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-04172394v1

Submitted on 27 Jul 2023

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

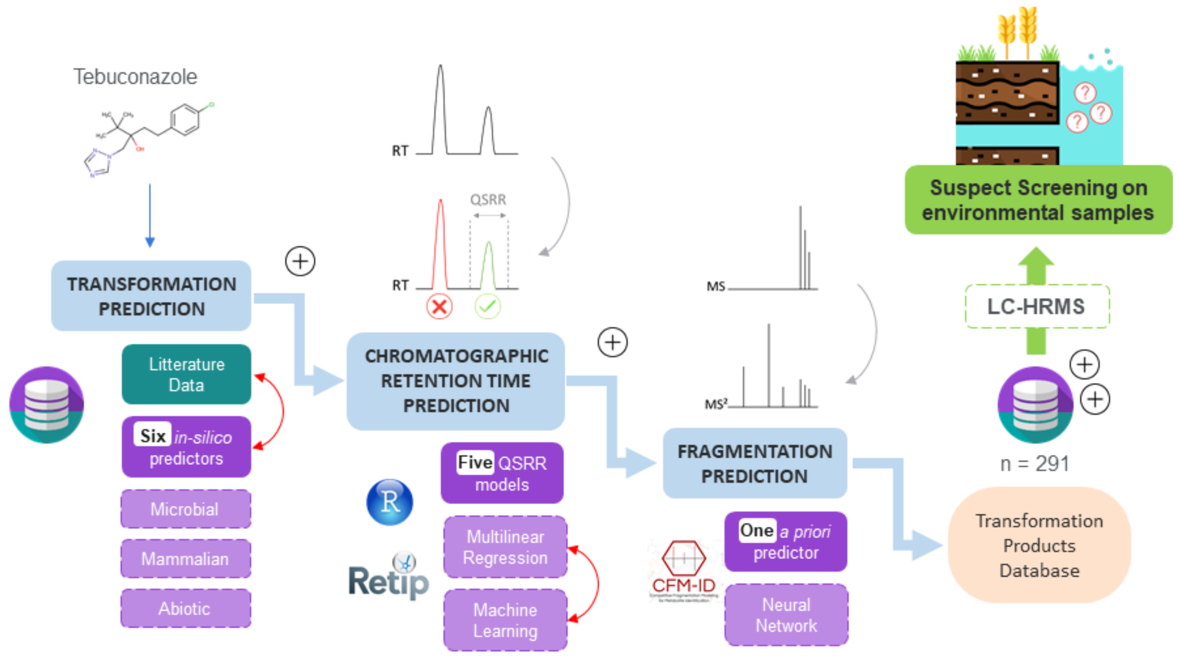
Kevin Rocco*, Christelle Margoum, Loïc Richard, Marina Coquery

INRAE, UR RiverLy, 69625 Villeurbanne, France

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

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

GRAPHICAL ABSTRACT



HIGHLIGHTS

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

ABSTRACT

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

KEYWORDS

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

ENVIRONMENTAL IMPLICATION

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

1. Introduction

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

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

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

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

Some *in silico* tools, called transformation predictors, can predict the formation of possible TPs by using the chemical identifiers of the parent compound as an input. These tools are based on various pre-established physicochemical reactions that can occur in various environmental compartments (e.g. aquatic, terrestrial or biological) via both abiotic and biotic transformation processes on scales running from microbial up to mammalian metabolism. The appropriate transformation predictor has to be selected based on the environmental degradation processes investigated. TPs predicted by these transformation predictors carry a relatively high rate of false-positives, but some predictors can use relative reasoning to address this issue [12]. The efficiency of these tools has already been proven. For instance, Jiao et al. [13] recently detected 14 new TPs of the fungicide pyrisoxazole using literature data and one in silico tool, Envipath [14], for database construction.

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

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

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

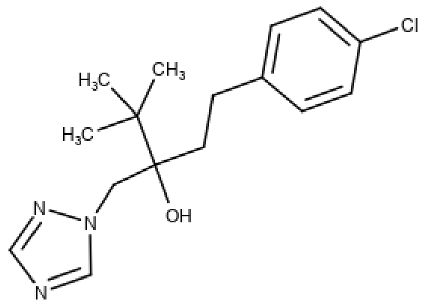
2. Materials and methods

2.1. *Experimental*

2.1.1 Pesticide selection

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

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

<table>
<tr><th colspan="2">Structure</th><th colspan="2">Compound name</th></tr>
<tr><td colspan="2" rowspan="5">
</td><td colspan="2">1-(4-chlorophenyl)-4,4-dimethyl-3-(1,2,4-triazol-1-ylmethyl)pentan-3-ol</td></tr>
<tr><td colspan="2">SMILES</td></tr>
<tr><td colspan="2">Clc1ccc(cc1)CCC(O)(C(C)(C)C)Cn2ncnc2</td></tr>
<tr><td colspan="2">InChiKey</td></tr>
<tr><td colspan="2">PXMNMQRDXWABCY-UHFFFAOYSA-N</td></tr>
<tr><td>DT50$_{\mathbf{Soil}}$ (EFSA 2014)</td><td>19.9–91.6 days</td><td>Formula</td><td>$C_{16}H_{22}ClN_3O$</td></tr>
<tr><td>DT50$_{\mathbf{Water-pH7}}$ (EFSA 2014)</td><td>590 days</td><td>Mass (g.mol^{-1})</td><td>307.8180</td></tr>
</table>

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

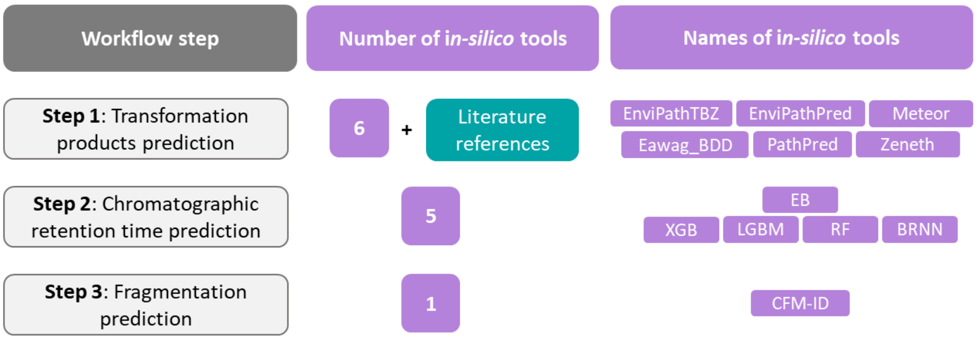
2.1.2. Instrumentation

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

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

2.2. Database creation for pesticide transformation products

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



Workflow step	Number of *in-silico* tools	Names of *in-silico* tools
Step 1: Transformation products prediction	6 + Literature references	EnviPathTBZ EnviPathPred Meteor Eawag_BDD PathPred Zeneth
Step 2: Chromatographic retention time prediction	5	EB XGB LGBM RF BRNN
Step 3: Fragmentation prediction	1	CFM-ID

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

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

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

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

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

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

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

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

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

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

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

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

The Retip package (v0.5.4.) [35] in R (v4.0.4) was used to build the ML-based QSRR models. The models created were based on the same training set as the MLR-based QSRR named ‘EB’ to facilitate cross-comparison (experimental compounds used in training or testing are listed in Supplementary data, Excel spreadsheet #2). The molecular descriptors for each analytical standard were predicted using the RCDK (v3.5.0.) package. As their prediction is not automated and requires special external software, the descriptors used for the EB model were not included in the construction of the ML-based QSRR models. After cleaning missing values, this resulted in 146 molecular descriptors (listed in Supplementary data, Excel spreadsheet #3) used for constructing the models. Four ML algorithms were used: XGBoost (XGB, an extreme gradient boosting algorithm for trees algorithms), Light Gradient Boosting Machine (LGBM), a random forest (RF, a decision-tree algorithm), and a Bayesian regularized neural network (BRNN). Ten-fold cross-validation was employed for all models [35].

The model performances for RT prediction were evaluated by a set of standard performance criteria calculations found in the literature on evaluation of QSRR models [16, 29]. Thus, the following performance criteria were calculated on the testing set: RMSE (root-mean-square error) (1), MAE (mean absolute error in minutes) (2), R² (coefficient of determination) (3), and $A^{95\%}$ (prediction accuracy with a 95% confidence interval). For the sake of harmonization and comparison between models, $A^{95\%}$ was recalculated for the EB model, following the calculations made by the Retip-package "get.score()". This function uses the "qnorm()" function, bundled as standard with R, in order to find the 95th percentile of a normal distribution whose mean and standard deviation correspond to the prediction errors.

$$(1)\ RMSE = \sum_{i=1}^{n} \sqrt{\frac{(ExpRT_i - PredRT_i)^2}{n}}$$

$$(2)\ MAE = \sum_{i=1}^{n} \frac{|ExpRT_i - PredRT_i|}{n}$$

$$(3)\ R^2 = 1 - \frac{\sum i(PredRT_i - ExpRT_i)^2}{\sum i(\overline{ExpRT_i} - ExpRT_i)^2}$$

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

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

In order to test the effectiveness of the fragments prediction and the proposed “blended” strategy, the predicted spectra were compared to experimental spectra for TBZ. The experimental spectra were acquired as described in section 2.1.2., resulting in the “home-ramp” spectra. Four LC-ESI-QToF spectra were compiled from the MassBank database (https://massbank.eu/MassBank): three at the energy levels used by CFM-ID (10eV/20eV/40eV) and one at an “optimized” energy ramp (21.8–32.6 eV). A score of mass spectra similarity between all these spectra was calculated using the OrgMassSpecR package (v0.5-3) in R (v4.0.4). In addition to this calculation, the number of common fragments between mass spectra was investigated. The tolerance used to align the *m/z* values of the spectral fragments was 0.001 *m/z*, which is consistent with the use of mass spectra from HRMS acquisition with a QToF.

2.3. Statistical analysis

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

3. Results and discussion

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

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

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

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

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

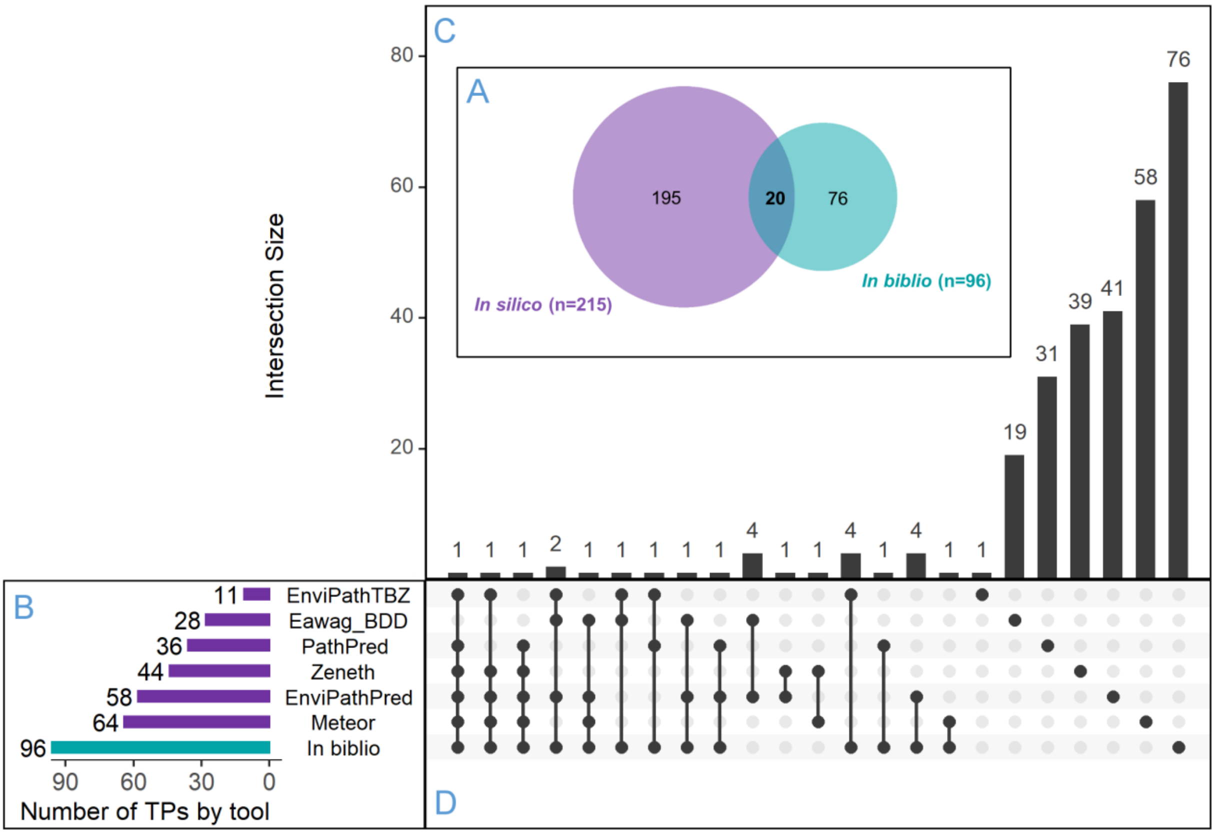


Figure 2. *Results of all-in-silico prediction and in biblio search with (A) the Venn diagram representing the overlap between overall in silico prediction tools and in biblio search of TBZ TPs. (B) Number of transformation products (TPs) and tebuconazole (TBZ) from the six in silico tools (purple) and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

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

3.2. Chromatographic retention time prediction by QSRR models

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

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

Table 2. *Performance values calculated on the testing set (n=69) for the five QSRR models tested in this 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

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

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

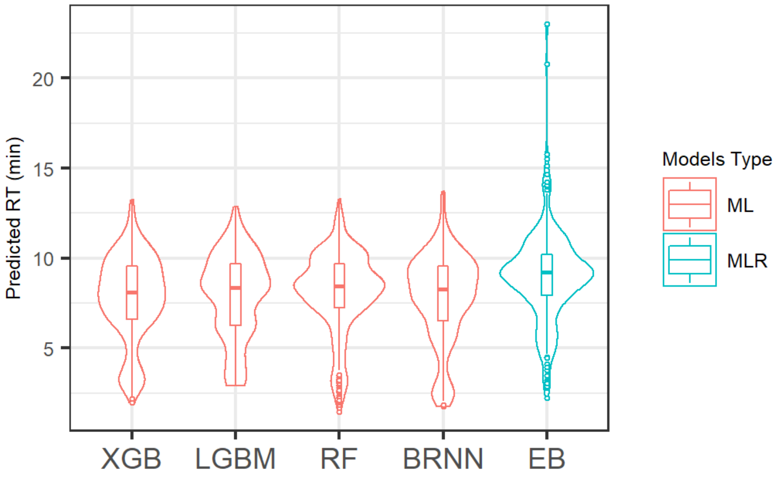


Figure 3. *Violin plots for predicted chromatographic retention times (RT, in minutes) for the five QSSR models (XGB, LGBM, RF, BRNN, and EB) applied to the database of the 291 tebuconazole transformation products. Tools are classified according to model type (machine learning: ML; multilinear regression: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

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

3.3. Tandem-mass spectra prediction by the fragmentation predictor

In order to test the effectiveness of the fragments prediction and the proposed "blended" strategy, we compared the predicted and experimental spectra of TBZ. The similarity scores calculated to evaluate the similarity of the spectra, as well as the number of common fragments between all the spectra discussed here, are presented in Table 3 (all values are compiled in a larger comparison matrix in Table S3). For visual observation of compared mass spectra, their head-to-tail plots are given in supplementary data - figures S2 and S3. The comparison of predicted *vs* experimental TBZ spectra revealed poor similarity scores at the corresponding fixed ionization energies (10, 20, 40 eV). This is connected to the small number of common fragments between the predicted and experimental spectra. In contrast, the comparison of 'blended' predicted spectra *vs* experimental energy-ramped spectra shows good similarity scores (0.84) as well as two common fragments. A low similarity score between the "Home-ramp" spectra and "MassBank-ramp" spectra (0.14), for the same number of common fragments, is explained by the way the score itself is calculated. Indeed, the calculation takes into account the intensity of the fragment, which biases this comparison, given the different ionization energy values of the ramps applied ("Home-ramp": 10–45 eV, "MassBank-ramp": 21.8–32.6 eV). Nevertheless, these calculated scores are important for theoretical comparison, and what matters most for suspect screening analysis in practice is the fragments found in samples corresponding to screened compounds. The highest number of common fragments was found between the experimental and "blended" predicted mass spectra, highlighting its effectiveness. Based on these comparison results for tebuconazole, we suggest the use of a fragmentation with an energy ramp, which revealed more predicted fragments than a fragmentation at fixed energies. To corroborate these findings, this spectra similarity comparison should be performed for a TBZ TP, such as the hydroxytebuconazole or 1,2,4-triazole. However, the MassBank database does not have QToF-acquired spectra of these very specific TPs.

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

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

The MS/MS spectra of the 291 TPs of TBZ incremented in the database developed here were predicted with CFM-ID (v4.0) at three ionization energy levels, which resulted in 873 spectra contained within 291 distinct text files. Applying the blended strategy on the spectra (Figure 4, supplementary data - figure S1) led to a set of 634 fragments compiled in the database. These fragments are often shared by multiple TPs; among these 634 predicted fragments, only 179 (around 30%) were unique. Indeed, the 291 TPs were predicted from a single compound, TBZ, and so most of them logically share similar parts of molecular structures (database available at the following address: https://doi.org/10.57745/Y3JLTV), resulting in similar fragmentation patterns. Furthermore, a single TP may share the same most abundant fragment at two different energy levels, which limits the number of different fragments per compound. As a result, one to three predicted fragments per compound were incorporated in the database. Nevertheless, incrementing the associated fragments of TBZ TPs enhanced the database and is expected to limit mismatches during subsequent suspect screening analysis. For example, TP_096 and TP_220 share the same chemical formula and are predicted to elute at similar RTs (7.61 and 7.34 minutes, respectively), but they disassemble into different fragments according to fragmentation model used (supplementary data – table S2). If this predicted difference in fragmentation pattern is verified during the analysis, it will allow discrimination of the two TPs.

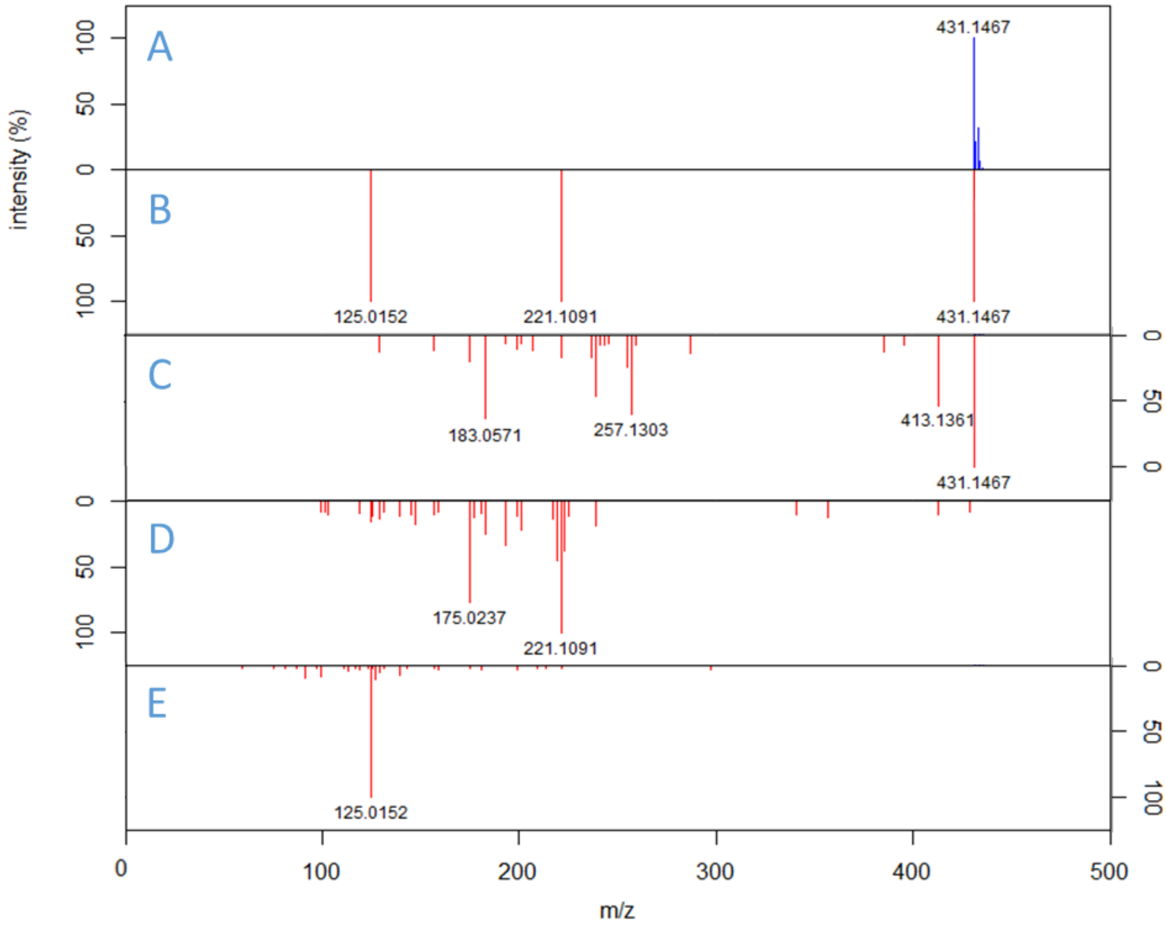


Figure 4. *Head-to-tail plot of different mass spectra of the tebuconazole transformation product TP_095 from the database. (A) Predicted isotope pattern with no ionization energy applied. (B) 'Blended' spectra, emerging from the predicted spectrum of different energy levels used in the fragmentation prediction. (C) Predicted spectra on energy = 10eV. (D) Predicted spectra on energy = 20eV. (E) Predicted spectra on energy = 40eV.*

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

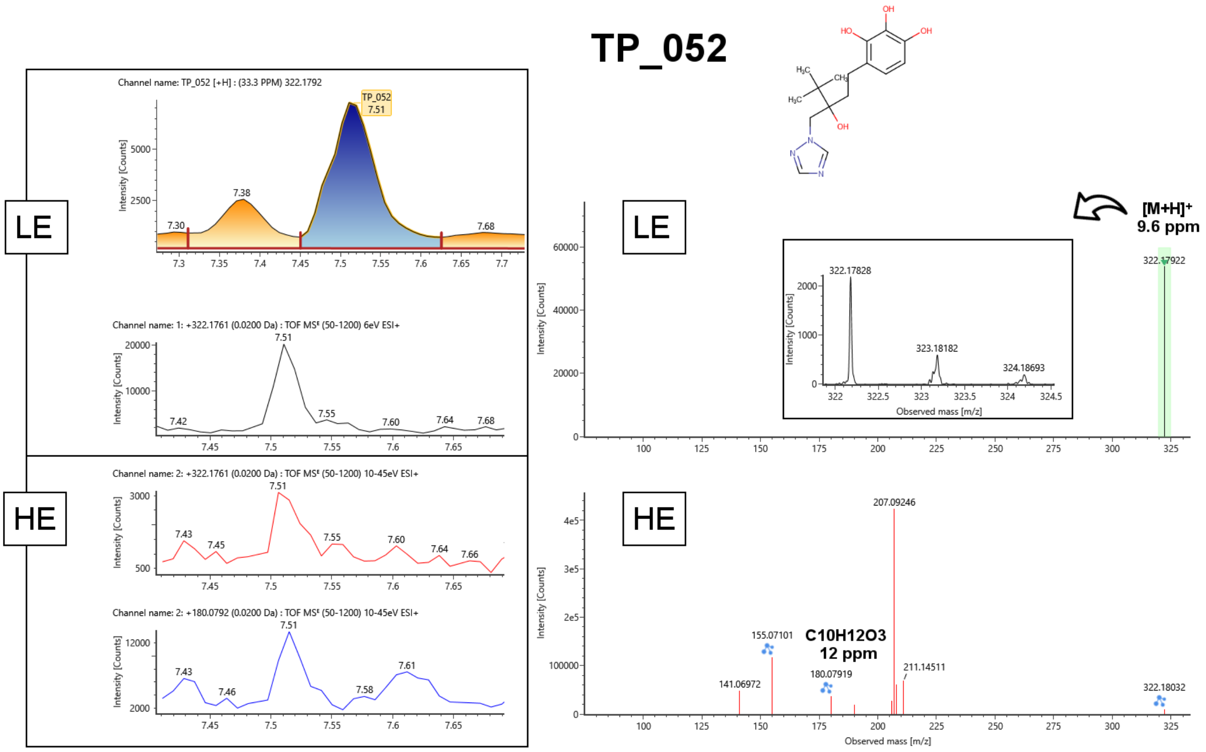
With these points in mind, using CFM-ID predictions and incorporating predicted fragments into the database still increases the elucidation power of the database. Indeed, it provides an additional *a priori* filter on the fragmentation pattern during suspect analysis and thus enables some outliers to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

3.4. Application of the workflow to environmental samples

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

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

Among the 7 different TPs found, 6 come from *in silico* prediction (Figure S5), 4 of which originate from the 'EnviPath' predictor [14]. These results demonstrate the ability of 'EnviPath' to generate accurate TPs for river waters, and justify its exclusive use in recent works [13, 40, 48]. Nevertheless, the application of several other *in silico* transformation predictors, as in the workflow proposed here, led to a more exhaustive detection of TPs. The two remaining TPs from *in silico* predictions were predicted by the transformation predictors 'PathPred' [32] and 'Zeneth' [34]. The whole identification process was enhanced by the use of predicted chromatographic RTs, with the accuracy of the XGB prediction used as a threshold. Using this threshold over the 24 hits among the injections, 16 outliers candidates were eliminated for 7 retained TPs. CFM-ID failed to predict enough fragments of the detected TPs to make it useful in the discrimination of compounds in our suspect screening strategy. This is probably due to the very low concentrations of TPs in these water samples, which resulted in fragment intensities that were below the analytical background. These suspected transformation products could be qualified with a certitude at level 4 ("tentative candidates") to 3 ("unequivocal molecular formula") [9], as for some of them, no fragmentation pattern was detected. In order to reach the level 2B ("diagnostic probable structure"), further search of specific fragments need to be performed. This could be done manually, or with *a posteriori* tool such as MetFrag [19] which make predictions on acquired fragmentation spectra. It is important to note that these detection results have relatively large mass error values for a HRMS instrument, with a mean of 5.7 Da (Supplementary data, Excel spreadsheet #5). This lack of accuracy can cause identification problems, as illustrated for the TP_052 on figure 5. No fragmentation pattern was confirmed for this compound mainly due to mass error value higher than 10 ppm on the predicted fragments. This large mass error values are potentially due to a strong matrix effect in the surface water samples. Nonetheless, targeted analysis operated on a liquid chromatography – tandem mass spectrometry (UHPLC TQ-XS, Waters) confirmed the presence of tebuconazole in the same samples. These results highlight the effectiveness of the proposed workflow in the search for unknown TPs in environmental matrices. Applied on TBZ, the created database of TPs was used on a set of previously analyzed surface water samples, and led to the detection of 6 previously-unseen TPs for this matrix.



(A) (B)

Figure 5. *Tentative Identification of TP_052. (A) Extracted ion chromatogram (EIC) of protonated TP_052 at Low Energy (LE) with a 33 ppm mass error window (resulting from the UNIFI treatment), and at a 0.02 Da mass error window (from a manual extraction). EIC of predicted fragments of TP_052 at High Energy (HE) with a 0.02 Da mass error window. (B) Mass and detected isotopic pattern of TP_052, on LE and HE mass spectra generated by UNIFI. Blue symbols on HE spectra show which fragment is taken in account in fragmentation prediction that UNIFI operates.*

4. Conclusions

This study proposed a comprehensive workflow for the implementation of detailed and ready-to-use databases to support suspect screening analyses of unknown compounds in agricultural watersheds. This novel workflow, combining several *in silico* tools, was applied on tebuconazole. It allowed the creation of a database of 291 tebuconazole transformation products, incremented with their predicted chromatographic retention times and fragment patterns.

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

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

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

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

Acknowledgments

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[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 of machine learning-based retention time prediction for suspect screening of pesticides and pesticide transformation products in LC-HRMS, Chemosphere, 271 (2021) 129447.

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

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

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

[44] A.D. McEachran, K. Mansouri, S.R. Newton, B.E.J. Beverly, J.R. Sobus, A.J. Williams, A comparison of three liquid chromatography (LC) retention time prediction models, Talanta, 182 (2018) 371-379.

[45] M. Ibáñez, V. Borova, C. Boix, R. Aalizadeh, R. Bade, N.S. Thomaidis, F. Hernández, UHPLC-QTOF MS screening of pharmaceuticals and their metabolites in treated wastewater samples from Athens, Journal of Hazardous Materials, 323 (2017) 26-35.

[46] R.W. Becker, D.S. Araújo, C. Sirtori, N.P. Toyama, D.A. Tavares, G.A. Cordeiro, S.F. Benassi, A.C. Gossen, B. do Amaral, Pesticides in surface water from Brazil and Paraguay cross-border region: Screening using LC-QTOF MS and correlation with land use and occupation through multivariate analysis, Microchemical Journal, 168 (2021) 106502.

[47] B. Mathon, A. Dabrin, I. Allan, S. Lardy-Fontan, A. Togola, J.-P. Ghestem, C. Tixier, J.-L. Gonzalez, M. Ferreol, L. Dherret, A. Yari, L. Richard, A. Moreira, M. Eon, B. Delest, E. Noel-Chery, M. El Mossaoui, E. Alasonati, P.-F. Staub, N. Mazzella, C. Miège, Surveillance prospective – évaluation de la pertinence des échantillonneurs intégratifs passifs (EIP) pour la surveillance réglementaire des milieux aquatiques, Rapport AQUAREF 2020, (2020) 172.

[48] J. Zhou, D. Wang, F. Ju, W. Hu, J. Liang, Y. Bai, H. Liu, J. Qu, Profiling microbial removal of micropollutants in sand filters: Biotransformation pathways and associated bacteria, Journal of Hazardous Materials, 423 (2022) 127167.