccpl.2024.100073

Varney, J., Barrett, J., Scarlata, K., Catsos, P., Gibson, P. R., & Muir, J. G. (2017). FODMAPs: Food composition, defining cutoff values and international application. *Journal of Gastroenterology and Hepatology*, *32*(S1), 53–61. https://doi.org/10.1111/jgh.13698

Wang, M., & Feng, L. (2023). A carbon based-screen-printed electrode amplified with two-dimensional reduced graphene/Fe3O4 nanocomposite as electroanalytical sensor for monitoring 4-aminophenol in environmental fluids. *Chemosphere*, *323*, 138238. https://doi.org/10.1016/j.chemosphere.2023.138238

Wang, M., Peng, Y., Yan, H., Pan, Z., Du, R., & Liu, G. (2025). Bioequivalence and safety of two amisulpride formulations in healthy Chinese subjects under fasting and fed conditions: A Randomized, open-label, single-dose, crossover study. *Drugs in R&D*. https://doi.org/10.1007/s40268-025-00508-7

Wang, S., Song, R., Wang, Z., Jing, Z., Wang, S., & Ma, J. (2018). S100A8/A9 in inflammation. *Frontiers in Immunology*, *9*. https://doi.org/10.3389/fimmu.2018.01298

Westgate, P. J., & Park, C. (2010). Evaluation of proteins and organic nitrogen in wastewater treatment effluents. *Environmental Science & Technology*, *44*(14), 5352–5357. https://doi.org/10.1021/es100244s

Wilmes, P., & Bond, P. L. (2006). Metaproteomics: Studying functional gene expression in microbial ecosystems. *Trends in Microbiology*, *14*(2), 92–97. https://doi.org/10.1016/j.tim.2005.12.006

Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Vázquez-Fresno, R., Sajed, T., Johnson, D., Li, C., Karu, N., Sayeeda, Z., Lo, E., Assempour, N., Berjanskii, M., Singhal, S., Arndt, D., Liang, Y., Badran, H., Grant, J., … Scalbert, A. (2017). HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Research*, *46*(D1), D608–D617. https://doi.org/10.1093/nar/gkx1089

Wishart, D. S., Guo, A., Oler, E., Wang, F., Anjum, A., Peters, H., Dizon, R., Sayeeda, Z., Tian, S., Lee, B. L., Berjanskii, M., Mah, R., Yamamoto, M., Jovel, J., Torres-Calzada, C., Hiebert-Giesbrecht, M., Lui, V. W., Varshavi, D., Varshavi, D., … Gautam, V. (2021). HMDB 5.0: The human metabolome database for 2022. *Nucleic*

Acids Research, *50*(D1), D622–D631. https://doi.org/10.1093/nar/gkab1062

Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., Djoumbou, Y., Mandal, R., Aziat, F., Dong, E., Bouatra, S., Sinelnikov, I., Arndt, D., Xia, J., Liu, P., Yallou, F., Bjorndahl, T., Perez-Pineiro, R., Eisner, R., … Scalbert, A. (2012). HMDB 3.0—the Human metabolome database in 2013. *Nucleic Acids Research*, *41*(D1), D801–D807. https://doi.org/10.1093/nar/gks1065

Wishart, D. S., Knox, C., Guo, A. C., Eisner, R., Young, N., Gautam, B., Hau, D. D., Psychogios, N., Dong, E., Bouatra, S., Mandal, R., Sinelnikov, I., Xia, J., Jia, L., Cruz, J. A., Lim, E., Sobsey, C. A., Shrivastava, S., Huang, P., … Forsythe, I. (2009). HMDB: A knowledgebase for the human metabolome. *Nucleic Acids Research*, *37*(Database), D603–D610. https://doi.org/10.1093/nar/gkn810

Wishart, D. S., Lewis, M. J., Morrissey, J. A., Flegel, M. D., Jeroncic, K., Xiong, Y., Cheng, D., Eisner, R., Gautam, B., Tzur, D., Sawhney, S., Bamforth, F., Greiner, R., & Li, L. (2008). The human cerebrospinal fluid metabolome. *Journal of Chromatography B*, *871*(2), 164–173. https://doi.org/10.1016/j.jchromb.2008.05.001

Wishart, D. S., Tzur, D., Knox, C., Eisner, R., Guo, A. C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., Fung, C., Nikolai, L., Lewis, M., Coutouly, M.-A., Forsythe, I., Tang, P., Shrivastava, S., Jeroncic, K., Stothard, P., … Querengesser, L. (2007). HMDB: The human metabolome database. *Nucleic Acids Research*, *35*(Database), D521–D526. https://doi.org/10.1093/nar/gkl923

Wohlgemuth, G., Haldiya, P. K., Willighagen, E., Kind, T., & Fiehn, O. (2010). The Chemical Translation Service—a web-based tool to improve standardization of metabolomic reports. *Bioinformatics*, *26*(20), 2647–2648. https://doi.org/10.1093/bioinformatics/btq476

Yeboah, G. K., Lobanova, E. S., Brush, R. S., & Agbaga, M.-P. (2021). Very long chain fatty acid-containing lipids: A decade of novel insights from the study of ELOVL4. *Journal of Lipid Research*, *62*, 100030. https://doi.org/10.1016/j.jlr.2021.100030

Yu, Q., Wang, D., Dong, P., & Zheng, L. (2025). Probiotics combined with trimebutine for the treatment of irritable bowel syndrome patients: A systematic review and meta-analysis. *Journal of Gastroenterology and Hepatology*, *40*(3), 677–691. https://doi.org/10.1111/jgh.16858

Yubero-Serrano, E. M., Lopez-Moreno, J., Gomez-Delgado, F., & Lopez-Miranda, J. (2018). Extra virgin olive oil: More than a healthy fat. *European Journal of Clinical Nutrition*, *72*(S1), 8–17. https://doi.org/10.1038/s41430-018-0304-x

Zahedi, A., Monis, P., Deere, D., & Ryan, U. (2021). Wastewater-based epidemiology—surveillance and early detection of waterborne pathogens with a focus on SARS-CoV-2, Cryptosporidium and Giardia. *Parasitology Research*, *120*(12), 4167–4188.

https://doi.org/10.1007/s00436-020-07023-5

Zhang, P., Shen, Y., Guo, J.-S., Li, C., Wang, H., Chen, Y.-P., Yan, P., Yang, J.-X., & Fang, F. (2015). Extracellular protein analysis of activated sludge and their functions in wastewater treatment plant by shotgun proteomics. *Scientific Reports*, *5*(1). https://doi.org/10.1038/srep12041

Zhang, P., Zhu, J., Xu, X.-Y., Qing, T.-P., Dai, Y.-Z., & Feng, B. (2019). Identification and function of extracellular protein in wastewater treatment using proteomic approaches: A minireview. *Journal of Environmental Management*, *233*, 24–29. https://doi.org/10.1016/j.jenvman.2018.12.028

Zhang, X., Fang, A., Riley, C. P., Wang, M., Regnier, F. E., & Buck, C. (2010). Multi-dimensional liquid chromatography in proteomics—A review. *Analytica Chimica Acta*, *664*(2), 101–113. https://doi.org/10.1016/j.aca.2010.02.001

Zhao, J., Lu, J., Zhao, H., Yan, Y., & Dong, H. (2023). In five wastewater treatment plants in Xinjiang, China: Removal processes for illicit drugs, their occurrence in receiving river waters, and ecological risk assessment. *Chemosphere*, *339*, 139668. https://doi.org/10.1016/j.chemosphere.2023.139668

Zheng, W., Chai, J., Wu, J., Zhang, J., & Qi, H. (2025). Ultrasensitive and real-time detection of kanamycin residues in milk using an aptasensor based on microfluidic capacitive strategy. *Biosensors*, *15*(5), 322. https://doi.org/10.3390/bios15050322

Zhu, F., Wu, X., Zhao, L., Liu, X., Qi, J., Wang, X., & Wang, J. (2017). Lipid profiling in sewage sludge. *Water Research*, *116*, 149–158. https://doi.org/10.1016/j.watres.2017.03.032

Zou, W., Liu, L., & Chen, J. (2012). Structure, mechanism and regulation of an artificial microbial ecosystem for vitamin C production. *Critical Reviews in Microbiology*, *39*(3), 247–255. https://doi.org/10.3109/1040841x.2012.706250

Zuccato, E., Chiabrando, C., Castiglioni, S., Bagnati, R., & Fanelli, R. (2008). Estimating community drug abuse by wastewater analysis. *Environmental Health Perspectives*, *116*(8), 1027–1032. https://doi.org/10.1289/ehp.11022