



UNIVERSITY OF THE PELOPONNESE & NCSR "DEMOKRITOS"  
MSC PROGRAMME IN DATA SCIENCE

# A Deep Learning approach for modeling the spatial distribution of contaminants in the Black

## Sea

by

Nikiforos Alygizakis

A thesis submitted in partial fulfillment  
of the requirements for the MSc  
in Data Science

**Supervisor:** Theodoros Giannakopoulos  
Principal Researcher

**Co-supervisors:** Anastasia Krithara, Charilaos Akasiadis  
Research associate, Research associate

Athens, May 2022

A Deep Learning approach for modeling the spatial distribution of contaminants in  
the Black Sea

Nikiforos Alygizakis

MSc. Thesis, MSc. Programme in Data Science

University of the Peloponnese & NCSR “Demokritos”, May 2022

Copyright © 2022 Nikiforos Alygizakis. All Rights Reserved.



UNIVERSITY OF THE PELOPONNESE & NCSR "DEMOKRITOS"  
MSC PROGRAMME IN DATA SCIENCE

# A Deep Learning approach for modeling the spatial distribution of contaminants in the Black

## Sea

by

Nikiforos Alygizakis

A thesis submitted in partial fulfillment  
of the requirements for the MSc  
in Data Science

**Supervisor:** Theodoros Giannakopoulos  
Principal Researcher

**Co-supervisors:** Anastasia Krithara, Charilaos Akasiadis  
Research associate, Research associate

Approved by the examination committee on May, 2022.

(Signature)

(Signature)

(Signature)

.....  
Theodoros Giannakopoulos  
Principal Researcher

.....  
Anastasia Krithara  
Research associate

.....  
Charilaos Akasiadis  
Research associate

Athens, May 2022





## Declaration of Authorship

- (1) I declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where states otherwise by reference or acknowledgment, the work presented is entirely my own.
- (2) I confirm that this thesis presented for the degree of Bachelor of Science in Informatics and Telecommunications, has
  - (i) been composed entirely by myself
  - (ii) been solely the result of my own work
  - (iii) not been submitted for any other degree or professional qualification
- (3) I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(Signature)

.....  
Nikiforos Alygizakis

Athens, May 2022



# Acknowledgments

I acknowledge Dr. Theodoros Giannakopoulos for the guidance and support. I also want to thank Dr. Anastasia Krithara and Dr. Charilaos Akasiadis for agreeing to be part of the examination committee and for evaluating the thesis. I gratefully acknowledge the crew of Mare Nigrum vessel, the researchers who participated in the Joint Black Sea Surveys in 2016 and 2017, the team responsible for the logistics (Dr. Peter Oswald, Michaela Mazanova, Peter Mazan, Martina Oswaldova among others) and the research group of Prof. Nikolaos Thomaidis for the analysis of seawater extracts in the facilities of National and Kapodistrian University of Athens. The data was generated in context of the EU/UNDP Project: Improving Environmental Monitoring in the Black Sea—Selected Measures (EMBLAS-Plus), ENPI/2018/389-859. Finally, I want to acknowledge the coordinator of the project, Dr. Jaroslav Slobodnik for the approval to use the data of the project.

To my family.



## Περίληψη

**Η** Μαύρη Θάλασσα είναι ένα σημαντικό οικοσύστημα, το οποίο επηρεάζεται από διάφορες ανθρωπογενείς πιέσεις, όπως ναυτιλιακές δραστηριότητες, εισροές λυμάτων από μεγάλες παράκτιες πόλεις και κυρίως φορτία από μεγάλα ποτάμια (π.χ. Δούναβης, Δνεϊστερος, Δνείπερος). Η χημική ρύπανση που μεταφέρουν τα ποτάμια στη Μαύρη Θάλασσα είναι σημαντική, δεδομένου ότι μόνο ο ποταμός Δούναβης απορρίπτει 6.550 m<sup>3</sup>/s. Αυτή η μελέτη εστιάζει στην ουκρανική υφαλοκρηπίδα (το βορειοδυτικό τμήμα της Μαύρης Θάλασσας) και διερευνά τις πηγές χημικών ουσιών από τα ποτάμια. Για την επίτευξη αυτού του στόχου, χρησιμοποιήθηκαν δεδομένα από δείγματα που συλλέχθηκαν από τις κοινές έρευνες της Μαύρης Θάλασσας (KEMΘ). Οι KEMΘ πραγματοποιήθηκαν το 2016 και 2017 στο πλαίσιο του έργου EU / UNDP EMBLAS II ([www.emblasproject.org](http://www.emblasproject.org)). Κατά τη διάρκεια της εκστρατείας KEMΘ, συλλέχθηκαν δείγματα θαλασσινού νερού, τα οποία εκχυλίστηκαν και αναλύθηκαν με αναλυτικές μεθόδους υψηλής απόδοσης, όπως υγρή χρωματογραφία φασματομετρία μάζας υψηλής διακριτικής ικανότητας. Η χημική ανάλυση δημιούργησε δεδομένα τύπου XML, τα οποία υποβλήθηκαν σε επεξεργασία χρησιμοποιώντας αλγόριθμους ανοιχτού κώδικα. Το αποτέλεσμα της επεξεργασίας ήταν η δημιουργία ενός σετ δεδομένων με τα ανιχνευμένα χημικά σήματα και την έντασή τους στους σταθμούς δειγματοληψίας. Το σετ δεδομένων χρησιμοποιήθηκε για τη δημιουργία εικόνων, που αντιπροσωπεύουν τη χωρική κατανομή των σημάτων. Οι εικόνες στη συνέχεια χρησιμοποιήθηκαν ως είσοδος σε ένα μοντέλο ταξινόμησης συνελικτικών νευρωνικών δικτύων βαθιάς μάθησης. Στόχος της μελέτης ήταν να δημιουργηθεί μια ολοκληρωμένη υπολογιστική διαδικασία για την εκτίμηση του δυναμικού ρύπανσης των σημαντικότερων ποταμών που συμβάλλουν (Δνείπερος και Δούναβης) στη θάλασσα της Ουκρανίας. Τέλος, κατασκευάστηκε μια ιστοσελίδα για τη διευκόλυνση της οπτικοποίησης και της αξιολόγησης των απο-

τελεσμάτων. Η δημιουργία τέτοιων μοντέλων μπορεί επίσης να χρησιμεύσει για την ιεράρχηση άγνωστων σημάτων, που είναι το κλειδί για την επίτευξη της λεγόμενης μη στοχευμένης ανάλυσης.

# Abstract

**B**lack Sea (BS) is an important ecosystem, which is affected by various anthropogenic pressures, such as shipping activities, wastewater inputs from large coastal cities and most importantly loads by major rivers (e.g., Danube, Dniester, Dnieper). The chemical pollution that rivers transfer to the BS is significant considering that the Danube river alone discharges 6,550 m<sup>3</sup>/s to the BS. This study focuses on the Ukrainian shelf (the northwestern part of the Black Sea) and investigates the river sources of chemicals in the shelf. To achieve this objective, data generated by the Joint Black Sea Surveys (JBSS) was used. JBSS took place in 2016 and 2017 in context of the EU/UNDP EMBLAS II project ([www.emblasproject.org](http://www.emblasproject.org)). During the JBSS campaign, seawater samples were collected, extracted and analyzed by high-throughput analytical methods such as liquid chromatography high-resolution mass spectrometry (LC-HRMS). The analysis resulted in data, which was processed using open-source algorithms to generate a dataset with the detected chemical signals and their intensity in the sampling stations. The dataset was used to generate images, representing the spatial distribution of the signals. The figures were then used as an input to a deep learning convolutional neural network classification model. The aim of the study was to create an end-to-end solution for the estimation of the pollution potential of the major contributing rivers (Dnieper and Danube) in the Ukrainian shelf. Finally, a dashboard to facilitate data visualization and results' evaluation was built. The generation of such models can also serve to the prioritization of unknown chemical signals, which is the key for non-target screening.

---

# Contents

List of Tables	iii
List of Figures	iv
List of Abbreviations	vi
<b>1 Introduction</b>	<b>1</b>
1.1 Problem description	1
1.2 Thesis structure	3
<b>2 Theoretical background</b>	<b>5</b>
2.1 Establishing the basic definitions	5
2.2 Deep learning	7
2.3 Activation functions	9
2.4 Loss functions	10
2.5 Training process	11
2.6 Evaluation metrics	13
2.7 Performance evaluation	15
2.8 Convolutional neural networks	15
<b>3 Chemicals and computational methods</b>	<b>19</b>
3.1 Chemicals and reagents	19
3.2 Sample collection and preparation	20
3.3 Instrumental analysis	22

## CONTENTS

---

3.4	Computational workflow	22
<b>4</b>	<b>Results</b>	<b>27</b>
4.1	Deep-learning model selection	27
4.2	Interactive dashboard for spatial distribution visualization	29
4.3	Tentatively Identified substance and their sources	31
<b>5</b>	<b>Conclusions and Future Work</b>	<b>37</b>

# List of Tables

3.1	Internal standards spiked in the JBSS samples, their CAS number, their Std. InChIKey and the sampling campaign in which the substances were spiked.	21
3.2	Instrumental setup for liquid chromatography and mass spectrometer (positive and negative ionization).	23
4.1	Characteristic examples of tentatively identified compounds with riverine origin (both Danube and Dnieper).	32
4.2	Characteristic examples of tentatively identified compounds and their origin.	34





# List of Figures

2.1	Representation of a neuron, which is the building unit of neural networks. The inputs $x$ are multiplied by the weights $w$ , and a bias is added. The outcome activates a non-linear function $f$	7
2.2	The layers of a deep neural network. A neural network may have one input layer, at least one hidden layer and one output layer	9
2.3	A typical convolution neural network (CNN) architecture.	16
3.1	Sampling points of the Joint Black Sea Surveys (JBSS2016 and JBSS2017). Both sampling campaigns were conducted in summer of 2016 and 2017 respectively.	20
3.2	Examples of simulated spatial distribution based on the non-target dataset. The plots depict sequentially cases of a signal that comes from Danube, from Dnieper and from unknown origin.	24
3.3	Illustration of the identification workflow, presenting an example of a successful identification.	26
4.1	Confusion matrices for the two best performing models using three different splits at the training set. The convolution neural network (CNN) proved to be the most precise with accuracy higher than 99%.	28
4.2	Screenshot from the developed application. Application is available at <a href="https://norman-data.eu/BS">https://norman-data.eu/BS</a>	30



# List of Abbreviations

AI	Artificial intelligence
BGD	Batch gradient descent
BS	Black Sea
CAS	Chemical abstracts service
CECs	Contaminants of emerging concern
CNN or ConvNet	Convolutional neural network
DEET	N,N-diethyl-meta-toluamide
EU/UNDP	European Union/United Nations Development Programme
FN	False negative
FP	False positive
GAMs	Generalized additive models
GS	Gradient descent
GUI	Graphical User interface
ICPDR	International commission for the protection of the Danube River
JBSS	Joint Black Sea Survey

## LIST OF ABBREVIATIONS

---

JDS4	Joint Danube Survey 4
LC-HRMS	Liquid chromatography high-resolution mass spectrometry
MAE	Mean absolute error
MBGD	Mini-batch gradient descent
MB	MassBank
ML	Machine learning
MLP	Multi-layer perceptron
MLR	Multiple linear regression
MSE	Mean squared error
OBI-Warp	Ordered bijective interpolated warping
OPLS-DA	Orthogonal projections to latent structures discriminant analysis
PCA	Principal component analysis
PCA-MLR	Principal component analysis - multiple linear regression
PPPs	Plant protection products
QTOF	Quadrupole time of flight
RC	Regenerated cellulose syringe filters
ReLU	Rectified linear unit
RTI	Retention time index
SGD	Stochastic gradient descent
SPE	Solid phase extraction

Std. InChIKey	Standard international chemical identifier key
TP	True positive
TPs	Transformation products
UPLC	Ultra pressure liquid chromatography
UV	Ultra violet
KEMΘ	Κοινές έρευνες της Μαύρης Θάλασσας

## LIST OF ABBREVIATIONS

---

# Chapter 1

## Introduction

The first report on the occurrence of organic contaminants of emerging concern (CECs) in the marine aquatic environment dates back to 1987 [1]. From very early, researchers recognized the importance of source detection. Chemicals may enter the marine environment from many anthropogenic sources such as agriculture surface run-offs [2], aquaculture [3], discharges from wastewater treatment plants [4], emissions from shipping activity [5], harbor activities [6], manufacture and construction [7]. A major source for many chemicals in the marine environment is the rivers [8–10]. The rivers transfer chemical pollution from urban settlements, manufacturing and agricultural regions to the marine environment [11]. For example, the Black Sea receives significant riverine inputs from rivers, especially Danube (average annual discharge of 6550 m<sup>3</sup>/s) [12], Dnieper (average annual discharge 1670 m<sup>3</sup>/s) [13] and Dniester (average annual discharge 310 m<sup>3</sup>/s) [14].

### 1.1 Problem description

Tracing sources involve appropriate design and execution of sampling campaigns at gradient distances from the sources [4, 8, 9, 15, 16] and use of appropriate statistical tests. The most commonly used approaches involve the use of principal component analysis (PCA) [15, 17–23] and linear regression between the concentration levels and salinity [4, 8, 10, 15, 24, 25]. Other less widely used approaches involve the application of pair-wise correlation analysis [22], OPLS-DA [23], PCA-MLR [26],

network analysis and decision trees [27] among others.

The majority of the marine studies reporting the occurrence of CECs attempt source detection or discuss potential origins of the chemical pollution but focus on specific classes of CECs using quantitative determinations through target screening, where reference standards are available [26, 28]. However, target screening is based on the preselection of certain contaminants and the use of reference standards and cover only a small proportion of CECs [29, 30]. The introduction of high-throughput analytical instrumentation such as liquid chromatography coupled to high-resolution mass spectrometry (LC-HRMS), has given unimaginable capabilities to the researchers, who are able to conduct non-target screening [31, 32]. These new approaches are able to cover a very wide universe of CECs contained in complex environmental samples given the limitations in extraction and instrumental sensitivity [33]. The challenge of these high-throughput methods is that they produce a high number of signals (typical many thousands for each ionization), the structural elucidation of which is not feasible because they require high time and effort [34]. Therefore, the key step to non-target screening is the application of prioritization strategies, so that elucidation efforts are focused on the most relevant chemicals based on the goals of the study [35–37].

The goal of this study was to create an open-source workflow for untargeted source detection in the marine aquatic ecosystem of the Ukrainian shelf (northwest Black Sea). This study reports a spatial distribution approach to support prioritization activities in non-target screening. To the authors' knowledge, this is the first study in the literature that uses deep learning as a prioritization approach. The created workflow was applied in real seawater samples collected from the Joint Black Sea Surveys (JBSS) that took place in 2016 and 2017. The deep learning models were trained using the data of 2016 and were used to make predictions for both sampling periods. Finally, an interactive Dash application was programmed to visualize the results and support the identification of unknown substances.



## 1.2 Thesis structure

The thesis is organized in four chapters. Chapter 1 introduced the importance of tracking sources of chemical pollution in the marine environment and explained the need of developing robust prediction models for their detection. Chapter 2 provides the theoretical background in machine learning and deep neural networks. Chapter 3 summarizes the materials, instrumentation and methods used to generate the dataset and describes the modeling workflow. Moreover, it describes the computational resources that were used and the workflow that was developed. Chapter 4 presents the results of the analysis and discusses the findings. The model selection and the description of the interactive dashboard application are discussed in detail. Finally, some examples of chemicals that were tentatively identified and their sources are presented. Chapter 5 provides a summary of the work and discusses potential future research directions.



# Chapter 2

## Theoretical background

### 2.1 Establishing the basic definitions

The term of Artificial Intelligence (AI) was established in 1956 at the Dartmouth College conference. Since then, AI has played a critical role in the human society because of the development of the useful real-world applications such as image recognition, natural language processing, driving assistance, art generation, drug discovery among many other examples. AI received and will continue to receive an increasing attention from the researchers due to the improved computational capabilities and the possibility to store huge amounts of data (e.g., images uploaded to social media platforms, hours of uploaded footage in video platforms, payment transactions etc.).

AI has developed from self-trained systems without human interference. As a scientific endeavor, machine learning (ML) grew out of the quest for AI. This particular domain of AI is called ML. ML is a set of applied statistical methods that can learn from data given a task and discover patterns or learn to map the data on classes. ML differentiates from the classical rule-based approach of an algorithm, which is an explicitly defined procedure. Therefore, ML is not an algorithmic process but a set of methods that are not programmed but trained on sets of data. A more formal definition of ML is the following: "A computer program is considered to be learning from experience  $E$  in relation to a class  $T$  task and a performance measure

P if its performance in class T tasks as valued by P are improved by experience E” [38].

There are variations in the types of machine learning. However, it is possible to divide them in categories based on the nature of the problem in the following broad categories:

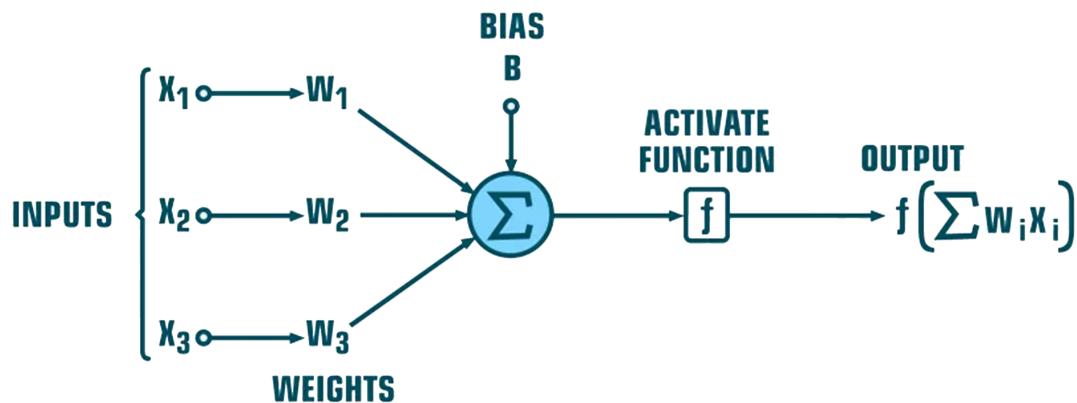
- **Supervised Learning:** The algorithm is trained to find associations at the input data that is accompanied with their corresponded labels. The set of these pairs is called *training set* and the process of calculating such a function from the dataset is called *training*. The training set consists of examples, which are also known as *instances*. The label is produced in most of the cases by human during a process called *annotation process*. During the training phase, the model uses the training set to learn and extract useful patterns. The objective of supervised learning is the calculation of a function that sufficiently generalizes to the input data, so that it is able to assign to the correct labels in new input data. It is critical that the new input data consists of instances that the model did not use during training [39, 40].
- **Unsupervised Learning:** The algorithm is trained with the aim to identify patterns in the data without feedback from labels. A typical example of unsupervised learning is clustering. This method aims at grouping the input data in groups whose members are similar to each other, and different from members of other groups [41].
- **Reinforcement Learning:** In reinforcing learning, the datasets are not labelled. The operation of reinforcement learning is based on a rewarding/punishing system [42]. The choices made to guide to an outcome are judged based on whether they had a positive or a negative contribution. In this way, the model can choose whether to make the same choices again or to follow a different path of action. The different path may involve a change in one or more choices [43]

For most applications, the input samples are subject to preprocessing. This step transforms the samples in a new parameter space to ease and accelerate training. The preprocessing stage is also known as *feature extraction*. In most of the tasks, in

which the raw data is multi and high dimensional, preprocessing step is essential.

## 2.2 Deep learning

A sub-domain of ML is deep learning (DL). DL has gained significant growth in popularity over the last years. In conventional ML, the algorithms rely on feature engineering and sophisticated feature selection techniques, which may require considerable domain knowledge. In contrary, DL methods are able to process unstructured data (images, text, voice, etc.) without applying any prior feature extraction. DL uses non-linear activated modules able to transform the raw data in high-level feature representations. This capability has led to revolutionary developments in many domains including computer vision and natural language processing. DL solved complex problems that researchers were unable to resolve for many years. DL turned out to have outstanding results at discovering complex patterns in high-dimensional data.



**Figure 2.1:** Representation of a neuron, which is the building unit of neural networks. The inputs  $x$  are multiplied by the weights  $w$ , and a bias is added. The outcome activates a non-linear function  $f$

The basic building unit that constructs a neural network is called *neuron* or *perceptron* [44]. A neuron is able to perform two operations: a linear (or affine) transformation to its input and a non-linear function application to the output. A

neuron is depicted in Figure 2.1 and is described by the equation:

$$z = f(w * x + b)$$

The input is a vector  $x$  and is multiplied by the weight matrix  $w$ , which is a tunable parameter. The bias is added to the multiplication outcome. These operations comprise the linear transformation of the perceptron. Afterwards, every component of the result vector is passed through a non-linear function  $f$ , which is called *activation function*. Training of a DL network is the tuning of neuron parameters  $w$  and  $b$  based on the labels of the instances included in the training set. The most popular activation functions that are used are the sigmoid, tanh, ReLU and softmax and will be described in Section 2.3.

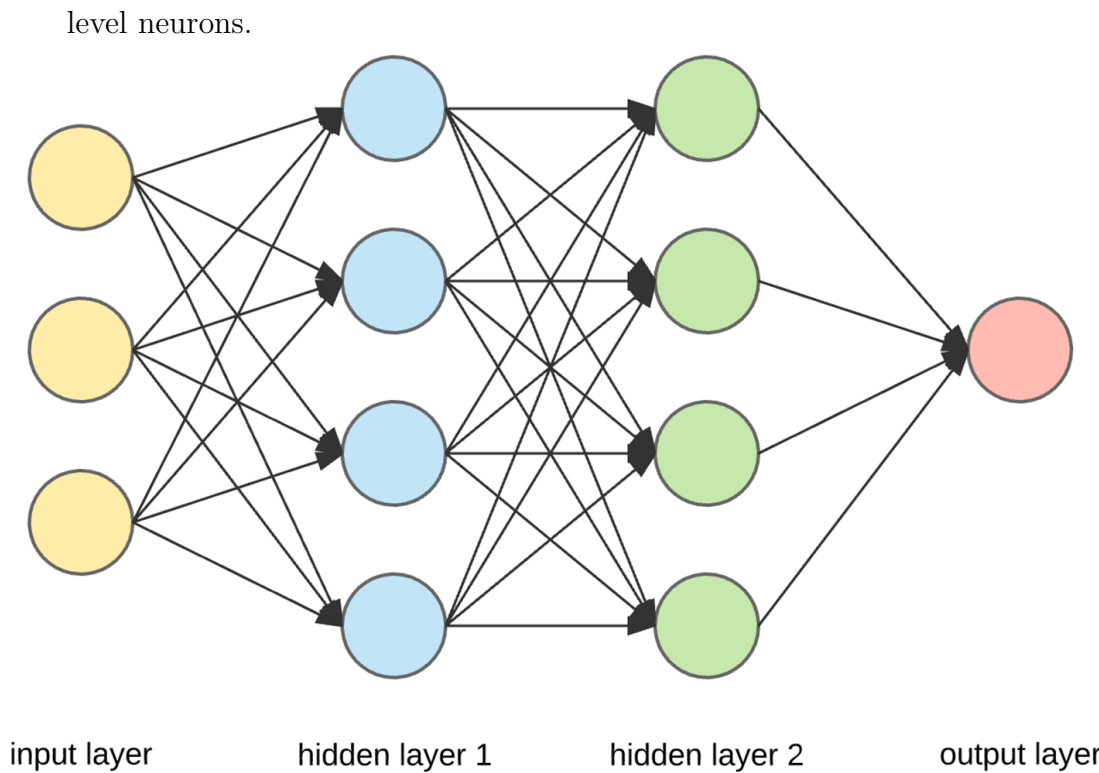
In the most common case, a typical neural network consists of multiple layers. A neural network is illustrated in Figure 2.2.

The output of one neuron becomes the input of another neuron in the next layer. The network is able to estimate non-linear functions in such hierarchical networks. If the process is repeated, the network becomes a multi-layer perceptron (MLP) network. MLP networks are able to learn complex functions. The neurons are organized by levels (Figure 2.2) which are divided into 3 categories:

- Input layer
- Hidden layer
- Output layer

A key property of neural networks is the way of connection between neurons of all layers. Based on this property, there are the following categories:

- **Fully Connected:** Networks in which each neuron of a level connects to all neurons of the next level.
- **Partially Connected:** Networks that exist at each level neurons that are not connected to all the neurons of the next level.
- **Feed-forward:** Networks in which neural connections do not form a circle. In this case, no neuron promotes its output in neurons of previous levels.
- **Feedback:** Networks that have neurons that advance their output to previous



**Figure 2.2:** The layers of a deep neural network. A neural network may have one input layer, at least one hidden layer and one output layer

## 2.3 Activation functions

Training a multi-layer neural network requires the choice of the network architecture and the activation functions. During the training process, the output of a neuron is passed to the next layer only if activated by a certain function. Sigmoidal activations functions are usually employed. The sigmoidal activation function is provided by the following equation:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

However, for a deep neural network, there is a computational advantage when using a non-sigmoidal *rectified linear unit (ReLU)* [45]. Moreover, the ReLU is preferred over a sigmoidal for classification problems [46]. The ReLU handles better the output of sigmoid function and the vanishing gradient phenomenon which may occur during the training process. The sigmoidal activation function is provided by

the following equation:

$$ReLU(x) = \max(0, x)$$

Another widely-used activation function is the hyperbolic tangent function (tanh) with the following mathematical expression:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

The activation function of the final layer of the network is based on the response type. No activation function is required, when the prediction is a continuous variable. However, when the prediction is a classification output, then softmax activation function is mainly used. Softmax activation is declared by the following equation.

$$\text{softmax}(z)_i = \frac{e^{z_i}}{\sum_{j=1}^C e^{z_j}}$$

Softmax activation normalizes the input into a probability distribution.

## 2.4 Loss functions

Training a neural network requires the definition of the loss function. During training, the predicted output is compared to the ground truth to evaluate its precision. For this purpose, a loss function is utilized. For problems in which the prediction is a continuous variable, the mean squared error (MSE) is frequently used. MSE between the predicted value ( $\hat{y}$ ) and the ground truth ( $y$ ) can be calculated using the following equation

$$MSE(y, \hat{y}) = \frac{1}{n} * \sum_{n=1}^n (y_i - \hat{y}_i)^2$$



Alternative loss function is the mean absolute error (MAE) with the following mathematical expression:

$$MAR(y, \hat{y}) = \frac{1}{n} * \sum_{n=1}^n |y_i - \hat{y}_i|$$

The loss function that is used in classification problems is called categorical cross-entropy and is defined according to the following equation:

$$J(y, \hat{y}) = - \sum_{n=1}^n y_i * \log \hat{y}_i$$

The user can define any loss function he/she wants, as long as the function is minimized when the predicted value is as close enough to the ground truth.

## 2.5 Training process

The learning process is an iterative procedure that is executed in learning cycles called *epochs*. The training phase stops when a criterion is fulfilled. The objective of the training is to find the network's parameters (weights  $w$  and biases  $b$ ) that minimize the loss function. To achieve this objective, the use of an optimization algorithm, which is called *optimizer*, is needed [47]. The simplest optimizer is the gradient descent (GS). GS is one of the most widely used optimizer together with its variations.

In GS, all training samples are involved in the calculation of the cost. GS is successful when the shape of the objective function is convex. This method minimizes the cost function using the first partial derivative. To achieve this goal, it performs two steps iteratively. Firstly, it computes the slope (gradient) that is the first-order derivative of the function at the current point. In the second step, it moves in the opposite direction of the slope increase from the current point by the computed amount. In each iteration of the algorithm, the partial derivative multiplied by a numeric parameter called learning rate ( $\alpha$ ) is subtracted from each parameter of the network. The mathematical formula is the following

$$W_i^{e+1} = W_i^e - a * \frac{\partial J(y, \hat{y})}{\partial W_i^e}$$

The choice of the learning rate is crucial. When using small learning rates, the model will require more steps to converge. High learning values might lose the minima and the model may never converge. This method can be considered as a criterion to signal the end of the training process.

The GS algorithm is executed after a forward propagation has been executed for all the data in the network. However, part of the dataset is fed to the network before an update is made. The reason of this action is that GS can become computationally expensive in applications with millions of trainable parameters. Based on this criterion, we distinguish the following basic cases:

- **Batch Gradient Descent (BGD)**: In BGD, all the training data is taken into consideration to take a single step. The average of the gradients of all the training examples is considered and then the mean gradient is used to update the parameters.
- **Stochastic Gradient Descent (SGD)**: In contrast to BGD, the network parameters are updated for each input sample separately. SGD uses a random sample  $i$  rather than all samples, to update the gradient per iteration. In SGD, unnecessary calculations do not happen. However, problems may arise when approaching at a local minimum of the curve and especially when the slopes of the minimum are steep. This method was proposed by Robbins in 1951 [48].
- **Mini-Batch Gradient Descent (MBGD)**: MBGD is characterized by random sample selection but in larger groups. It combines randomness and fast calculation of the stochastic case using more input data. In this way, it reduces noise from individual noisy input samples.

For deep learning networks, an instance is passed through the network of neurons until the final layer and the loss function calculates the error between the prediction and the ground truth. Afterwards, the network is informed about the error and adapts its parameters (weights and biases). This process is iterative and resembles

an optimization process. Thus, it can be solved utilizing the methods that were previously mentioned. For network architectures with multiple layers, a chain rule is applied to calculate the gradients of the layers next.

$$\frac{\partial J(y, \hat{y})}{\partial W_i} = \frac{\partial J(y, \hat{y})}{\partial z_{L-1}} * \frac{\partial J(y, \hat{y})}{\partial z_{L-2}} * \dots * \frac{\partial J(y, \hat{y})}{\partial W_i}$$

For computational reasons, layers are updated backwards from the latter to the former. This operation is referred as backpropagation [48]. The algorithm implements a modification to each weight of the network and considers the error that occurs for a specific input, the corresponding desired output and the network recall. The network weights are tuned according to their contribution to the overall network error. This adjustment is made following the opposite direction from the data flow.

## 2.6 Evaluation metrics

Evaluation metrics are used to compare different models. The selection of the most appropriate evaluation metric depends on the nature of the modeling task and the dataset. The evaluation metrics must be applied to the test set, which includes data that the models have never seen during the training. The most basic metric evaluation of a model is accuracy. Accuracy expresses the success rate of the model in classifying the samples into the correct categories. It is expressed using the following equation:

$$Accuracy = \frac{Correct_{predictions}}{Total_{predictions}}$$

This metric is rarely used and its use must be avoided when the number of instances in the data classes is not balanced. The correct predictions of the larger class can overshadow the incorrect predictions of the smaller classes. For this reason, precision and recall are defined. Precision is the ratio of the correct prediction results

of a class to the total number of forecasts in this class. It is defined by the following equation:

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

Recall is the ratio of the correct prediction results of a class to the total number of samples in this class. It is defined by the following equation:

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

Good performance of models is indicated by high precision and recall. However, there is a trade-off between these metrics. A metric that combines precision and recall is the F1 score, which is defined according to the following equation:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} = \frac{TruePositive}{TruePositive + \frac{1}{2}(FalsePositive + FalseNegative)}$$

F1 score can be calculated for all classes. The macro-averaged F1 score (or macro F1 score) is computed by taking the arithmetic mean (unweighted mean) of all the per-class F1 scores. This method treats all classes equally regardless of their support values. Support refers to the number of actual occurrences of the class in the dataset. The weighted-averaged F1 score is calculated by taking the mean of all per-class F1 scores while considering each class's support. The 'weight' essentially refers to the proportion of each class's support relative to the sum of all support values. Micro averaging computes a global average F1 score by counting the sums of the True Positives (TP), False Negatives (FN), and False Positives (FP). We sum the TP, FP, and FN values in all classes and then plug them into the F1 equation to get the micro F1 score.

## 2.7 Performance evaluation

Bias and variance are two types of errors. The bias occurs from the false estimations during training. It is high in cases in which the structure of the network is too simple to learn a complex representation. High bias is also known as *underfitting*. Variance is the opposite to bias. It derives from the classifier's sensitivity in minor input changes (noise). Noise may come from the dataset or from the random behavior in the learning algorithm itself [49]. In case of high variance, the classifier learns the training set perfectly. This phenomenon is known as *overfitting* and causes issues in the generalization capability of the model.

Splitting of the dataset into training set and test set is performed to avoid underfitting and overfitting. Moreover, to avoid the risk to overfit the network's hyperparameters to a certain test set, a third subset called *validation set* is created. The validation set is created from the split of the training set (typically 80-20 % or 70-30 % depending on the size of the dataset). The network is trained as described in Section 2.5 but the validation set is used to fine-tune the hyperparameters (the network's weights).

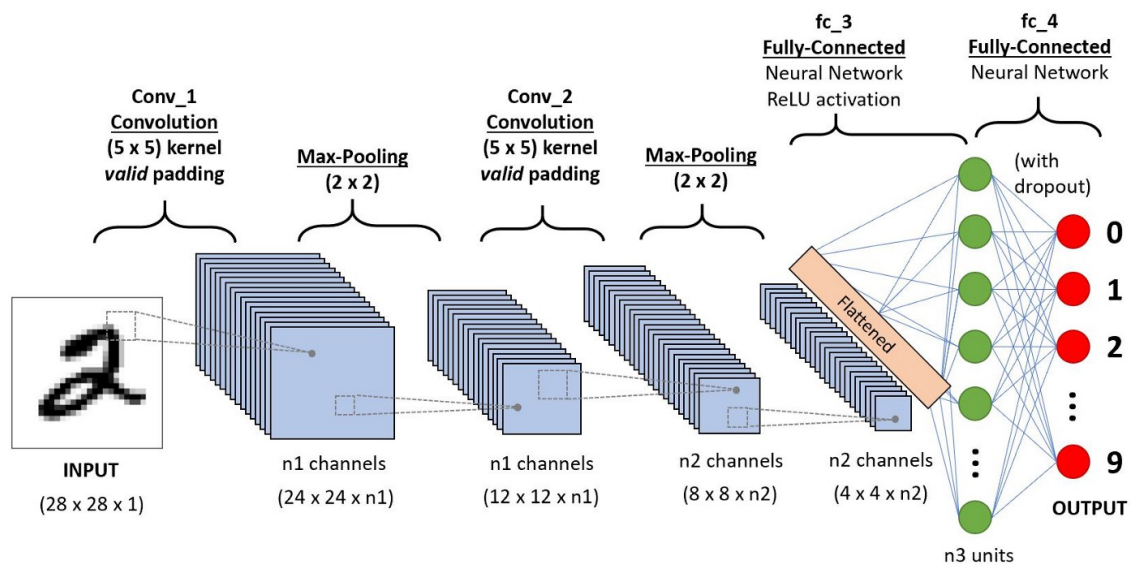
Training of a deep neural network can be affected by the choice of a number of hyperparameters. Some critical choices are the structure of the model, the weight initialization techniques, the learning rate, early stopping, dropout and batch normalization layers. The dropout is one of the oldest regularization techniques in DL [50]. At each training iteration, the network drops random neurons with a predefined probability (typically 20% to 50%). In practice, neuron outputs are set to zero. These neurons do not participate in the loss computation and thus they do not receive weight updates. Different neurons are dropped at each epoch. Dropout is a very common method to avoid overfitting in neural networks.

## 2.8 Convolutional neural networks

Convolutional neural network (CNN, or ConvNet) is a class of artificial neural networks. They are feed-forward fully-connected regularized versions of multi-layer perceptrons. They have a broad application in the fields of computer vision due to

their advantage to process data that come in its raw form and thus avoid manual feature selection.

The architecture of a CNN is analogous to that of the connectivity pattern of Neurons in the human brain. CNN was inspired by the visual cortex. Individual neurons respond to stimuli in a restricted region of the visual field. A collection of such fields overlap to cover the entire visual area. The architecture of a typical CNN is illustrated in Figure 2.3



**Figure 2.3:** A typical convolution neural network (CNN) architecture.

The role of the CNN is to reduce the images into a form which is easier to process, without losing features which are critical for getting a good prediction. Convolution is an operation for dimensionality reduction. The magnitude of the reduction depends on the size of the kernel. The kernel slides along the input matrix with a certain stride Value, generating a feature map, which in turn contributes to the input of the next layer. The filter moves until it parses the complete width, hops down to the beginning of the image with the same stride value and repeats the process until the entire image is traversed.

CNNs are not limited to only one convolutional layer. The first convolutional layer is responsible for capturing the low-level features such as edges, color, gradient orientation, etc. Additional convolutional layers, capture the high-Level features.

After a convolution operation, there is a pooling layer. The pooling layer is responsible for reducing the spatial size of the convolved feature. Pooling decreases the computational power required to process the data through dimensionality reduction. Pooling can act as noise suppressant. It discards the noisy activations altogether and also performs de-noising along with dimensionality reduction. The described process (convolution and pooling) enables the model to understand the features. Afterwards, there is a flatten layer that flattens the final output and feeds it to a regular neural network for classification purposes. The flattened output is passed to a feed-forward neural network and backpropagation applied to every iteration of training. Over a series of epochs, the model is able to distinguish between dominating and certain low-level features in images and classify them using the Softmax Classification technique. There are various architectures of CNNs (e.g. VGGNet, LeNet, ResNet), which have been key in building algorithms.





# Chapter 3

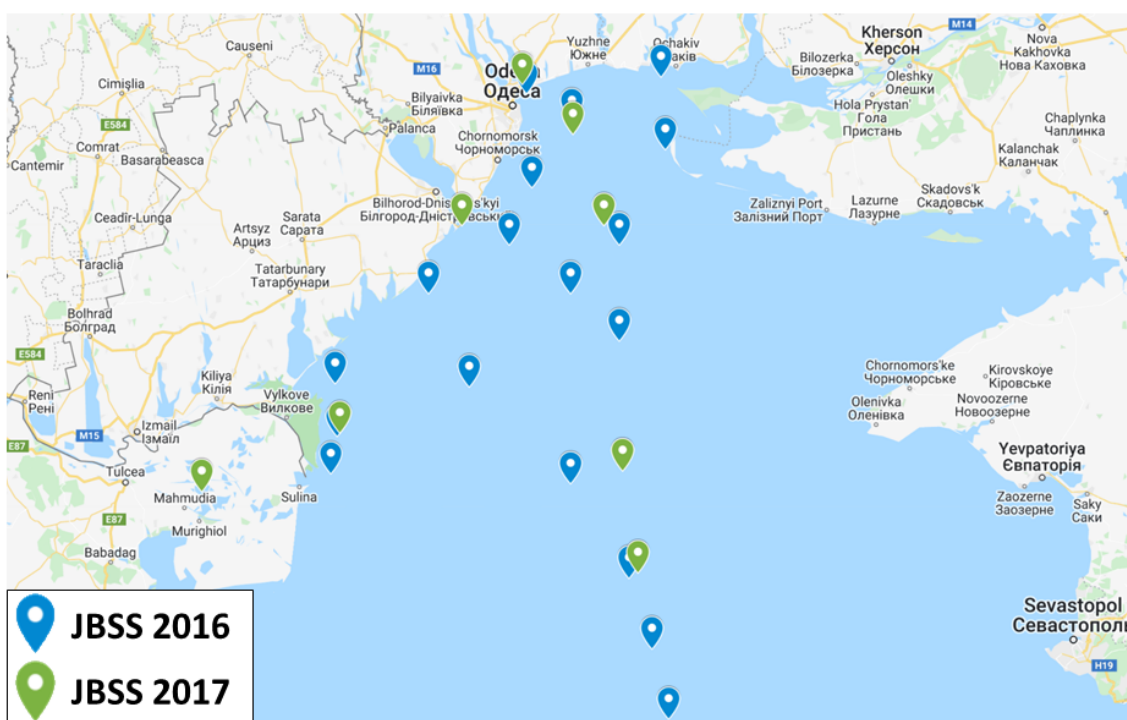
## Chemicals and computational methods

### 3.1 Chemicals and reagents

Acetonitrile and methanol were obtained from Merck (Darmstadt, Germany). 2-propanol was purchased from Fisher Scientific (Geel, Belgium). All solvents were of the highest possible analytical grade (UPLC grade). Distilled water was produced by the Milli-Q Direct-Q UV purification device manufactured from Millipore (Bedford, USA). Solid sodium hydroxide monohydrate for trace analysis of purity more than 99.9995 % , ammonium formate, ammonium acetate and formic acid 99 % were obtained from Fluka (Buchs, Switzerland). Empty solid phase extraction (SPE) polypropylene cartridges (6 mL) were obtained from Phenomenex (Torrance, USA). The same company provided the sorbent materials: Septra ZT (Strata-X), weak cation exchange Septra ZT-WCX (Strata-X-CW) and weak anion exchange ZT-WAX (Strata-X-AW). The polar sorbent Isolute ENV+ and the frits (pore size 20  $\mu\text{m}$ ) were bought from Biotage (Ystrad Mynach, UK). Before instrumental analysis, the extracts were filtered from regenerated cellulose syringe filters (RC) with 0.2  $\mu\text{m}$  pore size (15 mm diameter) and were obtained from Phenomenex (Torrance, USA).

## 3.2 Sample collection and preparation

55 seawater samples were collected during the JBSS2016 survey and 20 seawater samples were collected during the JBSS2017 survey. The study area for contamination was covered by three transects of Black Sea; the western side close to Danube (Ukrainian shelf), the eastern Black Sea close to Georgia and the central side including (open sea) sampling points across the length of Black Sea and out of reach of any coastal city. This study focused on the seawater samples from the Ukrainian shelf. Therefore, the relevant samples were 18 from JBSS2016 and 8 from the JBSS2017 (Figure 3.1).



**Figure 3.1:** Sampling points of the Joint Black Sea Surveys (JBSS2016 and JBSS2017). Both sampling campaigns were conducted in summer of 2016 and 2017 respectively.

The samples were collected by the scientific crew during the expedition at the deck of the Mare Nigrum vessel. Both sampling campaigns took place in summer (August and beginning of September). The samples were preprocessed immediately upon collection. The samples were spiked with internal standards (Table 3.1) and were cleaned-up and preconcentrated by SPE; 2.5 L of seawater, spiked with internal standards passed through the in-house four-sorbent cartridge (200 mg Strata-X, 150

**Table 3.1:** Internal standards spiked in the JBSS samples, their CAS number, their Std. InChIKey and the sampling campaign in which the substances were spiked.

Compound Name	CAS	Std. InChIKey	Campaign
5-Methyl-1H-benzotriazole-d6	1246820-65-4	LRUDIIUSNGCQKF-RLTMCGQMSA-N	2016
Amisulpride-d5	71675-85-9	NTJOBXMMWNYJFB-SGEUAGPISA-N	2016
Amphetamine-d6	205437-60-1	KWTSXDURSIMDCE-ZQLKWRTGSA-N	2016, 2017
Atenolol-d7	1202864-50-3	METKIMKYRQLGS-SVMCCORHSA-N	2016
Atorvastatin-d5	222412-82-0	XUKUURHRXDUEBC-BDXWSXJNSA-N	2016, 2017
Atrazine-d5	163165-75-1	MXWJVTOOROXXGIU-SGEUAGPISA-N	2016
Benzophenone-d10	22583-75-1	RWCCWEUUXYIKHB-LHNTUAQVSA-N	2016, 2017
BPA-d16	96210-87-6	IISBACLAFKSPIT-MAJJRYNQSA-N	2016, 2017
Carbamazepine-d8	1538624-35-9	FFGPTBGBLSHEPO-PKSNNKEVSA-N	2016, 2017
Cetirizine-d8	774596-22-4	ZKLPARSLTMPFCP-DTSBCCDKSA-N	2016, 2017
Ciprofloxacin-d8	1130050-35-9	MYSWGUAQZAJQK-SQUIKQQTSA-N	2016, 2017
Citalopram-d6	1246819-94-2	WSEQXVZVJXJVFP-WFGJKAKNSA-N	2016, 2017
Clozapine-d8	1185053-50-2	QZUDBNBUXVUHMW-JNJBWJDISA-N	2016
Codeine-d6	1007844-34-9	OROGSEYTTFOCAN-JLXZPSIDSA-N	2016
Decoquinat-d5	1453100-61-2	JHAYEQICABJSTP-ZTIZGVCASA-N	2016
Diuron-d6	1007536-67-5	XMTQQYYKAHVGBJ-WFGJKAKNSA-N	2016
Fenbendazole-d3	1228182-47-5	HDDSHPAODJUKPD-FIBGUPNXSA-N	2016
Iohexol-d5	66108-95-0	NTHXOOBQLCIOLC-OPCJXEHASA-N	2016, 2017
Ketamine-d4	1246815-97-3	HPQHIBKAPINDEN-KDWOUJHVSA-N	2016
Lamotrigine- <sup>13</sup> C <sub>3</sub> ,d3	1246815-13-3	PYZRQGRPPTADH-MKOZQUTQSA-N	2016, 2017
Mefenamic acid-d3	1189707-81-0	HYYBABOKPJLUIN-FIBGUPNXSA-N	2016, 2017
Metformin-d6	1185166-01-1	XZWYZLIPXDOLR-WFGJKAKNSA-N	2016, 2017
Metronidazole-d4	1261392-47-5	VAOCPAMSLUNLGC-RRVWJQJTSAN	2016, 2017
Ranitidine-d6	1185238-09-8	VMXUWOKSQNHCOA-RUESZMOGSA-N	2016, 2017
Ritonavir-d6	155213-67-5	NCDNCNXCXDHOMX-GMBJSHJASA-N	2016, 2017
Saccharin- <sup>13</sup> C <sub>6</sub>	1286479-01-3	CVHZOJJKTDOEJC-IDEBNHGSA-N	2016, 2017
Sulfadimidine-d4	1020719-82-7	ASWVTGNCAZCNR-LNFUJOGGSA-N	2017
Sulfamethazine-d4	1020719-82-7	ASWVTGNCAZCNR-LNFUJOGGSA-N	2016, 2017
Sucralose-d6	1459161-55-7	BAQAVOSOZGMPRM-UQRHTAPNSA-N	2016
Tramadol-d6	1109217-84-6	TVYLLZQTGLZFBW-DTPCVOBTSA-N	2016, 2017
Valsartan- <sup>13</sup> C <sub>5</sub> , <sup>15</sup> N	-	ACWBQPMHZXGDFX-UDHSYOPXSA-N	2016, 2017
Venlafaxin-d6	1020720-02-8	FETCANMPQJPEEP-WFGJKAKNSA-N	2016, 2017

mg Isolute ENV+, 100 mg Strata-X-AW and 100 mg Strata-X-CW), previously preconditioned with 3 mL methanol and 3 mL water. Cartridges were eluted with 4 mL 50:50 methanol:Ethyl acetate containing 2% of ammonia, followed by 2 mL 50:50 methanol:ethyl acetate containing 1.7% of formic acid. The extracts were stored in freezer until their instrumental analysis.

### 3.3 Instrumental analysis

The extracts were injected onto a Thermo Acclaim RSLC C18 column with dimensions 2.1 x 100 mm, particle size 2.2  $\mu\text{m}$  (Dreieich, Germany) preceded by a guard column of the same packaging material connected to a Bruker Maxis Impact QTOF (Maxis Impact, Bruker Daltonics, Bremen, Germany). For positive ionization, the aqueous phase consisted of water:methanol 90:10 with 5 mM ammonium formate and 0.01% formic acid and the organic phase was methanol with 5 mM ammonium formate and 0.01% formic acid. For negative ionization, the aqueous phase consisted of water:methanol 90:10 with 5 mM ammonium acetate and the organic phase was methanol with 5 mM ammonium acetate. Gradient for both ionizations for organic phase was 1% (0-1 min), 39% (1-3 min), 99.9% (3-14 min), 99.9% (14-16 min), 1% (16-16.1 min), 1% (16.1-20 min) and flow rate gradient was 0.2 mL  $\text{min}^{-1}$  (0-3 min), 0.4 mL  $\text{min}^{-1}$  (3-14 min), 0.48 mL  $\text{min}^{-1}$  (16-19 min), 0.2 mL  $\text{min}^{-1}$  (19.1-20 min). The instrumental conditions are available in Table 3.2. All samples were injected in positive and negative ionization in data-dependent (5 most abundant precursors) and data-independent (full scan collision energy and 25 eV) acquisition mode. The data-dependent chromatograms were processed in context of this study.

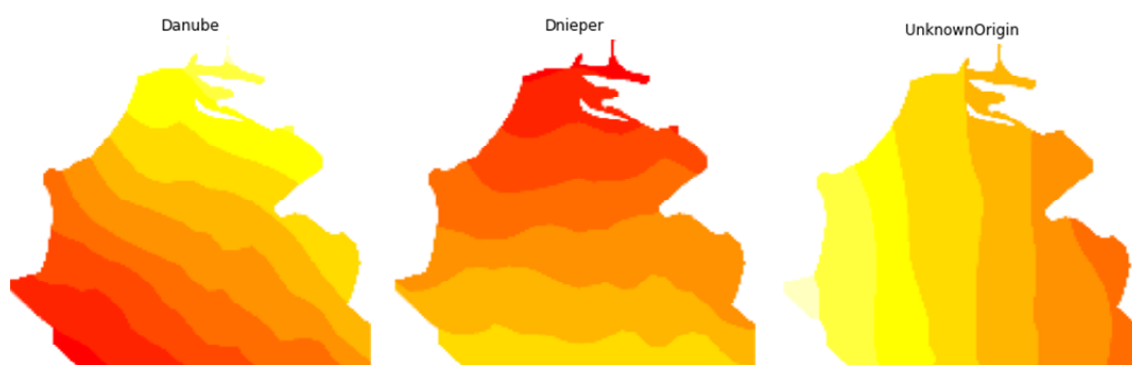
### 3.4 Computational workflow

The calibrant substance sodium formate and sodium acetate for positive and negative ionization respectively was injected in the beginning of each chromatographic run. The  $m/z$  peaks of the calibrant substance were used to recalibrate the whole chromatogram using HPC fitting algorithm, which is embedded in DataAnalysis 4.3. (Bruker Daltonics, Bremen, Germany). This calibration method ensures

**Table 3.2:** Instrumental setup for liquid chromatography and mass spectrometer (positive and negative ionization).

Gradient Elution Program		Positive Ionization	
Time (min)	% B	Electrospray Ionization Parameters	
0	1	Capillary Voltage	2500V
1	1	End plate offset	500V
3	39	Nebulizer	2 bar
14	99.9	Drying gas	8 L min <sup>-1</sup>
16	99.9	Drying temperature	200°C
16.1	1	(A) Water : Methanol 90:10 5mM HCOONH <sub>4</sub> with 0.01% HCOOH	
20	1	(B) Methanol 5mM HCOONH <sub>4</sub> with 0.01% HCOOH	
Gradient Elution Program		Negative Ionization	
Time (min)	% B	Electrospray Ionization Parameters	
0	1	Capillary Voltage	3500V
1	1	End plate offset	500 V
3	39	Nebulizer	2 bar
14	99.9	Drying gas	8 L min <sup>-1</sup>
16	99.9	Drying temperature	200°C
16.1	1	(A) Water : Methanol 90:10 5mM CH <sub>3</sub> COONH <sub>4</sub>	
20	1	(B) Methanol 5mM CH <sub>3</sub> COONH <sub>4</sub>	
Positive and Negative Ionization			
Gradient	0 min (1% B), 1 min (1% B), 3 min (39% B), 14 min (99.9% B), 16 min (99.9% B), 16.1 min (1% B), 20 min (1% B)		
Flow	0 min (200 ul min <sup>-1</sup> ), 2 min (200 ul min <sup>-1</sup> ), 14 min (400 ul min <sup>-1</sup> ), 16 min (480 ul min <sup>-1</sup> ), 20 min (200 ul min <sup>-1</sup> )		
Injection volume	5 µL		

mass accuracy below 2 mDa during the chromatographic run for  $m/z$  from 50 Da to 1000 Da. CompassXport 3.0.9.2 (Bruker Daltonics, Bremen, Germany) was used for exporting files in mzML format. The mzML files were processed with an established processing workflow using xcms [51] and CAMERA R-packages [52]. The functions for peak detection, matching peaks in the samples and retention time alignment (OBI-Warp algorithm) are included in the xcms package, whereas functions for componentization based on retention time and peak shape and functions for annotation of adducts and isotopic peaks are included in the CAMERA R-package. The input parameters for peak-picking, grouping and retention time alignment were optimized based on three level incomplete factorial design (Box-Behnken design) [53] for the data generated for the specific instrumental setup [54]. Peaks detected in the blank samples (with an intensity ratio below one order of magnitude) were removed. The final dataset consisted of 30489 components (11432 for year 2016 and 19057 for the year 2017). Spatial generalized additive model (GAMs) was used to predict the signal intensity near the sampling stations given as input the coordinates [55]. The computational workflow is available as R script (“1.Dataset generation.R”<sup>1</sup>). The workflow generated 30489 figures, one for every chemical signal. Three signals with three different origins are depicted in Figure 3.2.



**Figure 3.2:** Examples of simulated spatial distribution based on the non-target dataset. The plots depict sequentially cases of a signal that comes from Danube, from Dnieper and from unknown origin.

Manual effort was made to create the training set with total size of 1406 instances. The training set consisted of 652 figures showing a clear origin of the chemical signal

---

<sup>1</sup><https://github.com/nalygizakis/CECinBS>

from Danube, 559 figures showing origin from Dnieper and 195 instances of unknown origin. 80% of the training set was used for training and 20% for validation. To verify the performance of the models, three different training set splits were performed. Moreover, a hold out dataset was also created to evaluate the performance of the classifiers in instances never seen before by the models. The hold out set consisted of 224 instances (109 examples with origin from Danube, 78 examples with origin from Dnieper, 37 examples of unknown origin). The datasets (all three splits) are available at Zenodo<sup>1</sup>. TensorFlow machine learning library, which is developed by Google Brain, was used in this study [56]. The python Jupyter notebooks “2.Image recognition.ipynb” and “3.Image recognition\_Other split.ipynb”<sup>2</sup> were used to investigate the most accurate deep learning classifier. Given that the training and evaluation sets were unbalanced (not equal instances per class), the overall F1 score was used as the evaluation metric of the accuracy. Four deep learning architectures were evaluated; a three-layer model (regarded as baseline model), a convolutional neural network (CNN), VGG Net and ResNet. The best performing classifier which was the CNN model was used to predict the source of all components (“4.Make predictions\_CNN.ipynb”<sup>2</sup>). The data is summarized and is presented in an interactive Dash application (script “5.Dash application.ipynb”<sup>2</sup>), which is online available at <https://norman-data.eu/BS/>. The dependencies to run the application are available at the file “requirements.txt”. The workflow is reproducible and all scripts can run on the Google Colab cloud infrastructure.

Finally, the non-target screening identification workflow (script “6.Identification script.R”<sup>2</sup>) was used to reveal the identity of chemicals with clear pollution sources in the Ukrainian shelf. The script uses the MassBank (MB) text records from the official GitHub repository<sup>3</sup> and performs 1) filtering based on the mass accuracy (comparison between the experimental m/z and the theoretical m/z provided by the molecular formula and the adduct in the MB records) and 2) spectral match between the experimental HRMS/MS spectra and the MB library spectra. The MB records were retrieved from the repository on the 19th of March 2022. The m/z filter

---

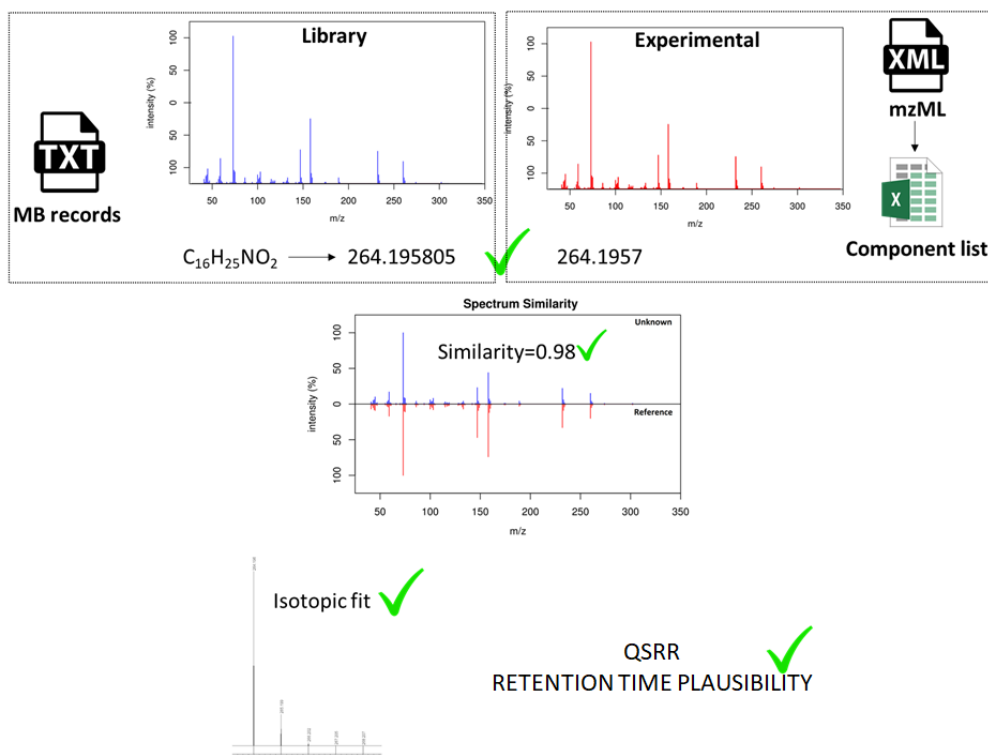
<sup>1</sup><https://doi.org/18610.5281/zenodo.6474592>

<sup>2</sup><https://github.com/nalygizakis/CECinBS>

<sup>3</sup><https://github.com/MassBank/MassBank-data>

### 3.4 : Computational workflow

was realized with help from functions included in the *enviPat* R-package [57] and the spectral match was achieved using the *OrgMassSpecR* R-package (Dodder and Mullen, 2017; Stein and Scott, 1994). The results obtained from this identification workflow (Figure 3.3) and using the following parameters (mass accuracy 2 mDa and spectral similarity  $\geq 0.95$ ) were further processed and manually verified using vendor software *DataAnalysis* v4.3 developed by Bruker Daltonics (Bremen, Germany). The workflow aimed at achieving reliable results with high efficiency. The elucidation was based on mass accuracy, isotopic pattern, plausibility of the chromatographic retention time and HRMS/MS spectral interpretation using the comparison with spectral libraries [34]. Finally, irrelevant substances such as naturally-occurring substances (aminoacids, nucleosides and proteins etc.) were excluded from the identifications, since their presence does not induce risk to the ecosystem. It is worth mentioning that the purpose of the identification workflow was not to exhaustively identify a high number of substances but to present some examples with characteristic spatial distribution.



**Figure 3.3:** Illustration of the identification workflow, presenting an example of a successful identification.



# Chapter 4

## Results

### 4.1 Deep-learning model selection

The tested deep learning architectures were the following: a three-layer model, a CNN model, the VGG Net and ResNet architectures. The three-layer model consisted of a layer to flatten the input, a densely-connected neural network layer of 1500 units and relu activation and another densely-connected with 3 units and softmax activation. This model was regarded as baseline model due to their simplicity in architecture. The average F1 score for all three classes was 0.78. More specifically,  $F1_{Danube}$  was 0.95, 0.97, 0.96,  $F1_{Dnieper}$  was 0.95, 0.95 and 0.84 and  $F1_{Unknownsource}$  was 0.84, 0.91 and 0.54 for the three splits of the training set respectively. The three-layer model was the second-best performing model. The confusion matrices for the two best performing models are given in Figure 4.1. The best performing model was the CNN model. This model consisted of a rescaling layer with input shape the dimensions of the figure and two convolution layers with a max pooling layer in each of them. The first 2D convolution layer had 16 filters (kernel size of 3, case-insensitive padding and relu activation function) and the second 2D convolution layer had 32 filters (same kernel size, padding and activation function). The next layer was a dropout layer with rate 0.2, a layer to flatten the output, a fully-connected dense layer with 128 units on top of it that is activated by relu activation function and finally a dense layer with three output units, representing the number of classes (Danube, Dnieper, Unknown source). As in the case of the three-layer

#### 4.1 : Deep-learning model selection

model, Adam optimizer was used with sparse categorical cross entropy as loss function. The model tuned 5,613,027 weights after 45 epochs with each epoch requiring time on average 22 sec per epoch resulting in a 16.5-minute training. The average f1 score for all three classes was 0.993 ( $F1_{Danube}$  was 1.00, 0.98, 1.00;  $F1_{Dnieper}$  was 1.00, 0.99 and 0.99;  $F1_{Unknownsource}$  was 1.00, 0.97 and 0.99 for the three splits respectively). Vggnet structure [58] and ResNet [59] could not achieve satisfactory results possibly due to domain differentiation and size of the training dataset. Complex neural networks require bigger datasets to fine-tune their weights. Moreover, the approach that was followed did not involve retraining of all weights but only tuning the weights of the last layer. Transfer learning of complex architectures was abandoned given the high performance of the much simpler custom-architecture CNN.



**Figure 4.1:** Confusion matrices for the two best performing models using three different splits at the training set. The convolution neural network (CNN) proved to be the most precise with accuracy higher than 99%.

The selected CNN model proved to be accurate enough to fit its purpose and

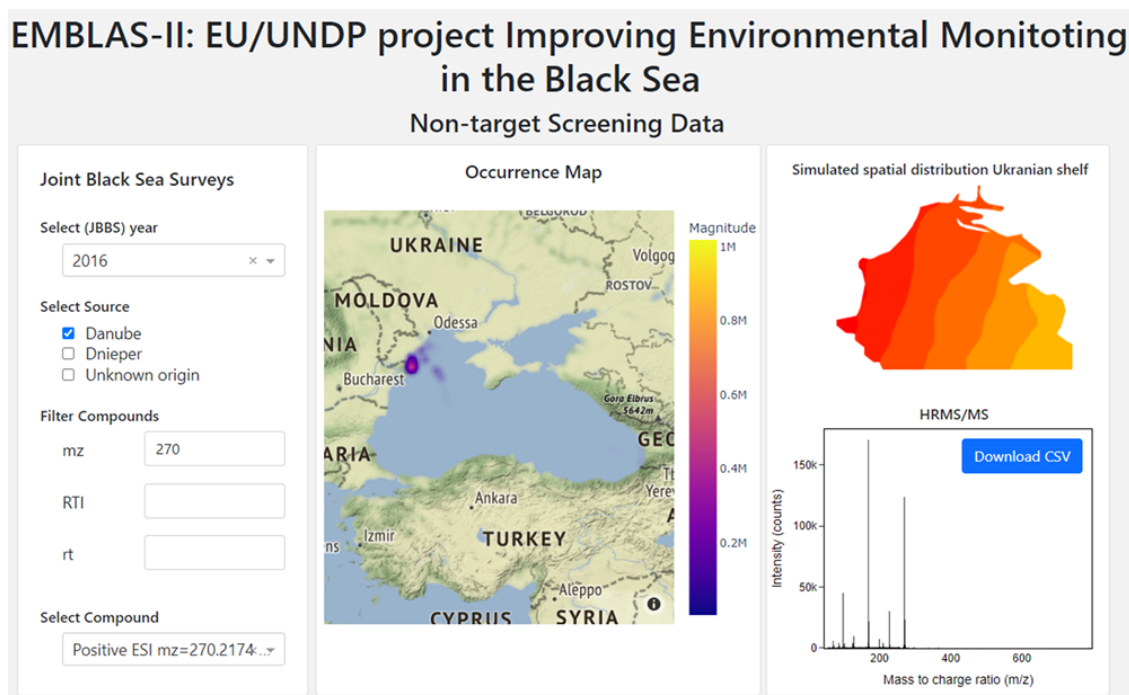
was used for allocation of the chemical signals in the spatial distribution categories. For the sampling campaign JBSS2016, 34.1% of chemical signals proved to originate from Dnieper (3893 signals), 49.7% of signals came from Danube (5686 signals), and 16.2% of chemical signals came from undefined sources (1853 signals). The results yield for JBSS2016 matched well those of the JBSS2017 campaign. 28.1% chemical signals came from Dnieper (5363 signals), 46.7% chemical signals originated from Danube (8894 signals) and 25.2% came from undefined sources (4800 signals).

Overall, more chemicals were detected in the JBSS2017 campaign comparing to JBSS2016 (11432 and 19057 in JBSS2016 and JBSS2017 respectively). In both campaigns, the majority of the detected compounds (on average 48.2%) proved to have Danube as their source. Dnieper proved to be an important source (on average 31.1%). Higher variability was observed for Dnieper percentage between the two major riverine sources. Finally, some chemicals had several different pollution sources. For these substances, different factors affect their occurrence and no clear conclusion for their source can be drawn. However, using CNN models demonstrated a clear pollution trends for more than 79% of detected chemicals. The initial hypothesis that the Danube River is one of the major pollution hotspots of the Black Sea is confirmed from the findings of this study.

## 4.2 Interactive dashboard for spatial distribution visualization

The results of the analysis were used to build an interactive interface to facilitate further interaction and exploitation of the results. The graphical user interface (GUI) of the application is shown in Figure 4.2. The dashboard was built using Dash platform, which is commonly used for producing and sharing enterprise-ready analytic application. It was proved to be of great help for the identification of the potential sources of pollution sources.

The application consists of three vertical panels (Figure 4.2). On the left panel, there is a selection module, which helps the user to select the compound of interest. The user can select the campaign of interest (JBSS2016 or JBSS2017) and filter the



**Figure 4.2:** Screenshot from the developed application. Application is available at <https://norman-data.eu/BS>

chemical signals based on the source (Danube, Dnieper or unknown origin) using the predictions from the CNN model. Moreover, the user can filter the signals based on their  $m/z$ , retention time and retention time index (RTI). RTI is the normalized retention time given the retention time of 18 recently proposed calibrant compounds [60]. The user optionally applies the desired filters. It must be noticed that it is not necessary to apply all these filters at the same time. The system will filter the signals that comply with the user filter selection and will update instantly the dropdown menu “Select Compound”. Once a specific signal is selected, the map (on the middle of Figure 4.2) will visualize its occurrence exactly as it is in the initial raw data. The signal intensity is automatically scaled in the map, so that the most intense signals are presented with purple color, while the weakest signals are presented with orange color. Moreover, the selection of a specific signal triggers the visualization of the simulated spatial distribution on the right panel. The HRMS/MS spectra is also visualized below the simulated spatial distribution, given that HRMS/MS was acquired by the instrument. A button to download the HRMS/MS as csv is available. This capability is useful, because the HRMS/MS is the fingerprint of the

structure of the substances and can be used as input in structure elucidation tools such as MetFrag [61], CSI:FingerID [62].

### 4.3 Tentatively Identified substance and their sources

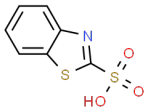

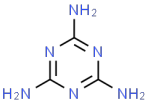

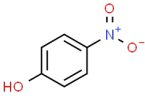

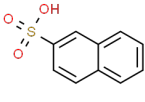

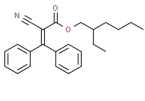

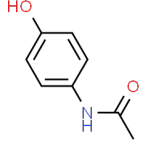

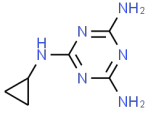

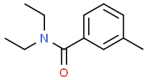

In total 35 compounds were tentatively identified (Table 4.1 and Table 4.2 at level 2A of probable structure by library spectrum match [63]. Table 1 presents some characteristic examples from each source category; substances introduced in the Black Sea by Danube river (12 compounds; ID 1 to 12), by Dnieper river (10 compounds; ID 13 to 22) and input from unknown sources (5 compounds; ID 23 to 27).

Danube river proved to be the input of many contaminants belonging in different chemical classes. Among the substances that were detected include industrial chemicals (1,2,3-benzotriazole), pharmaceuticals (metformin, carbamazepine, telmisartan, tiapride, sulpiride), antidepressant drugs (sertraline) and plant protection products (PPPs). The following three PPPs originating from Danube river were elucidated dimethenamid, terbutylazine and metolachlor. Both degraded substances and some transformation products (atenolol acid, 4-methyl-benzotriazole) were detected. The initial hypothesis that the Danube river is one of the major pollution hotspots of the Black Sea is supplemented with new findings through this analysis, which unequivocally excludes the contribution of selected compounds from additional significant pollution sources. A pollution link between the Danube and the Black Sea was verified using data from the Joint Danube Survey 4 (JDS4) organised by the ICPDR in 2019 [12]. The compounds detected in this study were also detected at the delta of the Danube river and more specifically at stations JDS50 (Reni) and JDS51 (Vilkove Chilia).

As expected, a wide variety of chemicals were proved to be introduced by Dnieper river. Compounds that were tentatively identified include industrial chemicals (triisobutyl phosphate, mono(2-ethylhexyl) phthalate), surfactants (lauryldiethanolamine, diethylene glycol monobutyl ether acetate), pharmaceuticals (repaglinide) and PPPs (isoxaben). TPs originating from the Dnieper were detected in the Ukrainian shelf. Some characteristic examples that were elucidated were iminostilbene, cyprodinil-

### 4.3 : Tentatively Identified substance and their sources

**Table 4.1:** Characteristic examples of tentatively identified compounds with riverine origin (both Danube and Dnieper).

ID	Compound	structure	Formula	RTI	Ion.	Year	Similarity (matched record)	origin
28	2-Benzothiazolesulfonic acid		C7H5NO3S2	191.7	-	2016	0.95 (SM820151)	
29	Melamine		C3H6N6	19.2	+	2017	1.00 (AU387001)	
30	4-nitrophenol		C6H5NO3	319.3	-	2016, 2017	0.96 (LU115652)	
31	2-Naphthalenesulfonic acid		C10H8O3S	190.1	-	2017	0.98 (EA065359)	
32	Octocrylene		C24H27NO2	918.2	+	2017	0.99 (AU250103)	
33	Paracetamol		C8H9NO2	138.4	+	2017	0.99 (EA024309)	
34	Cyromazine		C6H10N6	309.5	+	2016	0.99 (AU262503)	
35	DEET		C12H17NO	530.2	+	2016	0.98 (AU335302)	

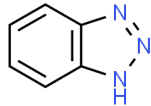

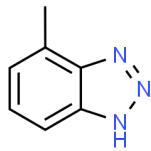

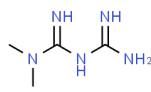

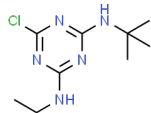

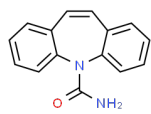

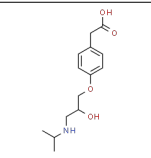

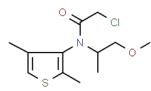

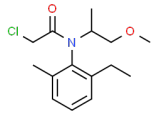

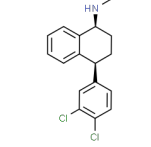

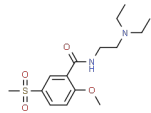

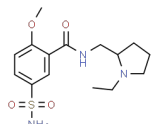

TP CGA 249287, 2-hydroxy-terbutylazine and 2-hydroxy-benzothiazole. The occurrence of these TPs indicates the consumption of the respective drug TPs (e.g. carbamazepine in the case of Iminostilbene [64]) and use of the parent PPPs in agriculture (e.g. Cyprodinil in the case of Cyprodinil-TP CGA 249287 [65]). The formation of TPs is highly dependent on the degradation rates, the flow of the river, and the climatic conditions. It is known that the flow rate of Dnieper is almost four times lower than Danube, indicating that contaminants have more time to fully degrade in Dnieper before reaching the sea.

It is worth to note that some substances have inputs from both rivers (examples presented in Table 4.1). In these cases, the origin of the chemicals in the shelf was determined based on the signal intensities. For example, when a substance yields higher signals close to the Danube delta compared to the Dnieper delta, then Danube was regarded as the major source. Some examples of substances that are introduced by both rivers in the Ukrainian shelf are the industrial chemicals (melamine, 4-nitrophenol, 2-naphthalenesulfonic acid), PPPs (cyromazine, DEET), UV filters (octocrylene) and pharmaceuticals (paracetamol).

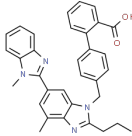
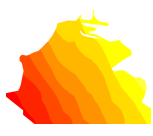
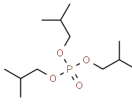

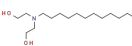

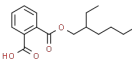

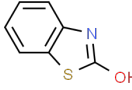

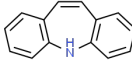

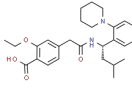

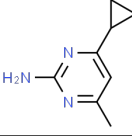

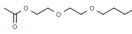

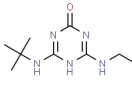

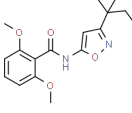

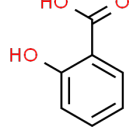

The third class that was investigated were chemicals with unknown origin. The term “unknown origin” indicates that there are unknown sources of input of these compounds in the sea or that there are multiple sources of input. The following compounds proved to have unknown origin based on the observed signals in the sampling stations: salicylic acid, caffeine, atrazine-2-hydroxy, isoproturon-didemethyl and O-Demethylmetoprolol. Salicylic acid and caffeine have multiple inputs from various anthropogenic activities (e.g. shipping activities). A reason that can drive a compound in this category is when signals are very low and thus close to the detection limits. This fact can obscure the origin of the compounds especially in the case in which a compound has been sporadically detected in some stations. Moreover, complex degradation mechanisms can also obscure the sources of the compounds. This is due to the fact that there are multiple degradation paths leading to a multitude of TPs, the signal of which is most of the times lower than the parent compound.

### 4.3 : Tentatively Identified substance and their sources

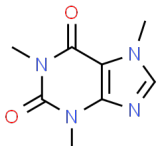

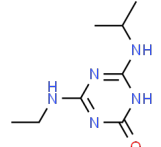

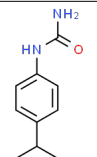

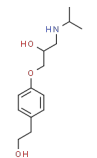

**Table 4.2:** Characteristic examples of tentatively identified compounds and their origin.

ID	Compound	structure	Formula	RTI	Ion.	Year	Similarity (matched record)	origin
1	1,2,3-Benzotriazole		C <sub>6</sub> H <sub>5</sub> N <sub>3</sub>	262.9	+	2016, 2017	0.99 (EA016662)	
2	4-Methylbenzotriazole		C <sub>7</sub> H <sub>7</sub> N <sub>3</sub>	348.4	+	2016	0.99 (AN124702)	
3	Metformin		C <sub>4</sub> H <sub>11</sub> N <sub>5</sub>	16.6	+	2016	0.95 (AU112701)	
4	Terbutylazine		C <sub>9</sub> H <sub>16</sub> ClN <sub>5</sub>	608.1	+	2016	0.99 (AU367602)	
5	Carbamazepine		C <sub>15</sub> H <sub>12</sub> N <sub>2</sub> O	474.5	+	2016	1.00 (AU112006)	
6	Atenolol acid		C <sub>14</sub> H <sub>21</sub> NO <sub>4</sub>	187.3	+	2016	0.98 (AU202506)	
7	Dimethenamid		C <sub>12</sub> H <sub>18</sub> ClNO <sub>2</sub> S	617.3	+	2016	0.98 (AU339102)	
8	Metolachlor		C <sub>15</sub> H <sub>22</sub> ClNO <sub>2</sub>	699.0	+	2016	1.00 (AU355606)	
9	Sertraline		C <sub>17</sub> H <sub>17</sub> Cl <sub>2</sub> N	581.4	+	2016	0.99 (AU150006)	
10	Tiapride		C <sub>15</sub> H <sub>24</sub> N <sub>2</sub> O <sub>4</sub> S	150.6	+	2016	0.99 (AU227803)	
11	Sulpiride		C <sub>15</sub> H <sub>23</sub> N <sub>3</sub> O <sub>4</sub> S	126.9	+	2016	0.97 (AU168502)	



ID	Compound	structure	Formula	RTI	Ion.	Year	Similarity (matched record)	origin
12	Telmisartan		C33H30N4O2	713.5	+	2016	1.00 (AU225106)	
13	Triisobutyl phosphate		C12H27O4P	768.5	+	2016	1.00 (AN118101)	
14	Lauryldiethanolamine		C16H35NO2	704.4	+	2016	0.95 (LU030801)	
15	Mono(2-ethylhexyl) phthalate (MEHP)		C16H22O4	638.5	-	2016	0.98 (LU123852)	
16	2-Hydroxybenzothiazole		C7H5NOS	399.6	+,-	2016	1.00 (AU405906)	
17	Iminostilbene		C14H11N	680.7	+	2016	0.96 (EA099209)	
18	Repaglinide		C27H36N2O4	982.2	-	2016	0.98 (EA099209)	
19	Cyprodinil-TP CGA 249287		C14H15N3O	114.7	+	2017	0.96 (EQ417302)	
20	Diethylene glycol monobutyl ether acetate		C10H20O4	492.0	+	2016	1.00 (AU502902)	
21	Terbutylazine-2-hydroxy		C9H17N5O	40.6	+	2016	0.98 (EA034702)	
22	Isoxaben		C18H24N2O4	450.8	+	2017	0.96 (EQ360202)	
23	Salicylic acid		C7H6O3	120.1	-	2016	0.98 (AU238262)	

### 4.3 : Tentatively Identified substance and their sources

ID	Compound	structure	Formula	RTI	Ion.	Year	Similarity (matched record)	origin
24	Caffeine		C <sub>8</sub> H <sub>10</sub> N <sub>4</sub> O <sub>2</sub>	214.8	+	2016	0.99 (FIO00569)	
25	Atrazine-2-hydroxy		C <sub>8</sub> H <sub>15</sub> N <sub>5</sub> O	684.5	+	2016	0.98 (SM841101)	
26	Isoproturon-didemethyl		C <sub>10</sub> H <sub>14</sub> N <sub>2</sub> O	163.6	+	2016	0.98 (EA028502)	
27	O-Demethylmetoprolol		C <sub>14</sub> H <sub>23</sub> NO <sub>3</sub>	75.0	+	2016	0.97 (ET280103)	

# Chapter 5

## Conclusions and Future Work

Sources of chemical pollution of the Ukrainian shelf of the Black Sea were investigated using a non-target screening workflow. Spatial distribution was proposed as a new prioritization approach. A novel method for the classification of 30,489 signals/-chemicals detected during the Joint Black Sea Surveys in 2016 and 2017 was applied using open-source tools and supervised machine learning. Deep learning proved to be highly accurate to build spatial distribution models. The developed workflow was able to detect and tentatively identify chemicals that reach the Ukrainian shelf from the Danube and Dnieper rivers. The CNN model enabled a detection and reliable prediction of the percentage of chemical components that clearly originate from the Danube, Dnieper and other unknown sources. The two large European rivers, Danube and Dnieper, were identified as major contributors of the chemical pollution in the northwest region of the Black Sea. Further development of prioritization methodologies and their integration in open-source workflows still remains a future goal for the non-target screening community. There is a multitude of statistical approaches and advanced visualization tools that have not yet been exploited at the selection of the most relevant chemical signals. The use of interactive applications such as the Dash application, developed in this study, is expected to play an increasing role in identification of so far unknown emerging substances pinpointed by non-target screening. The proposed non-target screening and prioritization workflow developed in this study were found as useful at detection of the sources of contaminants and their transformation products that are continuously introduced

---

into the marine ecosystem. The prioritised substances can be then subjected to in-depth structure elucidation and follow up risk assessment.

# References

- [1] K. A. Burns. Chlorinated hydrocarbons in the open mediterranean ecosystem and implications for mass balance calculations. *Mar Chem*, 20:337–359, 1987.
- [2] L. Mijangos, H. Ziarrusta, O. Ros, L. Kortazar, L. A. Fernandez, M. Olivares, O. Zuloaga, A. Prieto, and N. Etxebarria. Occurrence of emerging pollutants in estuaries of the basque country: Analysis of sources and distribution, and assessment of the environmental risk. *Water Res*, 147:152–163, 2018.
- [3] Shichun Zou, Weihai Xu, Ruijie Zhang, Jianhui Tang, Yingjun Chen, and Gan Zhang. Occurrence and distribution of antibiotics in coastal water of the bohai bay, china: Impacts of river discharge and aquaculture activities. *Environmental Pollution*, 159(10):2913–2920, 2011.
- [4] T. H. Fang, F. H. Nan, T. S. Chin, and H. M. Feng. The occurrence and distribution of pharmaceutical compounds in the effluents of a major sewage treatment plant in northern taiwan and the receiving coastal waters. *Mar Pollut Bull*, 64(7):1435–44, 2012.
- [5] Ilja Maljutenko, Ida-Maja Hassellöv, Martin Eriksson, Erik Ytreberg, Daniel Yngsell, Lasse Johansson, Jukka-Pekka Jalkanen, Mariliis Kõuts, Mari-Liis Kasemets, Jana Moldanova, Kerstin Magnusson, and Urmas Raudsepp. Modelling spatial dispersion of contaminants from shipping lanes in the baltic sea. *Marine Pollution Bulletin*, 173:112985, 2021.
- [6] K. Wille, H. Noppe, K. Verheyden, J. Vanden Bussche, E. De Wulf, P. Van Caeter, C. R. Janssen, H. F. De Brabander, and L. Vanhaecke. Validation and application of an lc-ms/ms method for the simultaneous quantification of 13 pharmaceuticals in seawater. *Anal Bioanal Chem*, 397(5):1797–808, 2010.

- [7] Xiaozhong Gao, Yuyang Lin, Juying Li, Yiping Xu, Zhengfang Qian, and Wenjie Lin. Spatial pattern analysis reveals multiple sources of organophosphorus flame retardants in coastal waters. *Journal of Hazardous Materials*, 417:125882, 2021.
- [8] M. Biel-Maeso, R. M. Baena-Nogueras, C. Corada-Fernandez, and P. A. Lara-Martin. Occurrence, distribution and environmental risk of pharmaceutically active compounds (phacs) in coastal and ocean waters from the gulf of cadiz (sw spain). *Sci Total Environ*, 612:649–659, 2018.
- [9] F. Hernandez, N. Calisto-Ulloa, C. Gomez-Fuentes, M. Gomez, J. Ferrer, G. Gonzalez-Rocha, H. Bello-Toledo, A. M. Botero-Coy, C. Boix, M. Ibanez, and M. Montory. Occurrence of antibiotics and bacterial resistance in wastewater and sea water from the antarctic. *J Hazard Mater*, 363:447–456, 2019.
- [10] H. Zhao, J. L. Zhou, and J. Zhang. Tidal impact on the dynamic behavior of dissolved pharmaceuticals in the yangtze estuary, china. *Sci Total Environ*, 536:946–954, 2015.
- [11] K. Fisch, J. J. Waniek, and D. E. Schulz-Bull. Occurrence of pharmaceuticals and uv-filters in riverine run-offs and waters of the german baltic sea. *Mar Pollut Bull*, 124(1):388–399, 2017.
- [12] I. Liška, F. Wagner, M. Sengl, K. Deutsch, J. Slobodník, and M. Paunovic. Joint danube survey 4 scientific report: A shared analysis of the danube river. international commission for the protection of the danube river, vienna. isbn: 978-3-200-07450-7., 2021.
- [13] J. Slobodnik, N. Alygizakis, and P. Oswald. Investigative monitoring of the dnierper river basin (available at [https://euneighbourseast.eu/wp-content/uploads/2021/07/ua\\_dnieper\\_river\\_basin\\_screening\\_final\\_report\\_eng.pdf](https://euneighbourseast.eu/wp-content/uploads/2021/07/ua_dnieper_river_basin_screening_final_report_eng.pdf), lastaccess5march2022), 2021.
- [14] Konstantina S. Diamanti, Nikiforos A. Alygizakis, Maria-Christina Nika, Martina Oswaldova, Peter Oswald, Nikolaos S. Thomaidis, and Jaroslav Slobodnik. Assessment of the chemical pollution status of the dniester river basin by wide-scope

- 
- target and suspect screening using mass spectrometric techniques. *Analytical and Bioanalytical Chemistry*, 412(20):4893–4907, 2020.
- [15] J. Du, H. Zhao, S. Liu, H. Xie, Y. Wang, and J. Chen. Antibiotics in the coastal water of the south yellow sea in china: Occurrence, distribution and ecological risks. *Sci Total Environ*, 595:521–527, 2017.
- [16] J. Magner, M. Filipovic, and T. Alsberg. Application of a novel solid-phase-extraction sampler and ultra-performance liquid chromatography quadrupole-time-of-flight mass spectrometry for determination of pharmaceutical residues in surface sea water. *Chemosphere*, 80(11):1255–60, 2010.
- [17] A. M. Ali, H. T. Ronning, W. Alarif, R. Kallenborn, and S. S. Al-Lihaibi. Occurrence of pharmaceuticals and personal care products in effluent-dominated saudi arabian coastal waters of the red sea. *Chemosphere*, 175:505–513, 2017.
- [18] S. Bayen, H. Zhang, M. M. Desai, S. K. Ooi, and B. C. Kelly. Occurrence and distribution of pharmaceutically active and endocrine disrupting compounds in singapore’s marine environment: influence of hydrodynamics and physical-chemical properties. *Environ Pollut*, 182:1–8, 2013.
- [19] G. F. Birch, D. S. Drage, K. Thompson, G. Eaglesham, and J. F. Mueller. Emerging contaminants (pharmaceuticals, personal care products, a food additive and pesticides) in waters of sydney estuary, australia. *Mar Pollut Bull*, 97(1-2):56–66, 2015.
- [20] J. J. Jiang, C. L. Lee, and M. D. Fang. Emerging organic contaminants in coastal waters: anthropogenic impact, environmental release and ecological risk. *Mar Pollut Bull*, 85(2):391–9, 2014.
- [21] H. Y. Kim, I. S. Lee, and J. E. Oh. Human and veterinary pharmaceuticals in the marine environment including fish farms in korea. *Sci Total Environ*, 579:940–949, 2017.
- [22] A. Lolic, P. Paiga, L. H. Santos, S. Ramos, M. Correia, and C. Delerue-Matos. Assessment of non-steroidal anti-inflammatory and analgesic pharmaceuticals in sea-waters of north of portugal: occurrence and environmental risk. *Sci Total Environ*, 508:240–50, 2015.

- [23] F. Vanryckeghem, S. Huysman, H. Van Langenhove, L. Vanhaecke, and K. Demeestere. Multi-residue quantification and screening of emerging organic micropollutants in the belgian part of the north sea by use of speedisk extraction and q-orbitrap hrms. *Mar Pollut Bull*, 142:350–360, 2019.
- [24] T. H. Fang, C. W. Lin, and C. H. Kao. Occurrence and distribution of pharmaceutical compounds in the danshuei river estuary and the northern taiwan strait. *Mar Pollut Bull*, 146:509–520, 2019.
- [25] P. A. Lara-Martin, E. Gonzalez-Mazo, M. Petrovic, D. Barcelo, and B. J. Brownawell. Occurrence, distribution and partitioning of nonionic surfactants and pharmaceuticals in the urbanized long island sound estuary (ny). *Mar Pollut Bull*, 85(2):710–9, 2014.
- [26] Y. Cui, Y. Wang, C. Pan, R. Li, R. Xue, J. Guo, and R. Zhang. Spatiotemporal distributions, source apportionment and potential risks of 15 pharmaceuticals and personal care products (ppcps) in qinzhou bay, south china. *Mar Pollut Bull*, 141:104–111, 2019.
- [27] Sang-Soo Baek, Daeun Yun, JongCheol Pyo, Daeho Kang, Kyung Hwa Cho, and Junho Jeon. Analysis of micropollutants in a marine outfall using network analysis and decision tree. *Science of The Total Environment*, 806:150938, 2022.
- [28] L. Vergeynst, H. Van Langenhove, P. Joos, and K. Demeestere. Suspect screening and target quantification of multi-class pharmaceuticals in surface water based on large-volume injection liquid chromatography and time-of-flight mass spectrometry. *Anal Bioanal Chem*, 406(11):2533–47, 2014.
- [29] C. Moschet, A. Piazzoli, H. Singer, and J. Hollender. Alleviating the reference standard dilemma using a systematic exact mass suspect screening approach with liquid chromatography-high resolution mass spectrometry. *Anal Chem*, 85(21):10312–20, 2013.
- [30] Cathrin Veenas, Anders Bignert, Per Liljelind, and Peter Haglund. Nontarget screening and time-trend analysis of sewage sludge contaminants via two-dimensional gas chromatography–high resolution mass spectrometry. *Environmental Science & Technology*, 52(14):7813–7822, 2018.



- 
- [31] Jennifer E. Schollée, Emma L. Schymanski, Michael A. Stravs, Rebekka Gulde, Nikolaos S. Thomaidis, and Juliane Hollender. Similarity of high-resolution tandem mass spectrometry spectra of structurally related micropollutants and transformation products. *Journal of the American Society for Mass Spectrometry*, 28(12):2692–2704, 2017.
- [32] Mark Strynar, Sonia Dagnino, Rebecca McMahan, Shuang Liang, Andrew Lindstrom, Erik Andersen, Larry McMillan, Michael Thurman, Imma Ferrer, and Carol Ball. Identification of novel perfluoroalkyl ether carboxylic acids (pfecas) and sulfonic acids (pfesas) in natural waters using accurate mass time-of-flight mass spectrometry (tofms). *Environmental Science & Technology*, 49(19):11622–11630, 2015.
- [33] Parvaneh Hajeb, Linyan Zhu, Rossana Bossi, and Katrin Vorkamp. Sample preparation techniques for suspect and non-target screening of emerging contaminants. *Chemosphere*, 287:132306, 2022.
- [34] Pablo Gago-Ferrero, Emma L. Schymanski, Anna A. Bletsou, Reza Aalizadeh, Juliane Hollender, and Nikolaos S. Thomaidis. Extended suspect and non-target strategies to characterize emerging polar organic contaminants in raw wastewater with lc-hrms/ms. *Environmental Science & Technology*, 49(20):12333–12341, 2015.
- [35] Andrea M. Brunner, Milou M. L. Dingemans, Kirsten A. Baken, and Annemarie P. van Wezel. Prioritizing anthropogenic chemicals in drinking water and sources through combined use of mass spectrometry and toxcast toxicity data. *Journal of Hazardous Materials*, 364:332–338, 2019.
- [36] Martin Krauss, Christine Hug, Robert Bloch, Tobias Schulze, and Werner Brack. Prioritising site-specific micropollutants in surface water from lc-hrms non-target screening data using a rarity score. *Environmental Sciences Europe*, 31(1):45, 2019.
- [37] Matthias Ruff, Miriam S. Mueller, Martin Loos, and Heinz P. Singer. Quantitative target and systematic non-target analysis of polar organic micro-pollutants along the river rhine using high-resolution mass-spectrometry – identification of unknown sources and compounds. *Water Research*, 87:145–154, 2015.
- [38] Tom Mitchell. *Machine learning*. McGraw-Hill Education, New York, USA, 1 edition, 1994.
-

- [39] Kevin Murphy. *Machine learning: a probabilistic perspective*. MIT press, 1 edition, 2012.
- [40] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3 edition, 2010.
- [41] Hinton Geoffrey and Terrence Sejnowski. *Unsupervised Learning: Foundations of Neural Computation*. MIT Press, 1 edition, 1999.
- [42] Richard S. Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT Press, Cambridge, London, 2 edition, 2018.
- [43] Martijn van Otterlo and Marco Wiering. *Reinforcement learning and markov decision processes. Reinforcement Learning. Adaptation, Learning, and Optimization*. Springer, Berlin, Heidelberg, 2012.
- [44] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.
- [45] Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.
- [46] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. *In Proceedings of the fourteenth international conference on artificial intelligence and statistics*, page 315, 2011.
- [47] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
- [48] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.
- [49] G. Thomas Dietterich and Eun Bae Kong. *Machine Learning Bias, Statistical Bias, and Statistical Variance of Decision Tree Algorithms*. Oregon State University, 1995. <http://citeseerx.ist.psu.edu/viewdoc/download?rep=rep1&type=pdf&doi=10.1.1.38.2702>.
- [50] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

- 
- [51] Colin A. Smith, Elizabeth J. Want, Grace O’Maille, Ruben Abagyan, and Gary Siuzdak. Xcms: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry*, 78(3):779–787, 2006.
- [52] Carsten Kuhl, Ralf Tautenhahn, Christoph Böttcher, Tony R. Larson, and Steffen Neumann. Camera: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Analytical Chemistry*, 84(1):283–289, 2012.
- [53] Gunnar Libiseller, Michaela Dvorzak, Ulrike Kleb, Edgar Gander, Tobias Eisenberg, Frank Madeo, Steffen Neumann, Gert Trausinger, Frank Sinner, Thomas Pieber, and Christoph Magnes. Ipo: a tool for automated optimization of xcms parameters. *BMC Bioinformatics*, 16(1):118, 2015.
- [54] N. A. Alygizakis, P. Gago-Ferrero, J. Hollender, and N. S. Thomaidis. Untargeted time-pattern analysis of lc-hrms data to detect spills and compounds with high fluctuation in influent wastewater. *J Hazard Mater*, 361:19–29, 2019.
- [55] George Tserpes, Chrissi-Yianna Politou, Panagiota Peristeraki, Argyris Kallianiotis, and Costas Papaconstantinou. Identification of hake distribution pattern and nursery grounds in the hellenic seas by means of generalized additive models. *Hydrobiologia*, 612(1):125–133, 2008.
- [56] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, R. Jozefowicz, Y. Jia, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, M. Schuster, R. Monga, S. Moore, D. Murray, C. Olah, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. Large-scale machine learning on heterogeneous systems. software available from tensorflow.org. 2015.
- [57] Martin Loos, Christian Gerber, Francesco Corona, Juliane Hollender, and Heinz Singer. Accelerated isotope fine structure calculation using pruned transition trees. *Analytical Chemistry*, 87(11):5738–5744, 2015.
-

- [58] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition (<https://doi.org/10.48550/arxiv.1409.1556>). *ICLR 2015*, 2015.
- [59] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015.
- [60] R. Aalizadeh, N. A. Alygizakis, E. L. Schymanski, M. Krauss, T. Schulze, M. Ibanez, A. D. McEachran, A. Chao, A. J. Williams, P. Gago-Ferrero, A. Covaci, C. Moschet, T. M. Young, J. Hollender, J. Slobodnik, and N. S. Thomaidis. Development and application of liquid chromatographic retention time indices in hrms-based suspect and nontarget screening. *Anal Chem*, 93(33):11601–11611, 2021.
- [61] Christoph Ruttkies, Emma L. Schymanski, Sebastian Wolf, Juliane Hollender, and Steffen Neumann. Metfrag relaunched: incorporating strategies beyond in silico fragmentation. *Journal of Cheminformatics*, 8(1):3, 2016.
- [62] Kai Dührkop, Huibin Shen, Marvin Meusel, Juho Rousu, and Sebastian Böcker. Searching molecular structure databases with tandem mass spectra using csi:fingerid. *Proceedings of the National Academy of Sciences*, 112(41):12580–12585, 2015.
- [63] Emma L. Schymanski, Junho Jeon, Rebekka Gulde, Kathrin Fenner, Matthias Ruff, Heinz P. Singer, and Juliane Hollender. Identifying small molecules via high resolution mass spectrometry: Communicating confidence. *Environmental Science & Technology*, 48(4):2097–2098, 2014.
- [64] A. B. Martínez-Piernas, S. Nahim-Granados, M. I. Polo-López, P. Fernández-Ibáñez, S. Murgolo, G. Mascolo, and A. Agüera. Identification of transformation products of carbamazepine in lettuce crops irrigated with ultraviolet-c treated water. *Environmental Pollution*, 247:1009–1019, 2019.
- [65] Karin Kiefer. *Polar Micropollutants and their Transformation Products in Groundwater: Identification with LC-HRMS and their Abatement in Water Treatment*. Thesis, 2021.