



**HAL**  
open science

# Enhanced database creation with in silico workflows for suspect screening of unknown tebuconazole transformation products in environmental samples by UHPLC-HRMS

Kevin Rocco, Christelle Margoum, Loïc Richard, Marina Coquery

## ► To cite this version:

Kevin Rocco, Christelle Margoum, Loïc Richard, Marina Coquery. Enhanced database creation with in silico workflows for suspect screening of unknown tebuconazole transformation products in environmental samples by UHPLC-HRMS. *Journal of Hazardous Materials*, 2022, 440 (129706), pp.1-21. 10.1016/j.jhazmat.2022.129706 . hal-04172394

**HAL Id: hal-04172394**

**<https://hal.inrae.fr/hal-04172394>**

Submitted on 27 Jul 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# 1 Enhanced database creation with *in silico* workflows for 2 suspect screening of unknown tebuconazole transformation 3 products in environmental samples by UHPLC-HRMS

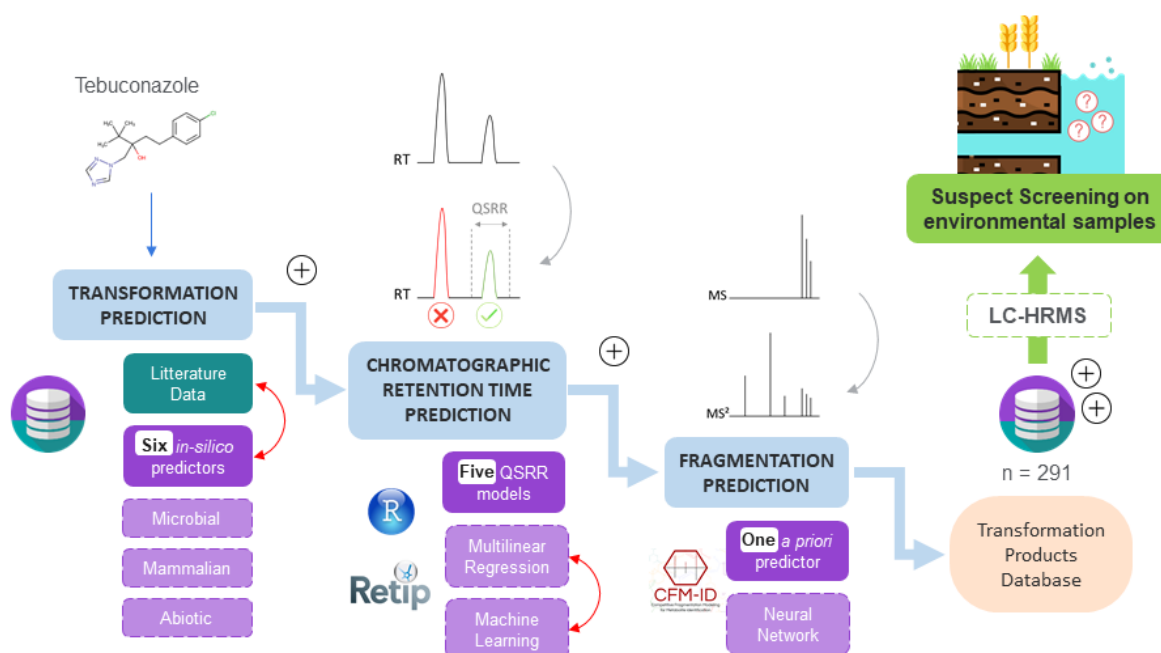
4 Kevin Rocco\*, Christelle Margoum, Loïc Richard, Marina Coquery

5 INRAE, UR RiverLy, 69625 Villeurbanne, France

6 \*Corresponding author: Phone: +33(0)4 72 20 11 06; E-mail: [kevin.rocco@inrae.fr](mailto:kevin.rocco@inrae.fr)

7 \*AUTHOR VERSION\* – Please refer to <https://doi.org/10.1016/j.jhazmat.2022.129706>

## 8 GRAPHICAL ABSTRACT



9

## 10 HIGHLIGHTS

- 11 . A suspect database of 291 TPs of tebuconazole was created.
- 12 . Twelve cutting-edge *in silico* predictors were used and compared.
- 13 . RT and *a priori* fragmentation predictions were conducted on predicted TPs.
- 14 . Comparison of prediction from transformation predictors revealed the known TPs.
- 15 . Workflow-aided retrospective analysis of surface-water samples highlighted new TPs.

16 **ABSTRACT**

17 The search and identification of organic contaminants in agricultural watersheds has become a crucial  
18 effort to better characterize watershed contamination by pesticides. The past decade has brought a  
19 more holistic view of watershed contamination via the deployment of powerful analytical strategies  
20 such as non-target and suspect screening analysis that can search more contaminants and their  
21 transformation products. However, suspect screening analysis remains broadly confined to known  
22 molecules, primarily due to the lack of analytical standards and suspect databases for unknowns such  
23 as pesticide transformation products. Here we developed a novel workflow by cross-comparing the  
24 results of various *in silico* prediction tools against literature data to create an enhanced database for  
25 suspect screening of pesticide transformation products. This workflow was applied on tebuconazole,  
26 used here as a model pesticide, and resulted in a suspect screening database counting 291  
27 transformation products. The chromatographic retention times and tandem mass spectra were  
28 predicted for each of these compounds using 6 models based on multilinear regression and more  
29 complex machine-learning algorithms. This comprehensive approach to the investigation and  
30 identification of tebuconazole transformation products was retrospectively applied on environmental  
31 samples and found 6 transformation products identified for the first time in river water samples.

32

---

33 **KEYWORDS**

34 pesticides; metabolites; computational tools; suspect screening analysis; biotic degradation

35

---

36 **ENVIRONMENTAL IMPLICATION**

37 The *in silico* workflow presented in our work represents an improvement in the suspect screening of  
38 transformation products, which are undeniable ubiquitous environmentally hazardous contaminants.  
39 Applied on the fungicide tebuconazole as a model compound, the workflow led to the detection of  
40 seven new transformation products in surface waters. Based on accessible and transposable *in silico*  
41 tools, the proposed workflow can be replicated to a wide range of organic substances and reused by  
42 other environmental analysis laboratories. We therefore believe in the relevance of publishing our  
43 work in Journal of Hazardous Material.

## 44 1. Introduction

45 Pesticides are chemical compounds used mainly in agriculture to control plant pests and  
46 improve crop yields. Once in the environment, pesticides can be degraded into transformation  
47 products (TPs) via both biotic and abiotic transformation processes [1, 2]. The chemical compounds  
48 formed by these transformations processes are generally lower, more persistent in the environment  
49 and more mobile than the parent compound, which can increase their transport to surface water and  
50 groundwater by runoff or seepage from agricultural soils [3, 4]. As a rule, these structural and property  
51 changes do not specifically increase the toxicity of TPs compared to parent compounds. However,  
52 within the multitude of products formed, some may be exceptions to this rule, which makes it  
53 important to identify them [2]. This blind-spot in identification means that the toxicity of pesticides  
54 and their TPs in water bodies is globally underestimated [5, 6]. Novel approaches are needed in order  
55 to identify these unknown TPs compounds.

56 The simultaneous quantification of pesticides and their known TPs in waterbodies has revealed  
57 the presence of TPs at higher levels of concentration and occurrence than their parent compounds. As  
58 an example, in headwater streams, Le Cor et al. [7] highlighted that pesticide TPs accounted for more  
59 than half of the substances detected and that TP concentrations were often ten times higher than the  
60 parent-compound concentrations ( $0.46 \pm 0.02 \mu\text{g/L}$  for the TP metazachlor-ESA *versus*  $0.047 \pm 0.007$   
61  $\mu\text{g/L}$  for the parent metazachlor). However, such targeted analyses are limited by the lack of standards  
62 for most pesticide TPs. To overcome this gap, powerful techniques such as high-resolution mass  
63 spectrometry (HRMS) have been developed over the last decade. Gas chromatography (GC) or liquid  
64 chromatography (LC) coupled with HRMS can serve to develop suspect and non-target screening (NTS)  
65 strategies that bring a more holistic understanding of the environmental fate of organic chemicals by  
66 untangling the unknowns [8].

67 Suspect screening strategies involve comparing key characteristics of compounds, compiled in  
68 a database (DB), to analytical data on actual environmental samples acquired by HRMS. The minimum  
69 data required to suspect a compound in a water sample is the exact mass of the compounds of interest.  
70 Levels of confidence in suspected presence can be increased with additional compound-related data  
71 such as mass fragmentation patterns (MS/MS spectra) and chromatographic retention times (RT) [9].  
72 This additional data is usually obtained by injecting analytical standards into a LC or GC-HRMS  
73 instrument or is already contained in commercial or public databases, such as the NORMAN Suspect  
74 List Exchange (<https://www.norman-network.com/nds/SLE/>). However, when analytical standards  
75 and databases are unavailable, analysts should consider using extensive suspect screening with  
76 enhanced databases built from *in silico* prediction tools. Recent developments in extensive suspect  
77 screening for pesticide TPs within water bodies has made it possible to identify many new focal  
78 compounds [10, 11], which underscores the value of creating improved databases for suspect  
79 screening analysis.

80 *In silico* tools are defined here as commercially or freely-available software or web platforms  
81 that use sophisticated algorithms to perform predictive tasks that would be too time-consuming or  
82 even impossible for a human to perform. The practicality of such *in silico* tools stems from their ability  
83 to predict compound properties solely from their chemical identifiers—as with the simplified  
84 molecular-input line-entry specification; SMILES—, thus overcoming the need for analytical standards.

85 Some *in silico* tools, called transformation predictors, can predict the formation of possible TPs  
86 by using the chemical identifiers of the parent compound as an input. These tools are based on various  
87 pre-established physicochemical reactions that can occur in various environmental compartments (e.g.  
88 aquatic, terrestrial or biological) via both abiotic and biotic transformation processes on scales running

89 from microbial up to mammalian metabolism. The appropriate transformation predictor has to be  
90 selected based on the environmental degradation processes investigated. TPs predicted by these  
91 transformation predictors carry a relatively high rate of false-positives, but some predictors can use  
92 relative reasoning to address this issue [12]. The efficiency of these tools has already been proven. For  
93 instance, Jiao et al. [13] recently detected 14 new TPs of the fungicide pyrisoxazole using literature  
94 data and one *in silico* tool, Envipath [14], for database construction.

95 Another important subset of *in silico* tools are chromatographic RT prediction tools, which are  
96 usually based on quantitative structure–activity relationship (QSAR) models principles, extended to so-  
97 called quantitative structure–retention relationship (QSRR) models. Predictions are made based on the  
98 assumption that there are relationships between the chemical structures of the compounds and their  
99 chromatographic RTs. These prediction tools are developed from predicted or experimental molecular  
100 descriptors—which are associated with experimental chromatographic RTs—of a group of compounds.  
101 This group is generally split into two: one called the “training set” that establishes the relationship  
102 between molecular descriptors and chromatographic RT, and the other called the “testing set” that is  
103 used for validation. This group can also be divided into three, with an addition to the training and  
104 testing set of a “validation set”, which deals with any overfitting produced during the QSRR  
105 construction [15]. The complexity of these QSRR models varies according to the amount and type of  
106 molecular descriptors required to build them, but also depending on the algorithms establishing the  
107 relationships, from multiple linear regression (MLR) to non-linear machine-learning (ML)-based QSRR.  
108 Taking into account the range of prediction error given by the QSRR model, the predicted  
109 chromatographic RTs can serve to eliminate outliers during suspect screening [16].

110 Other *in silico* tools can be used to annotate acquired MS/MS spectra *a posteriori*, such as  
111 SIRIUS [17], MAGMA [18] or MetFrag [19], in order to identify compounds or at least increase their  
112 confidence in detection during suspect and non-target analysis [11, 20]. A complementary approach  
113 consists of predicting MS/MS spectra before analytical acquisition (i.e. *a priori*) in order to enhance the  
114 suspect compounds database. This can be done with fragmentation predictors like competitive  
115 fragmentation modeling-ID (CFM-ID) that employ neural network algorithms for *a priori* prediction of  
116 MS/MS spectra based solely on SMILES compounds as an input [21, 22]. This addition of predicted  
117 MS/MS spectra strengthens the identification performance and limits compound mismatches during  
118 suspect screening analysis.

119 With that vision, a solution to better characterize water-body contamination by pesticide TPs  
120 could be to combine a selected set of these *in silico* tools, which are often used alone but, to our  
121 knowledge, have never been grouped into a comprehensive workflow. Here we address this gap by  
122 developing a comprehensive workflow for the creation of detailed databases for suspect screening of  
123 unknown compounds such as pesticide TPs in agricultural watersheds. Each step of this workflow  
124 allows the prediction of specific information about the TP compounds, such as their identity,  
125 chromatographic RT, and fragmentation spectra. The novelty of this approach is that it uses several *in*  
126 *silico* prediction tools based on innovative algorithms and cross-compares them together and against  
127 literature data. In addition to being easily transferable to other compounds or analytical conditions,  
128 this approach provides an enhanced ready-to-use database of a pesticide’s TPs for suspect screening  
129 analysis on environmental samples.

130

## 131 2. Materials and methods

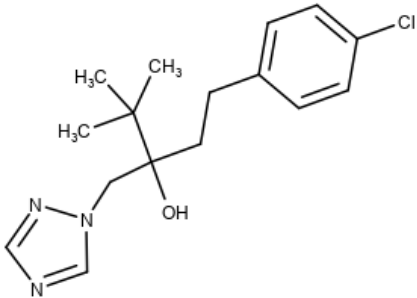
### 132 2.1. Experimental

#### 133 2.1.1 Pesticide selection

134 To demonstrate the potential of using a combination of *in silico* tools to create a suspect  
135 screening database of TPs, the triazole fungicide tebuconazole (TBZ) was used as a model compound.  
136 The main characteristics of this compound are presented in Table 1.

137

138 **Table 1.** Main chemical identifiers and environmental behavior of tebuconazole.

Structure		Compound name	
		1-(4-chlorophenyl)-4,4-dimethyl-3-(1,2,4-triazol-1-ylmethyl)pentan-3-ol	
		SMILES	
		<chem>Clc1ccc(cc1)CCC(O)(C(C)(C)C)Cn2ncnc2</chem>	
		InChiKey	
PXMNMQRDXWABCY-UHFFFAOYSA-N			
<b>DT50<sub>Soil</sub></b> (EFSA 2014)	19.9–91.6 days	<b>Formula</b>	C <sub>16</sub> H <sub>22</sub> ClN <sub>3</sub> O
<b>DT50<sub>Water – pH7</sub></b> (EFSA 2014)	590 days	<b>Mass (g.mol<sup>-1</sup>)</b>	307.8180

139

140 TBZ was selected primarily because it is one of the best-selling fungicides in the world and it  
141 has been applied for over twenty years in Europe due to its broad-spectrum activity [23, 24]. Moreover,  
142 the formation of TBZ TPs in the soil matrix has been extensively studied, mainly through the EU-funded  
143 Love-to-Hate project between 2013 and 2016 (<http://lovetohate.bio.uth.gr>). Over the course of this  
144 project, a series of analytical developments were carried out in order to identify the TPs of TBZ under  
145 laboratory [25] and field [26] exposure conditions. Furthermore, recent studies have shown that TBZ  
146 is one of the most frequently detected fungicides in surface waters worldwide [27], and some of its  
147 TPs have been identified *in situ* [28].

148

#### 149 2.1.2. Instrumentation

150 The analytical conditions used to construct the chromatographic RT prediction models and  
151 acquire the compound spectra are detailed elsewhere in Bride et al. [29]. Briefly, the conditions used  
152 consists in a chromatographic separation on a LC system (ACQUITY UPLC H-Class system, Waters) with  
153 a 100 mm × 2.1 mm, 1.8- $\mu$ m Acquity HSS T3 column (Waters, Milford, MA) at 30°C. The LC analyses  
154 were performed at a flowrate of 0.5 mL/min using water + 0.1% formic acid (A) and acetonitrile + 0.1%  
155 formic acid (B) as mobile phases. The gradient program consisted of an initial hold for 2 min at 2% B,  
156 followed by a linear gradient up to 99% B in 13 min, a hold for 2 min at 99% B, then a decrease from

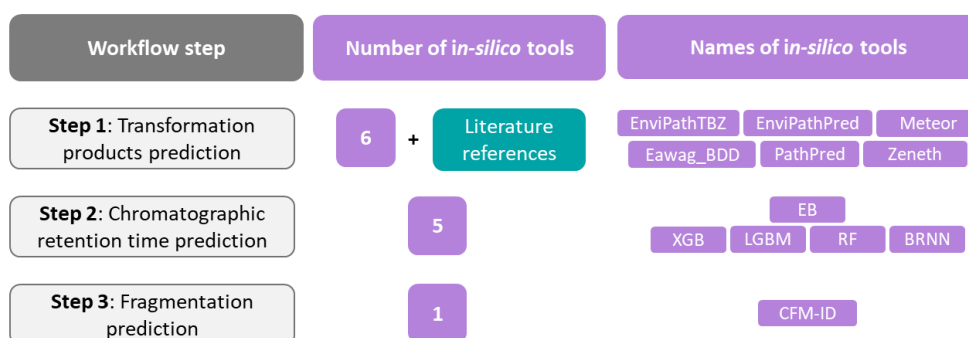
157 99% to 2% B in 1 min, and a final hold for 2 min at 2% B. This separation was completed by detection  
 158 with an Xevo G2-S (Waters) quadrupole time-of-flight (QToF) mass spectrometer. The QToF systems  
 159 was operated in MS<sup>E</sup> data-independent acquisition (DIA) mode (*i.e.* all ions simultaneously  
 160 fragmented) with an energy ramp of 10–45eV and a mass acquisition range of 50–1200 *m/z*.

161

## 162 2.2. Database creation for pesticide transformation products

163 The different steps of the workflow developed to create the database are detailed in this  
 164 section and schematized in Figure 1. The first step in this workflow is to implement the TPs to be  
 165 searched within the database. This step uses 6 *in silico* tools, defined as ‘transformation predictors’, in  
 166 order to predict the transformation of the parent compound into its TPs. As described in Figure 1, a  
 167 thorough literature review was performed to complement the TPs prediction implemented using *in*  
 168 *silico* transformation predictors. This literature search was performed on January 2021, on the Web-  
 169 of-Science and Scopus platforms using the search terms “tebuconazole AND transformation product\*”  
 170 or “tebuconazole AND metabolite\*”. The majority of the compounds listed by this search are from  
 171 publications derived from the Love-to-Hate project [25, 26]. All the TPs resulting from this literature  
 172 search were incorporated into our database under the term “*in biblio*” TPs in contrast to the “*in silico*”  
 173 predicted TPs. The second step in this workflow uses five *in-silico* tools, defined as QSRR, to predict the  
 174 chromatographic RTs of TPs. The third step in the workflow mobilizes a fragmentation predictor to  
 175 predict high-resolution tandem mass spectra.

176



177 **Figure 1.** Overview of the database creation workflow, including the numbers and names of *in silico*  
 178 tools used for the three workflows steps. The acronyms used for the *in silico* tools are spelled out in  
 179 section 2.2.

180

### 181 2.2.1. Step 1: Prediction of tebuconazole transformation products using transformation 182 predictors

183 We used 6 transformation predictors to predict TPs of TBZ: EnviPath, in its ‘EnviPathTBZ’ and  
 184 ‘EnviPathPred’ versions, plus ‘Meteor’, ‘Eawag\_BDD’, ‘PathPred’, and ‘Zeneth’ (Figure 1). Due to the  
 185 high chemical stability of TBZ in water (Table 1), most of the transformation predictors used are based  
 186 on degradation processes driven by microbial metabolism. Certain other transformation predictors are  
 187 used to predict abiotic hydrolysis and reduction, such as the ‘chemical transformation simulator’ (CTS)

188 [30]. This transformation predictors were not included in this study as they were ineffective in their  
189 prediction output, producing small numbers of irrelevant TPs.

190 Envipath is a transformation predictor for the microbial biotransformation of compounds that  
191 proposes a “store-and-view” system of experimentally-observed biotransformation pathways [14]. In  
192 the present study, the model includes two *in silico* transformation predictors: i) ‘EnvipathPred’, which  
193 results from the prediction of TBZ degradation by Envipath, and ii) ‘EnvipathTBZ’, which is a  
194 prerecorded TBZ degradation pathway stored within the platform.

195 The University of Minnesota Pathway Prediction System (UM-PPS, named ‘Eawag\_BDD’ in this  
196 study), which is hosted on the Eawag website (<http://eawag-bbd.ethz.ch/predict/>), predicts microbial  
197 catabolic reactions using substructure searching, a rule-base, and atom-to-atom compound mapping  
198 [31].

199 PathPred is a transformation predictor, hosted on the GenomeNet website, that predicts  
200 plausible biodegradation pathways of compounds based on enzyme-catalyzed reactions [32].

201 To complement these four transformation predictors that are based on microbial  
202 metabolisms, we used two other transformation predictors: Meteor Nexus [33] and Zeneth [34].  
203 Meteor Nexus is based on mammalian biotransformation reactions, while Zeneth is based on forced  
204 degradation pathways of compounds under various abiotic conditions (temperature, aerobic or  
205 anaerobic, with or without metal presence, or exposure to light). These transformation predictors  
206 were mobilized here to provide a more holistic picture of the range of TPs that can form in the  
207 environment. These two transformation predictors are the only *in silico* tools used in this study that  
208 are not freely-available.

209 The inputs needed for all these transformation predictors are the chemical identifiers of the  
210 parent compounds, such as SMILES, but the output format depends on the transformation predictor.  
211 OpenBabel (V2.4.1) was used to convert chemical identifiers (i.e. from .mol or SMILES to InChi) in order  
212 to harmonize the output and allow comparison of results between the 6 transformation predictors.  
213 The comparison between predicted TPs was done on InChiKey, a short-coded, compound-specific, one-  
214 way readable chemical identifier ([http://inchi.info/inchikey\\_overview\\_en.html](http://inchi.info/inchikey_overview_en.html)).

215

### 216 2.2.2. Step 2: Chromatographic retention time prediction by QSRR models

217 For step 2 of the workflow, two types of QSRR models were used for RT prediction: a QSRR  
218 model based on multiple linear regression (MLR), and four models based on machine-learning (ML)  
219 algorithms.

220 More information about the MLR-based QSRR model used can be found in Bride et al. [29].  
221 Briefly, this model (named ‘EB’ here) was built from 8 molecular descriptors selected for their  
222 relevance-for-purpose in LC (MW, logD, DBE, nbO, nbC, nbH, HBdD, logSw - described in  
223 Supplementary data, Excel spreadsheet #1), using 273 experimental chromatographic retention time  
224 (ERT). The ERTs were split into a training set and a testing set at a 65:35 ratio (training set size: 204  
225 ERTs, testing set size: 69 ERTs). This EB model enables chromatographic RT prediction within a range  
226 of  $\pm 1.96$  min (at 95% confidence intervals) for a 20-minutes chromatographic run. The prediction of  
227 the molecular descriptors used by the model is not automated.

228 The Retip package (v0.5.4.) [35] in R (v4.0.4) was used to build the ML-based QSRR models.  
229 The models created were based on the same training set as the MLR-based QSRR named ‘EB’ to



230 facilitate cross-comparison (experimental compounds used in training or testing are listed in  
231 Supplementary data, Excel spreadsheet #2). The molecular descriptors for each analytical standard  
232 were predicted using the RCDK (v3.5.0.) package. As their prediction is not automated and requires  
233 special external software, the descriptors used for the EB model were not included in the construction  
234 of the ML-based QSRR models. After cleaning missing values, this resulted in 146 molecular descriptors  
235 (listed in Supplementary data, Excel spreadsheet #3) used for constructing the models. Four ML  
236 algorithms were used: XGBoost (XGB, an extreme gradient boosting algorithm for trees algorithms),  
237 Light Gradient Boosting Machine (LGBM), a random forest (RF, a decision-tree algorithm), and a  
238 Bayesian regularized neural network (BRNN). Ten-fold cross-validation was employed for all models  
239 [35].

240 The model performances for RT prediction were evaluated by a set of standard performance  
241 criteria calculations found in the literature on evaluation of QSRR models [16, 29]. Thus, the following  
242 performance criteria were calculated on the testing set: RMSE (root-mean-square error) (1), MAE  
243 (mean absolute error in minutes) (2),  $R^2$  (coefficient of determination) (3), and  $A^{95\%}$  (prediction  
244 accuracy with a 95% confidence interval). For the sake of harmonization and comparison between  
245 models,  $A^{95\%}$  was recalculated for the EB model, following the calculations made by the Retip-package  
246 "get.score()". This function uses the "qnorm()" function, bundled as standard with R, in order to find  
247 the 95th percentile of a normal distribution whose mean and standard deviation correspond to the  
248 prediction errors.

$$249 \quad (1) \text{ RMSE} = \sum_{i=1}^n \sqrt{\frac{(\text{ExpRT}_i - \text{PredRT}_i)^2}{n}}$$

$$250 \quad (2) \text{ MAE} = \sum_{i=1}^n \frac{|\text{ExpRT}_i - \text{PredRT}_i|}{n}$$

$$251 \quad (3) R^2 = 1 - \frac{\sum_i (\text{PredRT}_i - \text{ExpRT}_i)^2}{\sum_i (\text{ExpRT}_i - \text{ExpRT}_i)^2}$$

252

### 253 2.2.3. Step 3: Tandem-mass spectra prediction by a fragmentation predictor

254 A fragmentation predictor, CFM-ID (v4.0), was used to predict the MS/MS spectra of the TPs  
255 predicted in step 1. This web-based model predicts *a priori* tandem mass spectra resulting from an  
256 electrospray ionization high-resolution tandem mass spectrometry (ESI-MS/MS). It was built using a  
257 neural network algorithm on a panel of experimental spectra of several compounds [36]. The  
258 prediction of compound spectra is carried out for three fragmentation levels, depending on their  
259 ionization energy value: low (10eV), medium (20eV), and high (40eV) energy. The SMILES of the TBZ  
260 TPs predicted in step 1 were taken as inputs. The model output for each SMILES consists of an  
261 individual text file containing the predicted spectra for the three energy levels (10eV/20eV/40eV)  
262 associated with potential intensities. The most abundant fragment of each predicted spectra was  
263 retained, resulting in a "blended" spectrum for each SMILES computed by the model. This blended  
264 strategy was performed using an in-house R script on the text file containing the compound spectra;  
265 the most abundant fragments of each spectrum predicted for a compound were compiled in an Excel  
266 spreadsheet. The most abundant fragment at each energy level was selected considering the use of a  
267 DIA mode ramping from 10 to 45eV. The associated predicted intensities were not included in the  
268 database as they are strongly influenced by the instrumentation and analytical conditions used.

269 In order to test the effectiveness of the fragments prediction and the proposed "blended"  
270 strategy, the predicted spectra were compared to experimental spectra for TBZ. The experimental

271 spectra were acquired as described in section 2.1.2., resulting in the “home-ramp” spectra. Four LC-  
272 ESI-QToF spectra were compiled from the MassBank database (<https://massbank.eu/MassBank>): three  
273 at the energy levels used by CFM-ID (10eV/20eV/40eV) and one at an “optimized” energy ramp (21.8–  
274 32.6 eV). A score of mass spectra similarity between all these spectra was calculated using the  
275 OrgMassSpecR package (v0.5-3) in R (v4.0.4). In addition to this calculation, the number of common  
276 fragments between mass spectra was investigated. The tolerance used to align the  $m/z$  values of the  
277 spectral fragments was 0.001  $m/z$ , which is consistent with the use of mass spectra from HRMS  
278 acquisition with a QToF.

279

### 280 2.3. Statistical analysis

281 All statistical analyses, comparisons and graphing of results were performed using R (v4.0.4)  
282 and Microsoft Excel (v16.0.4849.1000) software. The statistical relationship between sets of  
283 quantitative values was evaluated using Pearson’s correlation coefficient. Coefficients were  
284 considered significant at a  $p < 0.01$ .

285

## 286 3. Results and discussion

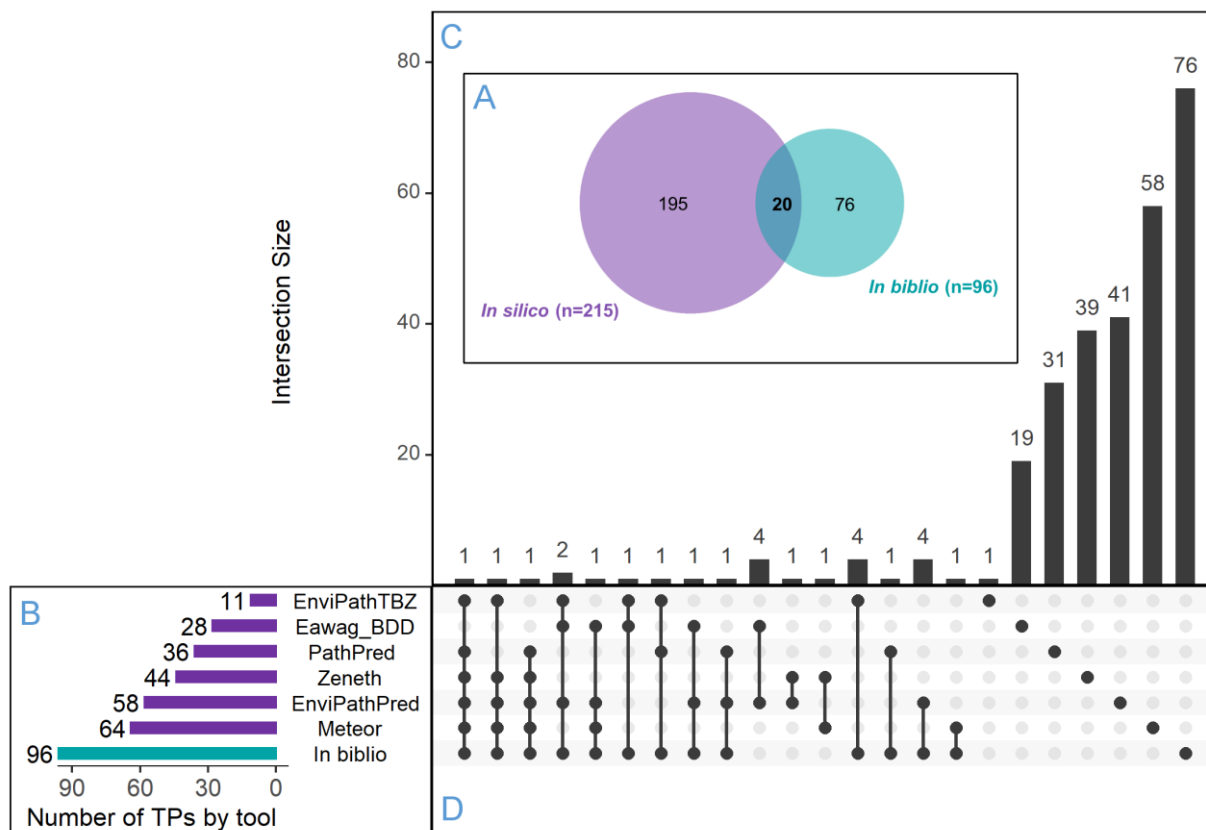
### 287 3.1. Comparison of *in silico* and *in biblio* predictions for transformation products

288 The six transformation predictors used were able to predict 215 distinct TPs for TBZ. Literature  
289 search yielded 97 TPs, predominantly from the work of Storck et al. [26], and El Azhari et al.[25] that  
290 included previous experimental studies on TBZ degradation. The full database of TBZ TPs created at  
291 this workflow step can be consulted at the following address: <https://doi.org/10.57745/Y3JLTV>

292 The overlap between the *in silico* transformation predictors and *in biblio* approaches was less  
293 than 7% (20 TPs in common, Figure 2 – A). This low overlap may be explained by the number and  
294 variety of transformation predictors used. These results are consistent with previous research, as Kern  
295 et al. [37] found a similar overlap of 8.4% between *in silico* prediction and literature data in a study on  
296 24 pesticides using one transformation predictor, UM-PPS (named ‘Eawag\_BDD’ in our study). The  
297 workflow proposed here differs from previous studies as it uses a large number of *in silico* prediction  
298 tools in combination. Given the range and variety of tools used, this low level of overlap is nevertheless  
299 unexpected and underscores the need for literature searches during the process of database creation  
300 for suspect screening of TPs.

301 The overlap in predicted TPs between the different *in silico* transformation predictors was also  
302 investigated (Figure 2). No TPs were predicted by all *in silico* transformation predictors. Four of the 6  
303 transformation predictors predicted the formation of 1,2,4-triazole, considered as the terminal TP [38].  
304 Also, four of the 6 transformation predictors predicted the formation of hydroxytebuconazole (5-(4-  
305 chlorophenyl)-2,2-dimethyl-3-(1H-1,2,4-triazol-1-ylmethyl)-1,3-pentanediol), one of the few TBZ TPs  
306 that can be readily purchased as an analytical standard. Despite these cases, the overall picture  
307 matched to the comparison between *in silico* transformation predictors and *in biblio* search. Indeed,  
308 most of the compounds predicted *in silico* do not have overlapping identities across the different  
309 transformation predictors used (Figure 2), with only 8% of compounds sharing identity overlap. This  
310 low level of overlap highlights the fact that models tend to over-predicting transformation products.

311 Nevertheless, this overlap should not be interpreted as a weakness of the transformation predictors  
 312 used for prediction, as it can be explained by the complementary of the transformation predictors  
 313 chosen for this study. As the selected transformation predictors cover a wide range of biotic processes  
 314 occurring in the environment, they can predict a large number of structurally-different TPs [12, 37].  
 315



316 **Figure 2.** Results of all-in-silico prediction and in biblio search with (A) the Venn diagram representing  
 317 the overlap between overall in silico prediction tools and in biblio search of TBZ TPs. (B) Number of  
 318 transformation products (TPs) and tebuconazole (TBZ) from the six in silico tools (purple) and the in  
 319 biblio search (cyan). (C) The bar chart shows the number of intersecting and non-intersecting TBZ TPs  
 320 between in silico tools and in biblio. (D) Table presenting the intersection between tools for each bar of  
 321 the bar chart.

322

323 Qualitatively speaking, such a large number of predicted TPs (n=215) could lead to possible  
 324 mismatching in identification or false-positives during subsequent suspect screening analyses of real  
 325 samples. This is especially true with isomers that may be tricky to differentiate, as reported by El Azhari  
 326 et al. [25]. Nonetheless, this *in silico* approach led to the identification of TPs that had never be  
 327 searched or detected before. Moreover, the cross-comparison of the predicted TBZ TPs obtained using  
 328 several *in silico* transformation predictors highlighted some well-known TPs, such as 1,2,4-triazole or  
 329 hydroxytebuconazole. Jiao et al. [13] recently detected 14 new TPs of the fungicide pyrisoxazole using  
 330 literature data and one *in silico* tool, Envipath [14], for database construction. All these findings  
 331 demonstrate that the creation of a TPs database using *in silico* transformation predictors can serve as  
 332 a complementary approach rather than a substitute for literature review.

333

### 334 3.2. Chromatographic retention time prediction by QSRR models

335 Results of the performance criteria calculations executed on the testing set (n=69,  
336 supplementary data - Excel spreadsheet #4) for the four ML-based QSRR algorithms (XGB, LightGBM,  
337 BRNN, and RF) are summarized in Table 2, along with the calculations for the MLR-based QSRR model  
338 (EB).

339 Among the four ML models, XGB showed the best performance with the lowest RMSE, MAE,  
340  $R^2$ , and  $A^{95\%}$  values for the testing set. These results are consistent with previous studies that have  
341 highlighted the good performance of gradient boosting models such as XGB among ML algorithms  
342 while emphasizing the importance of a large training set (> 100 experimental RT) for model building  
343 [39]. The prediction accuracy,  $A^{95\%}$ , computed for XGB (1.64 min for a 20-min chromatographic run or  
344  $\pm 8.2\%$  of the total chromatographic run) is in line with a recent study by Feng et al. (2021) who built  
345 an XGB model for RT prediction of pesticides and achieved an  $A^{95\%}$  of 1.14 min for a 15-minutes  
346 chromatographic run ( $\pm 7.6\%$  of the total chromatographic run), with 321 pesticides used as training  
347 set and 77 used as testing set [40]. This level of accuracy is also consistent with previous studies using  
348 other models (such as logP-based MLR, Artificial Neural Network, and QSRR-MLR) resulting in a  
349 prediction accuracy ranging from  $\pm 9\%$  to  $\pm 15\%$  of the total chromatographic run [41-44].

350

351 **Table 2.** Performance values calculated on the testing set (n=69) for the five QSRR models tested in this  
352 study. The acronyms used for the *in silico* tools are spelled out in section 2.2.

Performance criteria					
Model code	Algorithm	RMSE	MAE	$R^2$	$A^{95\%}$
XGB	ML	1.09	0.84	0.80	1.64
LGBM		1.13	0.78	0.86	1.81
BRNN		1.17	0.80	0.77	1.75
RF		1.23	0.95	0.75	1.72
EB	MLR	0.95	0.74	0.84	1.56

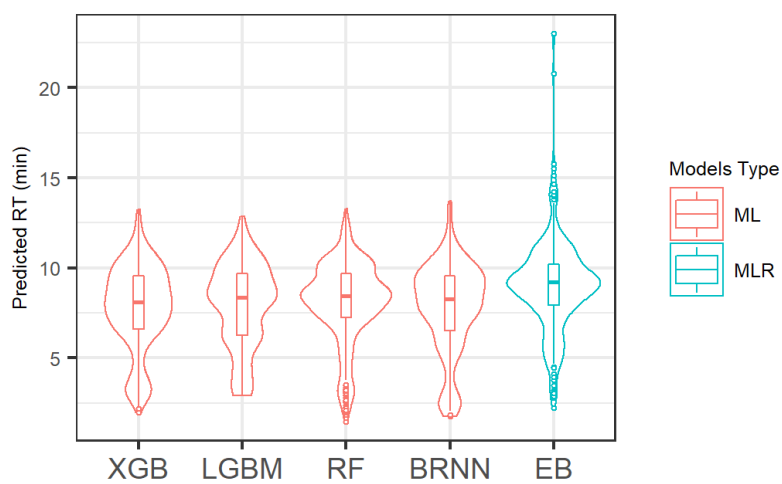
353

354

355 According to these performance results, the MLR-based QSRR (EB in Table 2) seems to be a  
356 better model than XGB. Indeed, it has the lowest RMSE and highest prediction accuracy of the five  
357 models tested, even though it was built from the least complex algorithm. This startling finding may  
358 be explained by the QSRR-based construction of the EB model, the high analytical relevance of the  
359 molecular descriptors used, and the optimization of the training and testing set used [29]. In order to  
360 compare the six models, we used the same training and testing set as described in Bride et al. [29].  
361 These sets were optimized for the construction of a MLR-based QSRR model and may not be fit for the  
362 construction of a QSRR model based on ML algorithms. The main difference between literature and  
363 this work and lies in the ratio used for splitting the training and testing sets, which is closer to 80:20  
364 (training set:test set) in literature [35, 40] versus a 65:35 ratio used by Bride et al. [29] and here. This  
365 change in ratio is reflected by a larger training set thus theoretically more efficient ML-based QSRR  
366 models.

367 The five QSRR models, compared on the training set of known compounds, were used to  
368 predict the RTs of the 291 TPs databased (Figure 3 and supplementary data - Excel spreadsheet #5).  
369 The predictions made by the models that performed best, i.e. XGB and EB, show an acceptable  
370 Pearson's correlation of 0.82 (supplementary data - Table S1). A more troubling result is the large  
371 number of outliers predicted by the EB model (Figure 3), with some values exceeding the  
372 chromatographic run time (>20 minutes). This may point to limitations of the MLR-based QSRR model,  
373 which may not be suited for this set of TPs. Indeed, the predicted properties of the TPs must be outside  
374 the field of application of the MLR-based QSRR model. As the MLR model (EB) is built solely on 8  
375 molecular descriptors, its field of application is easily surpassed, which limits its potential for use in  
376 predicting RTs of unknown compounds.

377



378 **Figure 3.** Violin plots for predicted chromatographic retention times (RT, in minutes) for the five QSRR  
379 models (XGB, LGBM, RF, BRNN, and EB) applied to the database of the 291 tebuconazole  
380 transformation products. Tools are classified according to model type (machine learning: ML;  
381 multilinear regression: MLR).

382

383 Based on the results of the present study, we suggest preferentially using the XGB model  
384 among the ML and MLR-based QSRR models for predicting chromatographic RTs. This is mainly  
385 because the XGB model had the best overall performances on the testing set, with the lowest RMSE,  
386 the highest  $A^{95\%}$ , and the fewest outliers in its prediction for this set of TPs. Moreover, like the other  
387 ML-based QSRR models tested here, the XGB model can be easily constructed from data obtained  
388 using different LC methods [35] and it can be automated for the molecular descriptors search using  
389 the RCDK package. All these factors make the XGB model easily transposable and less time-consuming  
390 for RT predictions than the MLR-based QSRR models like EB.

391

### 392 3.3. Tandem-mass spectra prediction by the fragmentation predictor

393 In order to test the effectiveness of the fragments prediction and the proposed “blended”  
394 strategy, we compared the predicted and experimental spectra of TBZ. The similarity scores calculated  
395 to evaluate the similarity of the spectra, as well as the number of common fragments between all the  
396 spectra discussed here, are presented in Table 3 (all values are compiled in a larger comparison matrix  
397 in Table S3). For visual observation of compared mass spectra, their head-to-tail plots are given in

398 supplementary data - figures S2 and S3. The comparison of predicted vs experimental TBZ spectra  
 399 revealed poor similarity scores at the corresponding fixed ionization energies (10, 20, 40 eV). This is  
 400 connected to the small number of common fragments between the predicted and experimental  
 401 spectra. In contrast, the comparison of 'blended' predicted spectra vs experimental energy-ramped  
 402 spectra shows good similarity scores (0.84) as well as two common fragments. A low similarity score  
 403 between the "Home-ramp" spectra and "MassBank-ramp" spectra (0.14), for the same number of  
 404 common fragments, is explained by the way the score itself is calculated. Indeed, the calculation takes  
 405 into account the intensity of the fragment, which biases this comparison, given the different ionization  
 406 energy values of the ramps applied ("Home-ramp": 10–45 eV, "MassBank-ramp": 21.8–32.6 eV).  
 407 Nevertheless, these calculated scores are important for theoretical comparison, and what matters  
 408 most for suspect screening analysis in practice is the fragments found in samples corresponding to  
 409 screened compounds. The highest number of common fragments was found between the  
 410 experimental and "blended" predicted mass spectra, highlighting its effectiveness. Based on these  
 411 comparison results for tebuconazole, we suggest the use of a fragmentation with an energy ramp,  
 412 which revealed more predicted fragments than a fragmentation at fixed energies. To corroborate  
 413 these findings, this spectra similarity comparison should be performed for a TBZ TP, such as the  
 414 hydroxytebuconazole or 1,2,4-triazole. However, the MassBank database does not have QToF-  
 415 acquired spectra of these very specific TPs.

416

417 **Table 3.** Comparison of experimental and predicted mass spectra for tebuconazole. For each set of  
 418 mass spectra compared, the score obtained by the "SpectrumSimilarity()" function is given along with  
 419 the number of fragments in common (in brackets). This table is an excerpt from the full comparison  
 420 matrix detailed in Supporting Information (Table S3).

	<b>Experimental</b>				
	<b>Home-ramp</b>	<b>MassBank- ramp</b>	<b>MassBank- 10 eV</b>	<b>MassBank- 20 eV</b>	<b>MassBank- 40 eV</b>
<b>Experimental: MassBank-ramp</b>	0.14 (3)				
<b>Predicted: Blended</b>	0.10 (2)	0.84 (2)			
<b>Predicted: 10 eV</b>			0.91 (1)	0.93 (1)	0.00 (0)
<b>Predicted: 20 eV</b>			0.00 (0)	0.00 (0)	0.00 (0)
<b>Predicted: 40 eV</b>			0.00 (0)	0.00 (0)	0.00 (0)

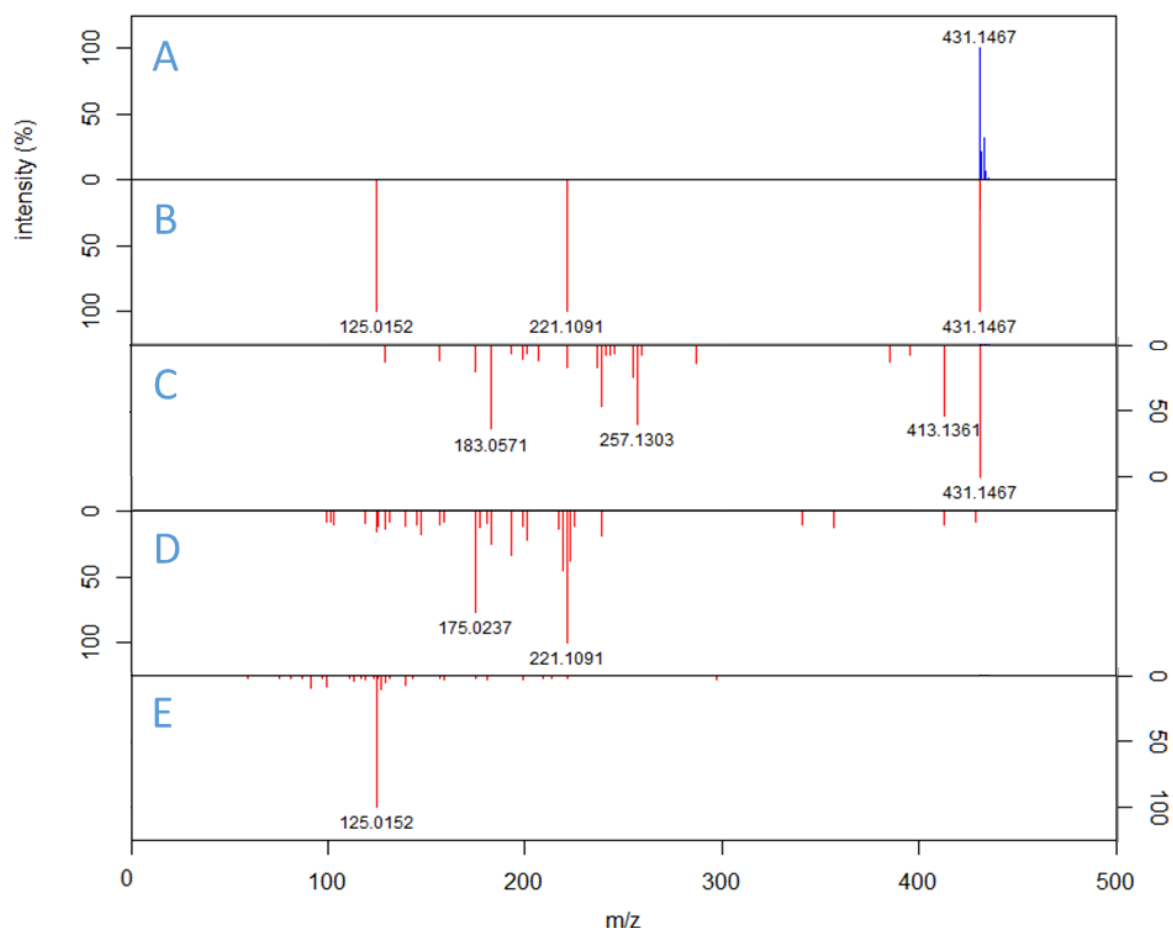
421

422

423 The MS/MS spectra of the 291 TPs of TBZ incremented in the database developed here were  
 424 predicted with CFM-ID (v4.0) at three ionization energy levels, which resulted in 873 spectra contained  
 425 within 291 distinct text files. Applying the blended strategy on the spectra (Figure 4, supplementary  
 426 data - figure S1) led to a set of 634 fragments compiled in the database. These fragments are often  
 427 shared by multiple TPs; among these 634 predicted fragments, only 179 (around 30%) were unique.  
 428 Indeed, the 291 TPs were predicted from a single compound, TBZ, and so most of them logically share  
 429 similar parts of molecular structures (database available at the following address:  
 430 <https://doi.org/10.57745/Y3JLTV>), resulting in similar fragmentation patterns. Furthermore, a single  
 431 TP may share the same most abundant fragment at two different energy levels, which limits the  
 432 number of different fragments per compound. As a result, one to three predicted fragments per

433 compound were incorporated in the database. Nevertheless, incrementing the associated fragments  
434 of TBZ TPs enhanced the database and is expected to limit mismatches during subsequent suspect  
435 screening analysis. For example, TP\_096 and TP\_220 share the same chemical formula and are  
436 predicted to elute at similar RTs (7.61 and 7.34 minutes, respectively), but they disassemble into  
437 different fragments according to fragmentation model used (supplementary data – table S2). If this  
438 predicted difference in fragmentation pattern is verified during the analysis, it will allow discrimination  
439 of the two TPs.

440



441 **Figure 4.** Head-to-tail plot of different mass spectra of the tebuconazole transformation product  
442 TP\_095 from the database. (A) Predicted isotope pattern with no ionization energy applied. (B)  
443 'Blended' spectra, emerging from the predicted spectrum of different energy levels used in the  
444 fragmentation prediction. (C) Predicted spectra on energy = 10eV. (D) Predicted spectra on energy =  
445 20eV. (E) Predicted spectra on energy = 40eV.

446

447 The main limitation of the use of predicted fragments in this study is the sensitivity of the  
448 instrument used here. Indeed, no precursor ions were isolated with the DIA mode used, which leads  
449 to exhaustive fragmentation spectra that are not specific to a compound but specific to the scan  
450 previously acquired. In addition, TPs are often present at trace amounts in environmental samples,  
451 which could result in fragments of TPs close to or below the analytical background noise, thus negating  
452 their identification during suspect screening.

453 With these points in mind, using CFM-ID predictions and incorporating predicted fragments  
454 into the database still increases the elucidation power of the database. Indeed, it provides an  
455 additional *a priori* filter on the fragmentation pattern during suspect analysis and thus enables some  
456 outliers to be ruled out. This *a-priori* filter, obtained by prediction by CFM-ID, can be strengthened by  
457 a comparison with an *a posteriori* prediction based on experimentally-acquired spectra, using tools  
458 such as MetFrag. In a complementary way, a common fragmentation pathway approach as applied by  
459 Ibáñez and al. (2017)[45] as well as Wielens Becker and al. (2020)[46] could be considered, given the  
460 large number of common fragments shared between the tebuconazole TPs, as predicted by CFM-ID.  
461 Applying this complementary approach could reveal TPs missed during the prediction step or confirm  
462 those already identified.

463

### 464 3.4. Application of the workflow to environmental samples

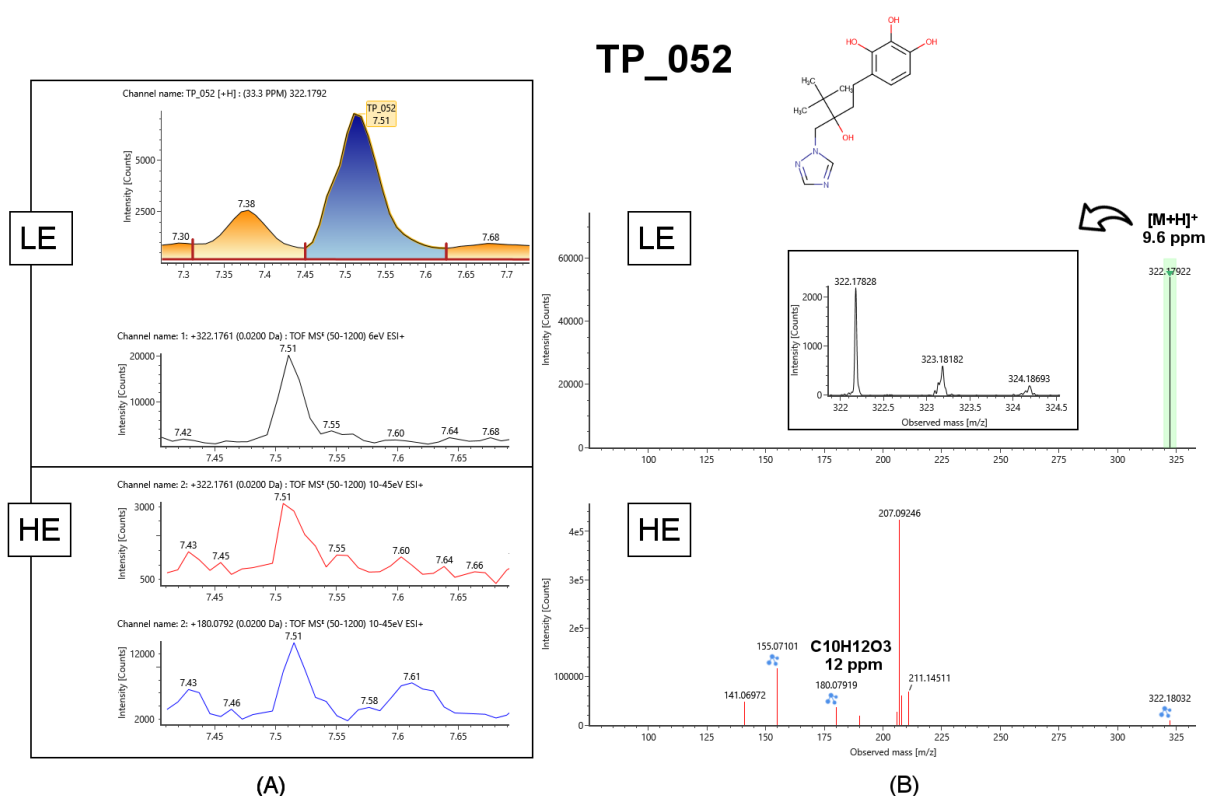
465 To illustrate the efficiency of the database created here, we ran retrospective suspect  
466 screening for TBZ TPs on environmental samples. The selected samples used here were collected  
467 within the framework of the French prospective surveillance network [47] (supplementary data –  
468 figure S4). Surface waters collected from 20 sites in France were filtered, in order to analyze the  
469 dissolved fraction, extracted, and then analyzed by LC-HRMS in our laboratory in 2018. The data were  
470 collected using the same acquisition method as described in the instrumentation section (2.1.2.), and  
471 the resulting information was purpose-stored to allow retrospective screening.

472 The whole suspect screening workflow was applied to the water samples using Waters' UNIFI  
473 software and the created database containing information on TBZ and 291 of its TPs. Identification of  
474 TBZ and its TPs was performed with the following threshold criteria: (1) mass accuracy:  $\leq 10$  ppm; (2)  
475 chromatographic RT:  $\leq 2$  minutes; (3) isotopic pattern match  $m/z$  RMS  $\leq 10$  ppm, isotopic pattern  
476 match intensity RMS  $\leq 20\%$ ; (4) uniqueness; no detection in the analytical or field blanks. TBZ was  
477 detected at 8 of the 20 sites and its TPs were detected at 5 sites (information about the detected  
478 compounds and detailed detection results and can be found in supplementary data, Excel spreadsheet  
479 #6 and Excel spreadsheet #7). The TBZ TPs were only detected in samples from agricultural catchments  
480 where TBZ was also quantified. To the best of our knowledge, six of seven TPs suspected in the present  
481 study were detected for the first time in surface waters samples.

482 Among the 7 different TPs found, 6 come from *in silico* prediction (Figure S5), 4 of which  
483 originate from the 'EnviPath' predictor [14]. These results demonstrate the ability of 'EnviPath' to  
484 generate accurate TPs for river waters, and justify its exclusive use in recent works [13, 40, 48].  
485 Nevertheless, the application of several other *in silico* transformation predictors, as in the workflow  
486 proposed here, led to a more exhaustive detection of TPs. The two remaining TPs from *in silico*  
487 predictions were predicted by the transformation predictors 'PathPred' [32] and 'Zeneth' [34]. The  
488 whole identification process was enhanced by the use of predicted chromatographic RTs, with the  
489 accuracy of the XGB prediction used as a threshold. Using this threshold over the 24 hits among the  
490 injections, 16 outliers candidates were eliminated for 7 retained TPs. CFM-ID failed to predict enough  
491 fragments of the detected TPs to make it useful in the discrimination of compounds in our suspect  
492 screening strategy. This is probably due to the very low concentrations of TPs in these water samples,  
493 which resulted in fragment intensities that were below the analytical background. These suspected  
494 transformation products could be qualified with a certitude at level 4 ("tentative candidates") to 3  
495 ("unequivocal molecular formula") [9], as for some of them, no fragmentation pattern was detected.  
496 In order to reach the level 2B ("diagnostic probable structure"), further search of specific fragments



497 need to be performed. This could be done manually, or with *a posteriori* tool such as MetFrag [19]  
 498 which make predictions on acquired fragmentation spectra. It is important to note that these detection  
 499 results have relatively large mass error values for a HRMS instrument, with a mean of 5.7 Da  
 500 (Supplementary data, Excel spreadsheet #5). This lack of accuracy can cause identification problems,  
 501 as illustrated for the TP\_052 on figure 5. No fragmentation pattern was confirmed for this compound  
 502 mainly due to mass error value higher than 10 ppm on the predicted fragments. This large mass error  
 503 values are potentially due to a strong matrix effect in the surface water samples. Nonetheless, targeted  
 504 analysis operated on a liquid chromatography – tandem mass spectrometry (UHPLC TQ-XS, Waters)  
 505 confirmed the presence of tebuconazole in the same samples. These results highlight the effectiveness  
 506 of the proposed workflow in the search for unknown TPs in environmental matrices. Applied on TBZ,  
 507 the created database of TPs was used on a set of previously analyzed surface water samples, and led  
 508 to the detection of 6 previously-unseen TPs for this matrix.



509 **Figure 5.** Tentative Identification of TP\_052. (A) Extracted ion chromatogram (EIC) of protonated  
 510 TP\_052 at Low Energy (LE) with a 33 ppm mass error window (resulting from the UNIFI treatment), and  
 511 at a 0.02 Da mass error window (from a manual extraction). EIC of predicted fragments of TP\_052 at  
 512 High Energy (HE) with a 0.02 Da mass error window. (B) Mass and detected isotopic pattern of TP\_052,  
 513 on LE and HE mass spectra generated by UNIFI. Blue symbols on HE spectra show which fragment is  
 514 taken in account in fragmentation prediction that UNIFI operates.

515

#### 516 4. Conclusions

517 This study proposed a comprehensive workflow for the implementation of detailed and ready-  
 518 to-use databases to support suspect screening analyses of unknown compounds in agricultural  
 519 watersheds. This novel workflow, combining several *in silico* tools, was applied on tebuconazole. It

520 allowed the creation of a database of 291 tebuconazole transformation products, incremented with  
521 their predicted chromatographic retention times and fragment patterns.

522 The six transformation predictors allowed to predict a large number of TPs (215), including  
523 several TPs that have never been searched before. This large number of predicted compounds  
524 highlights the over-prediction that models may perform. We demonstrated that *in silico* prediction is  
525 a complementary approach to literature review. The low overlap between the prediction process and  
526 literature data (7%) and between the various transformation predictors (8%) should be considered as  
527 an opportunity to extend the range of transformation products investigated. Moreover, the cross-  
528 comparison of the transformation predictors may be useful in order to single out well known TPs. Given  
529 the chemical properties of TBZ, we only used one *in silico* transformation predictor for abiotic  
530 degradation ('Zeneth'). Depending on the compounds studied, the workflow described here may need  
531 to be complemented by other suitably appropriate prediction tools. However, abiotic degradation is  
532 often considered difficult to predict and suffers from a lack of a freely-available transformation  
533 predictor.

534 Concerning the prediction of chromatographic retention times, XGB, a machine learning-based  
535 QSRR, was the model that performed the best, with the lowest of RMSE values and highest prediction  
536 accuracy. We therefore advocate preferentially using XGB to predict the retention times of further  
537 unknown compounds.

538 Regarding fragments prediction, CFM-ID was used to predict *a priori* the MS/MS spectra of  
539 tebuconazole transformation products. This approach mobilizing *a priori in silico* fragmentation  
540 prediction together with a blended strategy on predicted spectra limited compound mismatching and  
541 thus enhanced the database created. This *a priori* approach could be further strengthened by *a*  
542 *posteriori* prediction of fragments on LC-HRMS spectra acquired from environmental samples.

543 The strength of the complete workflow presented here lies in the hyphenated use of several  
544 cutting-edge *in silico* tools—most of which are freely available—transposable to different LC-MS  
545 methods and to various organic contaminants, whether they already known or still unknown. Used on  
546 tebuconazole, this workflow resulted in a database of 291 transformation products which was then  
547 applied on a set of 20 real-world surface-water samples acquired in 2018. This retrospective suspect  
548 screening analysis led to the detection of 6 transformation products that had never been detected  
549 before. We anticipate this novel workflow approach as a starting point for studies on other pesticides  
550 in different environmental samples such as surface waters or groundwaters and sediments or soils, in  
551 order to further demonstrate its effectiveness for *in situ* suspect screening of a wide range of pesticides  
552 transformation products.

553

## 554 **Acknowledgments**

555 This work was performed as part of the "TAPIOCA" project funded by the French National  
556 Office for Biodiversity (OFB) and the Ecophyto II program. The authors thank the OFB's Réseau de  
557 surveillance prospective' and Céline Guillemain for the acquisition of UHPLC-HRMS data on surface  
558 water samples. We thank Sylvain Merel for providing access to the predictions done using Meteor  
559 Nexus and Zeneth software packages. We also thank MetaForm Langues for English editing. We are  
560 grateful to reviewers and editor for their helpful comments.

561 **References**

562

563 [1] K. Fenner, S. Canonica, L.P. Wackett, M. Elsner, Evaluating Pesticide Degradation in the  
564 Environment: Blind Spots and Emerging Opportunities, *Science*, 341 (2013) 752-758.

565 [2] B.I. Escher, K. Fenner, Recent Advances in Environmental Risk Assessment of Transformation  
566 Products, *Environmental Science & Technology*, 45 (2011) 3835-3847.

567 [3] A.B.A. Boxall, C.J. Sinclair, K. Fenner, D. Kolpin, S.J. Maund, Peer Reviewed: When Synthetic  
568 Chemicals Degrade in the Environment, *Environmental Science & Technology*, 38 (2004) 368A-375A.

569 [4] C. Postigo, D. Barceló, Synthetic organic compounds and their transformation products in  
570 groundwater: Occurrence, fate and mitigation, *Science of The Total Environment*, 503-504 (2015) 32-  
571 47.

572 [5] B.J. Mahler, L.H. Nowell, M.W. Sandstrom, P.M. Bradley, K.M. Romanok, C.P. Konrad, P.C. Van  
573 Metre, Inclusion of Pesticide Transformation Products Is Key to Estimating Pesticide Exposures and  
574 Effects in Small U.S. Streams, *Environmental Science & Technology*, 55 (2021) 4740-4752.

575 [6] C. Moschet, I. Wittmer, J. Simovic, M. Junghans, A. Piazzoli, H. Singer, C. Stamm, C. Leu, J. Hollender,  
576 How a Complete Pesticide Screening Changes the Assessment of Surface Water Quality, *Environmental  
577 Science & Technology*, 48 (2014) 5423-5432.

578 [7] F. Le Cor, S. Slaby, V. Dufour, A. Iuretig, C. Feidt, X. Dauchy, D. Banas, Occurrence of pesticides and  
579 their transformation products in headwater streams: Contamination status and effect of ponds on  
580 contaminant concentrations, *Science of The Total Environment*, 788 (2021) 147715.

581 [8] B.I. Escher, H.M. Stapleton, E.L. Schymanski, Tracking complex mixtures of chemicals in our  
582 changing environment, *Science*, 367 (2020) 388.

583 [9] E.L. Schymanski, J. Jeon, R. Gulde, K. Fenner, M. Ruff, H.P. Singer, J. Hollender, Identifying Small  
584 Molecules via High Resolution Mass Spectrometry: Communicating Confidence, *Environmental  
585 Science & Technology*, 48 (2014) 2097-2098.

586 [10] E. Fonseca, A. Renau-Pruñonosa, M. Ibáñez, E. Gracia-Lor, T. Estrela, S. Jiménez, M.Á. Pérez-  
587 Martín, F. González, F. Hernández, I. Morell, Investigation of pesticides and their transformation  
588 products in the Júcar River Hydrographical Basin (Spain) by wide-scope high-resolution mass  
589 spectrometry screening, *Environmental Research*, 177 (2019) 108570.

590 [11] K. Kiefer, A. Müller, H. Singer, J. Hollender, New relevant pesticide transformation products in  
591 groundwater detected using target and suspect screening for agricultural and urban micropollutants  
592 with LC-HRMS, *Water Research*, 165 (2019) 114972.

593 [12] A.A. Bletsou, J. Jeon, J. Hollender, E. Archontaki, N.S. Thomaidis, Targeted and non-targeted liquid  
594 chromatography-mass spectrometric workflows for identification of transformation products of  
595 emerging pollutants in the aquatic environment, *TrAC Trends in Analytical Chemistry*, 66 (2015) 32-44.

596 [13] B. Jiao, Y. Zhu, J. Xu, F. Dong, X. Wu, X. Liu, Y. Zheng, Identification and ecotoxicity prediction of  
597 pyrisoxazole transformation products formed in soil and water using an effective HRMS workflow,  
598 *Journal of Hazardous Materials*, 424 (2022) 127223.

599 [14] J. Wicker, T. Lorschach, M. Gütlein, E. Schmid, D. Latino, S. Kramer, K. Fenner, enviPath – The  
600 environmental contaminant biotransformation pathway resource, *Nucleic Acids Research*, 44 (2016)  
601 D502-D508.

602 [15] R.I.J. Amos, P.R. Haddad, R. Szucs, J.W. Dolan, C.A. Pohl, Molecular modeling and prediction  
603 accuracy in Quantitative Structure-Retention Relationship calculations for chromatography, *TrAC*  
604 *Trends in Analytical Chemistry*, 105 (2018) 352-359.

605 [16] R. Aalizadeh, M.-C. Nika, N.S. Thomaidis, Development and application of retention time  
606 prediction models in the suspect and non-target screening of emerging contaminants, *Journal of*  
607 *Hazardous Materials*, 363 (2019) 277-285.

608 [17] K. Dührkop, M. Fleischauer, M. Ludwig, A.A. Aksenov, A.V. Melnik, M. Meusel, P.C. Dorrestein, J.  
609 Rousu, S. Böcker, SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure  
610 information, *Nature Methods*, 16 (2019) 299-302.

611 [18] L. Ridder, J.J.J. van der Hooft, S. Verhoeven, R.C.H. de Vos, R. van Schaik, J. Vervoort, Substructure-  
612 based annotation of high-resolution multistage MSn spectral trees, *Rapid Communications in Mass*  
613 *Spectrometry*, 26 (2012) 2461-2471.

614 [19] C. Ruttkies, E.L. Schymanski, S. Wolf, J. Hollender, S. Neumann, MetFrag relaunched: incorporating  
615 strategies beyond in silico fragmentation, *Journal of Cheminformatics*, 8 (2016) 3.

616 [20] E. Eysseric, F. Beaudry, C. Gagnon, P.A. Segura, Non-targeted screening of trace organic  
617 contaminants in surface waters by a multi-tool approach based on combinatorial analysis of tandem  
618 mass spectra and open access databases, *Talanta*, 230 (2021) 122293.

619 [21] Y. Djoumbou-Feunang, A. Pon, N. Karu, J. Zheng, C. Li, D. Arndt, M. Gautam, F. Allen, D.S. Wishart,  
620 CFM-ID 3.0: Significantly Improved ESI-MS/MS Prediction and Compound Identification, *Metabolites*,  
621 9 (2019) 72.

622 [22] A. Chao, H. Al-Ghoul, A.D. McEachran, I. Balabin, T. Transue, T. Cathey, J.N. Grossman, R.R. Singh,  
623 E.M. Ulrich, A.J. Williams, J.R. Sobus, In silico MS/MS spectra for identifying unknowns: a critical  
624 examination using CFM-ID algorithms and ENTACT mixture samples, *Analytical and Bioanalytical*  
625 *Chemistry*, 412 (2020) 1303-1315.

626 [23] P. Cabras, A. Angioni, V.L. Garau, M. Melis, F.M. Pirisi, E.V. Minelli, F. Cabitza, M. Cubeddu, Fate of  
627 Some New Fungicides (Cyprodinil, Fludioxonil, Pyrimethanil, and Tebuconazole) from Vine to Wine,  
628 *Journal of Agricultural and Food Chemistry*, 45 (1997) 2708-2710.

629 [24] S. Li, Q. Sun, Q. Wu, W. Gui, G. Zhu, D. Schlenk, Endocrine disrupting effects of tebuconazole on  
630 different life stages of zebrafish (*Danio rerio*), *Environmental Pollution*, 249 (2019) 1049-1059.

631 [25] N. El Azhari, E. Dermou, R.L. Barnard, V. Storck, M. Tourna, J. Beguet, P.A. Karas, L. Lucini, N.  
632 Rouard, L. Botteri, F. Ferrari, M. Trevisan, D.G. Karpouzias, F. Martin-Laurent, The dissipation and  
633 microbial ecotoxicity of tebuconazole and its transformation products in soil under standard laboratory  
634 and simulated winter conditions, *Science of The Total Environment*, 637-638 (2018) 892-906.

635 [26] V. Storck, L. Lucini, L. Mamy, F. Ferrari, E.S. Papadopoulou, S. Nikolaki, P.A. Karas, R. Servien, D.G.  
636 Karpouzias, M. Trevisan, P. Benoit, F. Martin-Laurent, Identification and characterization of  
637 tebuconazole transformation products in soil by combining suspect screening and molecular typology,  
638 *Environmental Pollution*, 208 (2016) 537-545.

639 [27] R.M. de Souza, D. Seibert, H.B. Quesada, F. de Jesus Bassetti, M.R. Fagundes-Klen, R. Bergamasco,  
640 Occurrence, impacts and general aspects of pesticides in surface water: A review, *Process Safety and*  
641 *Environmental Protection*, 135 (2020) 22-37.

642 [28] D. Kang, K. Doudrick, N. Park, Y. Choi, K. Kim, J. Jeon, Identification of transformation products to  
643 characterize the ability of a natural wetland to degrade synthetic organic pollutants, *Water Research*,  
644 187 (2020) 116425.

645 [29] E. Bride, S. Heinisch, B. Bonnefille, C. Guillemain, C. Margoum, Suspect screening of environmental  
646 contaminants by UHPLC-HRMS and transposable Quantitative Structure-Retention Relationship  
647 modelling, *Journal of Hazardous Materials*, 409 (2021) 124652.

648 [30] C. Tebes-Stevens, J.M. Patel, W.J. Jones, E.J. Weber, Prediction of Hydrolysis Products of Organic  
649 Chemicals under Environmental pH Conditions, *Environmental Science & Technology*, 51 (2017) 5008-  
650 5016.

651 [31] J. Gao, L.B.M. Ellis, L.P. Wackett, The University of Minnesota Pathway Prediction System: multi-  
652 level prediction and visualization, *Nucleic Acids Research*, 39 (2011) W406-W411.

653 [32] Y. Moriya, D. Shigemizu, M. Hattori, T. Tokimatsu, M. Kotera, S. Goto, M. Kanehisa, PathPred: an  
654 enzyme-catalyzed metabolic pathway prediction server, *Nucleic Acids Research*, 38 (2010) W138-  
655 W143.

656 [33] C.A. Marchant, K.A. Briggs, A. Long, In Silico Tools for Sharing Data and Knowledge on Toxicity and  
657 Metabolism: Derek for Windows, Meteor, and Vitic, *Toxicology Mechanisms and Methods*, 18 (2008)  
658 177-187.

659 [34] A.D.C. Parenty, W.G. Button, M.A. Ott, An Expert System To Predict the Forced Degradation of  
660 Organic Molecules, *Molecular Pharmaceutics*, 10 (2013) 2962-2974.

661 [35] P. Bonini, T. Kind, H. Tsugawa, D.K. Barupal, O. Fiehn, Retip: Retention Time Prediction for  
662 Compound Annotation in Untargeted Metabolomics, *Analytical Chemistry*, 92 (2020) 7515-7522.

663 [36] F. Wang, J. Liigand, S. Tian, D. Arndt, R. Greiner, D.S. Wishart, CFM-ID 4.0: More Accurate ESI-  
664 MS/MS Spectral Prediction and Compound Identification, *Analytical Chemistry*, 93 (2021) 11692-  
665 11700.

666 [37] S. Kern, K. Fenner, H.P. Singer, R.P. Schwarzenbach, J. Hollender, Identification of Transformation  
667 Products of Organic Contaminants in Natural Waters by Computer-Aided Prediction and High-  
668 Resolution Mass Spectrometry, *Environmental Science & Technology*, 43 (2009) 7039-7046.

669 [38] EFSA, Conclusion on the peer review of the pesticide risk assessment of the active substance  
670 tebuconazole, *EFSA Journal*, 12 (2014) 3485.

671 [39] R. Bouwmeester, L. Martens, S. Degroeve, Comprehensive and Empirical Evaluation of Machine  
672 Learning Algorithms for Small Molecule LC Retention Time Prediction, *Analytical Chemistry*, 91 (2019)  
673 3694-3703.

674 [40] C. Feng, Q. Xu, X. Qiu, Y.e. Jin, J. Ji, Y. Lin, S. Le, J. She, D. Lu, G. Wang, Evaluation and application  
675 of machine learning-based retention time prediction for suspect screening of pesticides and pesticide  
676 transformation products in LC-HRMS, *Chemosphere*, 271 (2021) 129447.

677 [41] R. Bade, L. Bijlsma, T.H. Miller, L.P. Barron, J.V. Sancho, F. Hernández, Suspect screening of large  
678 numbers of emerging contaminants in environmental waters using artificial neural networks for  
679 chromatographic retention time prediction and high resolution mass spectrometry data analysis,  
680 *Science of The Total Environment*, 538 (2015) 934-941.

681 [42] R. Bade, L. Bijlsma, J.V. Sancho, F. Hernández, Critical evaluation of a simple retention time  
682 predictor based on LogKow as a complementary tool in the identification of emerging contaminants in  
683 water, *Talanta*, 139 (2015) 143-149.

684 [43] M.-C. Nika, A.A. Bletsou, E. Koumaki, C. Noutsopoulos, D. Mamais, A.S. Stasinakis, N.S. Thomaidis,  
685 Chlorination of benzothiazoles and benzotriazoles and transformation products identification by LC-  
686 HR-MS/MS, *Journal of Hazardous Materials*, 323 (2017) 400-413.

- 687 [44] A.D. McEachran, K. Mansouri, S.R. Newton, B.E.J. Beverly, J.R. Sobus, A.J. Williams, A comparison  
688 of three liquid chromatography (LC) retention time prediction models, *Talanta*, 182 (2018) 371-379.
- 689 [45] M. Ibáñez, V. Borova, C. Boix, R. Aalizadeh, R. Bade, N.S. Thomaidis, F. Hernández, UHPLC-QTOF  
690 MS screening of pharmaceuticals and their metabolites in treated wastewater samples from Athens,  
691 *Journal of Hazardous Materials*, 323 (2017) 26-35.
- 692 [46] R.W. Becker, D.S. Araújo, C. Sirtori, N.P. Toyama, D.A. Tavares, G.A. Cordeiro, S.F. Benassi, A.C.  
693 Gossen, B. do Amaral, Pesticides in surface water from Brazil and Paraguay cross-border region:  
694 Screening using LC-QTOF MS and correlation with land use and occupation through multivariate  
695 analysis, *Microchemical Journal*, 168 (2021) 106502.
- 696 [47] B. Mathon, A. Dabrin, I. Allan, S. Lardy-Fontan, A. Togola, J.-P. Ghestem, C. Tixier, J.-L. Gonzalez,  
697 M. Ferreol, L. Dherret, A. Yari, L. Richard, A. Moreira, M. Eon, B. Delest, E. Noel-Chery, M. El Mossaoui,  
698 E. Alasonati, P.-F. Staub, N. Mazzella, C. Miège, Surveillance prospective – évaluation de la pertinence  
699 des échantillonneurs intégratifs passifs (EIP) pour la surveillance réglementaire des milieux aquatiques,  
700 *Rapport AQUAREF 2020*, (2020) 172.
- 701 [48] J. Zhou, D. Wang, F. Ju, W. Hu, J. Liang, Y. Bai, H. Liu, J. Qu, Profiling microbial removal of  
702 micropollutants in sand filters: Biotransformation pathways and associated bacteria, *Journal of*  
703 *Hazardous Materials*, 423 (2022) 127167.