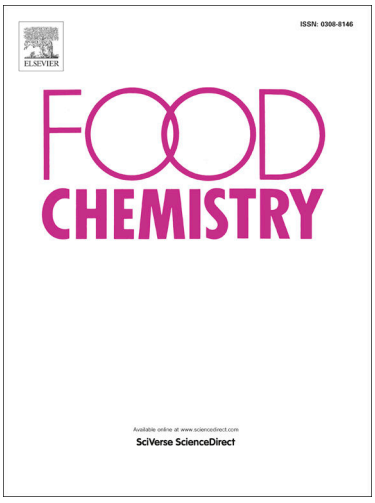
Journal Pre-proofs



Foodinformatic prediction of the retention time of pesticide residues detected in fruits and vegetables using UHPLC/ESI Q-Orbitrap

Cristian Rojas, José F. Aranda, Elisa Pacheco Jaramillo, Irene Losilla, Piercosimo Tripaldi, Pablo R. Duchowicz, Eduardo A. Castro

PII: S0308-8146(20)32216-0
DOI: https://doi.org/10.1016/j.foodchem.2020.128354
Reference: FOCH 128354

To appear in: *Food Chemistry*

Received Date: 16 May 2020
Revised Date: 14 September 2020
Accepted Date: 7 October 2020

Please cite this article as: Rojas, C., Aranda, J.F., Pacheco Jaramillo, E., Losilla, I., Tripaldi, P., Duchowicz, P.R., Castro, E.A., Foodinformatic prediction of the retention time of pesticide residues detected in fruits and vegetables using UHPLC/ESI Q-Orbitrap, *Food Chemistry* (2020), doi: https://doi.org/10.1016/j.foodchem.2020.128354

Foodinformatic prediction of the retention time of pesticide residues detected in fruits and vegetables using UHPLC/ESI Q-Orbitrap

Cristian Rojas[a,*], José F. Aranda[b], Elisa Pacheco Jaramillo[a], Irene Losilla[c], Piercosimo Tripaldi[a], Pablo R. Duchowicz[b,*], and Eduardo A. Castro[b]

[a] Grupo de Investigación en Quimiometría y QSAR, Facultad de Ciencia y Tecnología, Universidad del Azuay, Av. 24 de Mayo 7-77 y Hernán Malo, Cuenca, Ecuador

[b] Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas (INIFTA), CONICET, UNLP, Diag. 113 y 64, C.C. 16, Sucursal 4, 1900 La Plata, Argentina

[c] Departamento de Ciencias Biomédicas (Área de Microbiología), Facultad de Ciencias, Universidad de Extremadura, Badajoz 06071, Spain.

*Corresponding authors. E-mail: crojasvilla@gmail.com (Cristian Rojas); pabloducho@gmail.com (Pablo R. Duchowicz)

Abstract – The present work describes the development of an *in silico* model to predict the retention time (t_R) of a large Compound DataBase (CDB) of pesticides detected in fruits and vegetables. The model utilizes ultrahigh-performance liquid chromatography electrospray ionization quadrupole-Orbitrap (UHPLC/ESI Q-Orbitrap) mass spectrometry (MS). The available CDB was properly curated, and the pesticides were represented by conformation-independent molecular descriptors. In an attempt to improve the model predictions, the best four MLR models obtained were subjected to a consensus analysis. The optimal model was evaluated by means of the coefficient of determination and the residual standard deviation in calibration, validation, and prediction, along other internal and external validation criteria to accomplish the guidelines defined by the Organization for Economic Co-operation and Development. Finally, the *in silico* model was applied to predict the t_R of an external set of 57 pesticides.

Keywords: pesticide residues; fruits and vegetables; QSPR; consensus analysis, foodinformatics

1. Introduction

A pesticide is any substance or mixture of substances that aims to prevent, destroy, repel or control a pest. Pesticides are used as plant growth regulators, defoliants or desiccants, as well as nitrogen stabilizers. Thus, these compounds are used to control various pests and transmitters of diseases, such

as mosquitoes, ticks, rats and mice. Pesticides are also used in agriculture to control weeds, insect infestation and diseases (FAO, 2019). In some cases, pesticides generate residues, which are the substances that may remain in food after the use of the pesticides on crops; and therefore, these residues may be incorporated into the food chain. Many international bodies and countries are extremely concerned about pesticide residues. The tool used to guarantee the safety of consumers is the mandatory establishment of a Maximum Residue Limit (MRL). A MRL is the maximum amount of pesticide residue that is legally allowed in food (both inside and on the surface) resulting from the application of a pesticide in accordance with good agricultural practices. Adherence to the MRL is a guarantee of safety taking into account the best scientific information of the adverse health effects for the population, including vulnerable groups (FAO, 2019).

The gas and liquid chromatography (GC and LC) mass spectrometry (MS) techniques are widely applied for the determination of pesticide residues in food products (Poma *et al.*, 2019), particularly by applying the quick, easy, cheap, effective, rugged, and safe (QuEChERS) method (Anastassiades *et al.*, 2003; Lehotay *et al.*, 2010). On the other hand, in recent years several authors have demonstrated the importance of the ultrahigh-performance liquid chromatography and electrospray ionization quadrupole Orbitrap high-resolution (UHPLC/ESI Q-Orbitrap) mass spectrometry (MS) for the determination of pesticide residues in diverse samples of raw food products (Vu-Duc *et al.*, 2019; Wang *et al.*, 2019) and processed foods (Jia *et al.*, 2014). All of these methods generate analytical responses called retention times (t_R) or retention indices (I). The t_R is the primary parameter obtained in a chromatography system for peak identification, which measures the time required from the injection of the sample in the stationary phase until compound elution. This parameter considers the maximum (apex) of the peak belonging to a particular pesticide. The t_R for a given compound is not fixed, since many factors affect its determination; for instance, the mobile phase flow rate, temperature differences in the oven and the column, as well as column length and column degradation (Vu-Duc *et al.*, 2019).

The quantitative structure-property relationships (QSPRs) theory is usually employed for complementing experimental results from chemicals, as well as to provide reliable predictions when experimental data are not available (data gap filling). Thus, QSPRs are powerful mathematical tools that establish a predictive quantitative relationship between a property (for instance retention time) for a series of molecules (pesticides) and the chemical information provided by the molecular descriptors (Dearden, 2016; Kaliszan, 2007). Through the years, there has been an increased interest among researchers to use this approach, since it is useful to predict the t_R or I of un-evaluated and un-synthetized compounds and to prepare and optimize chromatographic experiments in order to separate complex mixtures and identify potential drug candidates from synthesized or computer-

designed chemicals. In addition, this approach enables the elucidation of the molecular mechanisms of retention phenomena in diverse stationary phases along with the design of new phases with required properties as well as to facilitate protein identification in proteomics studies (Kaliszan, 2007). Thus, several QSPR studies were reported in the literature to predict the t_R of pesticide residues (Dashtbozorgi *et al.*, 2013; Torrens & Castellano, 2014; Zdravković *et al.*, 2018). Our research group has also been interested in QSPR studies for the prediction of chromatographic retention indices in the field of food science (foodinformatics) (Rojas *et al.*, 2019; Rojas *et al.*, 2018), as well as the *in silico* modeling of the water solubility of pesticides (Fioressi *et al.*, 2019).

Consequently, in this work, an *in silico* model based on the QSPR approach was developed to predict the t_R for 823 pesticide residues identified in fruits and vegetables by means of UHPLC/ESI Q-Orbitrap in the Hypersil Gold column. In order to make the model applicable, the five principles established by the Organization for Economic Co-operation and Development (OECD, 2014) was followed. Pesticides were represented by conformation-independent molecular descriptors and fingerprints. For the development of the ordinary least squared (OLS) models the V-WSP unsupervised variable reduction and the replacement method (RM) descriptor subset selection were combined. In an attempt to improve the model predictions, the best four models obtained were subjected to a consensus analysis. The optimal model was thoroughly evaluated by several internal and external validation approaches, along with the applicability domain assessment. Additionally, the mechanistic interpretation of the molecular descriptors used to predict the t_R of the pesticide residues was given. Finally, the model was used to predict the retention time for an external set of pesticides and metabolites for which the t_R was not previously reported. To the best of our knowledge, no foodinformatic studies have been conducted for the prediction of retention times measured by the Hypersil Gold stationary phase for a large dataset of pesticide residues detected in fruits and vegetables.

2. Materials and Methods

2.1. Dataset description

In 2019, Wang *et al.* developed a large Compound DataBase (CDB) of 845 pesticides and their metabolites (Wang *et al.*, 2019). These authors used five fruits (apple, banana, grape, orange and strawberry) and five vegetables (carrot, potato, tomato, broccoli, and lettuce) for samples to determine pesticide residues by means of ultrahigh-performance liquid chromatography electrospray ionization quadrupole-Orbitrap (UHPLC/ESI Q-Orbitrap) mass spectrometry (MS). The UHPLC/ESI Q-Orbitrap system was composed of an Accela 1250 LC pump and an Accela open autosampler integrated with a Q Exactive mass spectrometer from Thermo-Fisher Scientific (Germany). They

compared different LC methods to improve sensitivity, and to obtain better chromatographic resolution. Thus, the Hypersil Gold selectivity column (100 × 2.1 mm, 1.9 μm), and the guard column Accucore aQ (10 × 2.1 mm, 2.6 μm) Defender cartridge were used, (both of them from Thermo Scientific, USA). This silica-based column is able to analyze low concentrations of pesticides in foods (i.e., analysis of impurities). A 4 mM ammonium formate and 0.10% formic acid in water (mobile phase A), and 4 mM ammonium formate and 0.10% formic acid in methanol (mobile phase B) were used as mobile phases with a gradient profile. The temperature of the UHPLC column was fixed at 45 °C, while the temperature of the autosampler was set at 5 °C. A 5 μL volume was used for the sample injection using a run time of 14 min. For each pesticide, the experimental retention time (t_R) was obtained from the chromatograms of a full MS scan based on the exact masses. During the retention time alignment, the t_R of 3-hydroxycarbofuran, a stable and well-characterized compound, was used as a reference standard. The experimental t_R was measured with a retention time tolerance of ± 0.5 min.

In a first step of the data curation, we verified the correct match between the pesticide name and the reported chemical formula. It was found that the formula for the *Fumesate* pesticide ($C_{11}H_{14}O_5S$) corresponds to *ethofumesate-2-hydroxy* (PubChem CID 536079), a *ethofumesate* metabolite; while the formula for the *Pyrethrin* pesticide ($C_{22}H_{28}O_5$) corresponds to *Pyrethrin II* (PubChem CID 5281555, CAS 121-29-9) (MacBean, 2012). Consequently, the *Pyrethrin II* and the *ethofumesate-2-hydroxy* metabolite were used. Moreover, ambiguous pesticides were excluded; that is, compounds having discrepancies between the name and the reported molecular formula. These included: 1) *Dinotefuran metabolite DN phosphate* ($C_7H_{15}N_3O$); 2) *Dodine* ($C_{13}H_{29}N_3$), 3) *Fentin* ($C_{18}H_{16}OSn$), and 4) *N-1-Naphthylacetamide* ($C_{10}H_7CH_2CONH_2$). On the other hand, eight pesticides were analyzed as fragments or metabolites (Wang *et al.*, 2019): *Aldicarb* (C_5H_9NS), *Chlorpropham* ($C_7H_6ClNO_2$), *Demeton-S-sulfone* ($C_6H_{15}O_5PS_2$), *Dialifos* ($C_{10}H_6NO_2Cl$), *Fentrazamide* ($C_{10}H_{16}O_2N_2$), *Isoprocarb* ($C_9H_{12}O$), *Methoxyfenozide* ($C_{18}H_{20}N_2O_3$), and *Bifenazate metabolite D23-15*. Since the exact nature of the kind of fragments that were used was unknown, and in order to avoid the use of wrong structures, these compounds were excluded in the initial analysis. However, these pesticides along with all the available fragments of these compounds were included in an external dataset for analysis.

2.2. Molecular structure visualization and dataset curation

The HyperChem version 8.0 (Hypercube) was used to draw and display the chemical structure of the 833 pesticides or their metabolites selected for this study. Since molecular structures available in chemistry publications and/or public and commercial databases are not exempt from errors, a

molecular structure curation was performed in order to verify the correctness of the inputs. Chemical curation constitutes a fundamental role during the development of a QSPR model because the presence of errors in the compound structures (i.e., lacking an atom, misplacing of atoms or swapping functional groups) influence the molecular descriptor calculation, which results in a detrimental effect on model performance; that is, differences between the predicted property and the expected value (Fourches *et al.*, 2010).

The new generation alvaMolecule software (Alvascience, 2020b) was used for pesticide curation. Thus, 60 pesticides were identified with unusual valence, one molecule with total charge, 35 structures exhibiting charged atoms, 4 with non-standard atom sets (H, C, N, O, P, S, F, Cl, Br and I), and 62 pesticides with no aromatic ring standardization. These pesticides were pretreated applying the following criteria implemented in alvaMolecule: standardize benzene rings into aromatic form, convert unusual covalent bonds to ionic forms, add charge to quaternary nitrogen atom, remove/add exceeding/missing hydrogens, and standardize nitro, azide and diazo groups. Since conformational analysis or energy minimization were not preformed, the clear chirality and clear bond direction options were applied in order to obtain the canonical SMILES (simplified molecular input line entry system) notation of each pesticide. In addition, the correctness of the chemical structures was verified in the PubChem library (Kim *et al.*, 2019) via an option implemented in alvaMolecule, as well as the PubChem CID, and the CAS registry number for each pesticide.

Then, the pesticide name, PubChem CID, CAS registry number, chemical formula, canonical SMILES, and the experimental retention times were merged into KNIME (Berthold *et al.*, 2008) to filter and curate the dataset. Initially, the CAS number was used as a filter criterion to identify three pairs of duplicated molecules; for instance, 1) *3,4,5-Trimethylphenyl methyl carbamate* and *Trimethacarb* (CAS 2686-99-9), 2) *Allethrin* and *Bioallethrin* (CAS 584-79-2), and 3) *Secbumeton* and *Sumitol* (CAS 26259-45-0). Subsequently, the criterion was set up to the canonical SMILES so as to identify seven pairs of pesticides exhibiting the same SMILES notation: 1) *3-Hydroxycycloate, cis-* and *3-Hydroxycycloate, trans-*; 2) *4-Hydroxycycloate, cis-* and *4-Hydroxycycloate, trans-*; 3) *Azoxystrobin* (CAS 131860-33-8) and *Azoxystrobin Z metabolite* (CAS 215934-32-0); 4) *Bioresmethrin* (CAS 28434-01-7) and *Resmethrin* (CAS 10453-86-8); 5) *Bromuconazole, cis-* and *Bromuconazole, trans-*; 6) *Esfenvalerate* (CAS 66230-04-4) and *Fenvalerate* (CAS 51630-58-1); and 7) *Fenbuconazole metabolite RH-9129* and *Fenbuconazole metabolite RH-9130*. For these duplicated pesticides (or metabolites), identified either by CAS number or canonical SMILES, the average t_R was used for the *in silico* modelling. Consequently, 823 structures were submitted in order to develop the QSPR model. Refer to Table S1 for details of the cured dataset.

2.3. Molecular descriptors calculation

Molecular descriptors (MDs) are numerical quantities (or results of some standardized experiments) obtained from logical and mathematical algorithms applied to a symbolic representation of chemicals (Todeschini & Consonni, 2009). MDs are the independent variables used to developed an *in silico* model. In order to develop a conformational independent QSPR model, 3,843 conformation-independent molecular descriptors were calculated along with 166 MACCS fingerprints in the new generation alvaDesc software (Alvascience, 2020a). In addition, 37 descriptors were calculated in DataWarrior (Sander *et al.*, 2015), 1,444 conformation-independent descriptors and 12,854 molecular fingerprints were calculated in the PaDEL-Descriptor freeware (Yap, 2011). A total of 271 descriptors were available in the cheminformatics functionality of the Chemistry Development Kit (CDK) library implemented in R, which is called RCDK (Guha, 2007).

Along with the computation of independent molecular descriptors, flexible molecular descriptors were computed in the CORAL freeware (http://www.insilico.eu/coral/). This program permits three structural representation (SR) approaches: chemical graphs, SMILES notation, and a hybrid between chemical graphs and SMILES. When using chemical graphs, it is possible to use the hydrogen-suppressed graph (HSG), hydrogen-filled graph (HFG) or a graph of atomic orbitals (GAO). Within the CORAL freeware, a QSPR model was quarried that correlated the experimental t_R and an adequate flexible descriptor (*DCW*) by means of a single-variable linear regression. The *DCW* is based on the summation of special coefficients called correlation weights (CW), calculated for each structural attribute (SA) type in the training set, which are obtained by means of the Monte Carlo (MC) simulation. The *DCW* descriptor depends on the threshold value (T) and the number of epochs (or iterations) used to optimize the algorithm. The T value defines uncommon SMILES attributes that do not contribute in predicting the property. Only SMILES attributes located above the T SMILES notations of the training set were classified as active. In this work the T value was set in the range from 1 to 2, and 20 as the maximum number of epochs (N).

2.4. Dataset splitting

The reliability of a *in silico* model is related to its predictive accuracy; that is, the ability to be used to predict the property of an external set of pesticides which were not considered during the calibration of the model. Moreover, a similar structure-property relationship during the splitting of the dataset has been stated to be an appropriate strategy in order to guarantee that the chemical space defined by the molecules in the training set should be representative of the validation and test set compounds. One of the strategies proved to achieve this goal is the Balanced Subsets Method (BSM) (Rojas *et al.*, 2015), which was applied elsewhere in foodinformatic studies when dealing with retention indices of

volatile organic compounds (VOCs) detected by SPME-GC-MS (Rojas *et al.*, 2019; Rojas *et al.*, 2018). In brief, the BSM approach creates clusters of molecules based on the *k*-means cluster analysis (*k*-MCA) by using conformation-independent molecular descriptors and the experimental retention time. The use of this kind of descriptors avoids the effect of the conformational analysis and the geometry optimization method used for calculating 3D descriptors and fingerprints. In order to guarantee the interpolation of the validation and test sets into the structure-property space of the training set, pesticides exhibiting the minimum and maximum t_R are automatically included in the training set. Subsequently, the algorithm creates a reduced matrix by removing the linearly dependent descriptors and the remaining ones are autoscaled. Then, a defined number of clusters (called n^0_{train}) are created by means of the *k*-MCA using the Euclidean distance as a distance metrics. The training set (n_{train}) includes the nearest object to the centroid in each n^0_{train} cluster and the pesticides with minimum and maximum values of the t_R. The validation set is defined following the same workflow as described for the training set, and the remaining molecules constitute the test set. The algorithm is repeated several times (number of iterations) in such a way as to minimize the distance among objects in the multidimensional space. Thus, the BSM method provides a balanced structure-property representation in the training, validation and test sets.

2.5. Development of the *in silico* model

2.5.1 Molecular descriptors pretreatment

In a first attempt to develop the QSPR model, constant descriptors, near constant descriptors, descriptors with both at least one missing value and all missing values were excluded from the initial pool of variables calculated in alvaDesc, DataWarrior, PaDEL-Descriptor and RCDK.

2.5.2. Molecular descriptors reduction

To reduce the size of the pool of MDs, the unsupervised variable reduction method based in the algorithm proposed by Wootton, Sergent, and Phan-Tan-Luu (V-WSP) was considered (Ballabio *et al.*, 2014). The idea behind the V-WSP approach is to reduce the presence of redundancy, multicollinearity, and noise in the initial pool of MDs by selecting an optimal pool of descriptors in such a way that they show a minimal correlation (defined by the user) from each other in the multidimensional space.

2.5.3. Molecular descriptors selection

The supervised selection of MDs was carried out by means of the replacement method (RM) variable subset selection (Duchowicz *et al.*, 2006), in order to find an optimal pool of descriptors. This optimal subset defines a parsimonious and predictive multiple linear regression (MLR) based on the ordinary least squares (OLS) by minimizing (optimizing) the residual standard deviation (*s*) estimator (Todeschini & Consonni, 2009). To this end, the RM randomly starts with a user-defined pool of descriptors *d* (seed) from the initial dataset of *D* variables, then each descriptor is replaced by the remaining ones (except the descriptors previously replaced) one at time, in such a way as to replace variables with the greatest relative error in their coefficients. Thus, the best model of each replaced descriptor is retained and becomes a new seed (path) for the subsequent replacements (except descriptors replaced in previous steps). In this way, the RM approach explores all the *d* paths, and is able to converge to the results achieved by the all subset method (ASM), although RM requires much less computational cost.

2.6. Consensus analysis and model validation

Consensus modeling is a strategy used to improve the predictive ability of a collection of QSPR models obtained during the supervised selection. In brief, an individual QSPR model might underestimate some predictions while overestimating other ones; on the other hand, consensus modeling considers a collection of models that could provide better predictions than single models. In this work, four different approaches available in the Intelligent Consensus Predictor (ICP) MLR tool (Roy *et al.*, 2018) have been applied: simple average of predictions (CM0), average of predictions from the qualified individual models (CM1), weighted average predictions (WAPs) from qualified individual models (CM2), and the best selection of predictions (compound-wise) from qualified individual models (CM3).

To evaluate the predictive performance of the best model, several statistical parameters were checked for both internal and external validation. In the cross-validation step, the leave-one-out (loo) and leave-many-out (lmo) procedures were applied. The absence of chance correlation in the model was evaluated by means of the Y-randomization technique (Rücker *et al.*, 2007), by permuting (randomly scrambling) the experimental t_R 10,000 times. The robustness of the *in silico* model was also controlled using the criteria proposed by Golbraikh and Tropsha (Golbraikh & Tropsha, 2002). Since the merit of a QSPR model is related to its ability to be used to correctly predict the property of the test set molecules (which were never considered during the calibration) the statistical parameters from the test set (R^2_{test} and s_{test}) were used to measure the predictive capability of the model. In addition, the Q^2_{F1}, Q^2_{F2} and Q^2_{F3} external validation criteria was calculated to assess the predictive ability of

the QSPR model (Todeschini *et al.*, 2016). All these parameters were used to avoid the selection of an overoptimistic and perhaps a wrong QSPR model.

2.7. Applicability Domain (AD) assessment

The applicability domain is a theoretical region in the chemical space defined by the descriptors (Hat matrix) in the calibrated QSPR model. Then, reliable predictions of the test set molecules are restricted to only pesticides falling inside this theoretical region (also called the interpolation chemical space); that is, those compounds that fall within this space are structurally similar to compounds of the training set. Among the diverse approaches reported in the literature for defining the AD of QSPR models, the leverage measure was used to verify whether any pesticide in the test set lies within or outside the theoretical region of the chemical space (Sahigara *et al.*, 2012). This approach is proportional to the Hotellings T^2 statistic and the Mahalanobis distance, and measures the distance of each test query to the centroid of the training molecules defined by the Hat matrix (**X** matrix of descriptors only). A warning leverage (threshold value) is set as $h^* = 3p/n$, where p is the number of parameters in the model and n is the number of training set compounds. Then, the leverage value of each test set pesticide (h_i), which is an indicator of the contribution on the predicted value (expected value), is compared to this threshold following this simple rule: if the $h_i \leq h^*$, the prediction of the query compound could be considered reliable (i.e., it is a model interpolation). Otherwise its predicted t_R is unreliable due to a model extrapolation ($h_i > h^*$); that is, the query compound is structurally distant from the centroid of the model. The AD of the external set of pesticides designed for the application of the *in silico* model was also checked.

2.8. Mechanistic interpretation

The mechanistic interpretation of the *in silico* model is an important requirement for the use of a QSPR model for regulatory purposes. It is related to the possibility of establishing a causality between a chemical (pesticide) described by the molecular descriptors and the corresponding experimental property (retention time) (Thoreau, 2016). For this purpose, in a QSPR model based on a multiple linear regression model, the absolute value of the standardized coefficient of each *j*th molecular descriptors (b_j^s) provides the importance (degree of contribution) of such descriptor in predicting the experimental property. Thus, it is possible to sort these standardized coefficients in a decreasing way, which correspond to the rank of the degree of contribution. Then, an explanation of each molecular descriptor and how it is related to the retention time is performed in terms of the definition of the descriptors (if possible). Since the MDs are defined by different theories, in some cases in an abstract

way, the term "if possible" refers to the difficulty of explaining the meaning of a particular descriptor. Consequently, the mechanistic interpretation contributes significantly to the knowledge of how the molecular descriptors describe the retention time phenomenon.

2.9. Application of the *in silico* model

Since the QSPR model was development keeping in mind the five principles stated by the Organisation for Economic Co-operation and Development (OECD, 2014), an external set of pesticides was designed by including the ambiguous pesticides excluded during the data verification and data curation, as well as some of their metabolites or fragments. Table 1 presents detailed information of these molecules. Thus, the *in silico* model was used in a real predictive setting to assess these external molecules, which will be utilized to identify other kinds of pesticides of particular interest. This external set of pesticides was also curated in the alvaMolecule program following the same workflow previously described for the pesticides in the dataset.

3. Software and code

HyperChem version 8 was used for drawing and displaying chemical structure of the pesticides. Molecular structure of pesticides was verified and curated in the alvaMolecule software. A KNIME workflow implemented by the authors was used for data filtering. Molecular descriptors were computed using alvaDesc, DataWarrior, PaDEL-descriptor, RCDK package implemented in R and CORAL-QSAR/QSPR. The V-WSP variable reduction routine was used in MATLAB language. Partition of the dataset by means of the BSM, supervised descriptor selection through the RM technique, as well as model fitting along with validation were also carried out in MATLAB by means of functions and codes implemented by the authors. Consensus analysis was carried out in the Intelligent Consensus Predictor (ICP) tool.

Table 1 should be inserted around here

3. Results and Discussion

3.1. Development of the *in silico* model

Initially, constant and near constant descriptors were excluded, as well as those with at least one missing value for each block of descriptors provided by each program. Thus, 2,515 alvaDesc descriptors, 37 DataWarrior descriptors, 5,702 PaDEL descriptors, and 125 RCDK descriptors were retained. Subsequently, the V-WSP unsupervised variable reduction was applied at a threshold value of 0.95 over each block of descriptors in order to reduce the ones with greatest correlation

(redundancy) in the initial datasets. Using these criteria, 1,379 alvaDesc descriptors, 30 DataWarrior descriptors, 3,314 PaDEL descriptors, and 95 RCDK descriptors were retained. Subsequently, the BSM was utilized in order to split the dataset of 823 pesticides represented by the conformation-independent MDs described above into a training set, a validation set and a test set. The training and validation sets were formed by 275 molecules, and the remaining 273 compounds constituted the test set (refer to Table S1 for splitting assignments). The CORAL-QSAR/QSPR software was used to optimize the *DCW* flexible descriptor by maximizing both the R^2_{train} and R^2_{val} in order to choose the most effective attributes for each structural representation (SR). The statistical parameters for the training set ($R^2_{train} = 0.83$ and $s_{train} = 0.91$) and the validation set ($R^2_{val} = 0.70$ and $s_{val} = 1.02$) suggested an appropriate descriptor for predicting the t_R. The *DCW* descriptor included HFG representations, which considered two variable types and 144 active attributes derived from the SR.

Table 2 should be inserted around here

Afterwards, the selection of MDs was carried out by means of the RM variable subset selection on the descriptors provided after V-WSP reduction. The RM was initially applied separately on each block of molecular descriptors; then, the best descriptors of each block were merged into a new set containing 80 MDs, included the optimal *DCW* flexible descriptor. Then, the RM was applied again to find the most suitable pool of descriptors that constituted the *in silico* model. During the descriptor selection, the training set was used to calibrate the models, while the validation set helped to avoid overfitting the models. The RM optimized the residual standard deviation (*s*) in the training and validation sets. For the selection of the best four models, a multicriteria approach was applied by considering the balanced ratio between the training set (R^2_{train} and s_{train}) and the validation set (R^2_{val} and s_{val}), as well as the number of *d* descriptors according to the Ockham's razor principle of parsimony (Hoffmann *et al.*, 1996). Table 2 summarizes the best MLR models containing from 2 to 5 conformational-independent descriptors selected by the RM approach.

Table 3 should be inserted around here

In an attempt to improve the predictive capability of the individual QSPR models, a consensus modeling was applied considering the CM0, CM1, CM2 and CM3 approaches. Table 3 summarizes the test set results found for both individual and consensus models, which clearly indicated that their prediction quality was acceptable. The best model based on the minimum $MAE_{95\%}$ was the IM4 (the

subscript 95 % indicates that the Q^2_{F1}, Q^2_{F2}, and *MAE* parameters were recalculated after removing the 5 % of high residual pesticides). This fact could be related to the consensus-like modeling during the RM supervised selection, i.e., the fusion of the best descriptors from each program. Thus, a foodinformatic model based on five ($d = 5$) conformation-independent descriptors was retained for further analysis.

$$t_R = 4.02 - 13.98\ Eta\_D\_epsiD + 0.37\ cLogP - 1.84\ Alkyl - Amines + 0.26\ MDEN.22 + 0.14\ DCW \quad \text{(Eq. 1)}$$

$n_{train} = 275$, $R^2_{train} = 0.87$, $s_{train} = 0.81$

$n_{val} = 275$, $R^2_{val} = 0.79$, $s_{val} = 0.82$

$n_{test} = 273$, $R^2_{test} = 0.74$, $s_{test} = 0.85$

Negligible differences for the training, validation and test sets indicated the absence of overfitting and the presence of a predictive *in silico* model. Consequently, the model derived by Eq. 1 was subjected to a more rigorous validation process. The cross-validation approach of leave-one-out ($R^2_{loo} = 0.86$ and $s_{loo} = 0.83$) and leave-many-out ($R^2_{lmo} = 0.82$ and $s_{lmo} = 0.85$) indicated good stability to internal perturbations. In addition, the $R^2_{rand} = 0.01$ and the $s_{rand} = 1.99$ parameters, obtained as the mean of 10,000 models (iterations) for the Y-randomization procedure confirmed the absence of change correlation in the *in silico* model ($R^2_{rand} \ll R^2_{train}$ and $s_{rand} \gg s_{train}$). The model also met the criteria of Golbraikh and Tropsha: $R^2_{loo} > 0.5$ (0.86); $R^2_{test} > 0.6$ (0.74); $1 - R^2_0 / R^2_{test} < 0.1$ (0.000) and $1 - R'^2_0 / R^2_{test} < 0.1$ (0.097); $0.85 \le k(1.00) \le 1.15$ and $0.85 \le k'(0.99) \le 1.15$; and $R^2_m > 0.5$ (0.73). Finally, the $Q^2_{F1} = 0.75$, $Q^2_{F2} = 0.74$ and $Q^2_{F3} = 0.82$ validation criteria also confirmed the predictive power of the *in silico* model.

Since the model accomplished all the cross-validation and external validation criteria, a robust (stable) and predictive conformation-independent *in silico* relationship was obtained to predict the retention time of pesticide residues (their metabolites or fragments) identified in fruits and vegetables samples. Details of the numerical t_R predicted by Eq. 1 are presented in Table S1, while descriptor values for the dataset of 823 pesticides are available in Table S2. Figure 1a shows the relationship between the experimental and predicted retention times obtained with Eq. 1, which clearly suggested a linear

relationship around the perfect fit line; while Figure 1b shows the dispersion plot of the residuals vs. the experimental t_R, which reflected a random distribution of the residuals around the zero line. Since the assumptions behind the OLS estimators in the MLR models were confirmed, a robust and predictive *in silico* model was achieved.

Figure 1 should be inserted around here

The QSPR model was also evaluated to identify possible outliers (i.e., molecules having poorly fitted t_R) by standardizing the residuals of the training set and defining a threshold value of ± 3*s*. Thus, pesticides having a standardized residual greater than this threshold were considered as outliers. There exist four pesticides labeled as outliers: *Carbofuran phenol* (PubChem CID 15278, CAS 1563-38-8), *Chinomethionate* (PubChem CID 17109, CAS 2439-01-2), *Pyribenzoxim* (PubChem CID 178117, CAS 168088-61-7) and *TDCPP* (PubChem CID 26177, CAS 13674-87-8). The correctness of the chemical formula and the experimental retention times were verified in several open libraries and sources, respectively. Since they were found to be correct, this particular behavior could be associated with the diverse factors involved during the analytical measurement. In fact, the UHPLC/ESI Q-Orbitrap analytical technique often requires extensive compound-dependent instrument parameter optimization, as well as a complete set of standards for preparing standard calibration curves for the identification and quantitation of pesticides present in the samples (Wang *et al.*, 2019).

The mechanism of action of the t_R phenomenon presented in Eq. 1 was constituted by four rigid molecular descriptors (*Eta_D_epsiD*, *cLogP*, *Alkyl-Amines* and *MDEN.22*) along with the *DCW* flexible descriptor. The maximum coefficient of determination ($R^2_{ij\max} = 0.68$) indicated a low to moderate correlation between the *cLogP* and the *DCW* descriptor pair, suggesting that descriptors in the model were not collinear. Consequently, each descriptor characterized particular aspects of the retention time phenomenon in the Hypersil Gold stationary phase that succeed when combined with the remaining MDs of the *in silico* model (Eq. 1). Additionally, the degree of contribution of each descriptor in predicting the t_R was analyzed by the standardization of the regression coefficients: *DCW* (0.53) > *cLogP* (0.30) > *Eta_D_epsiD* (0.21) > *Alkyl-Amines* (0.15) > *MDEN.22* (0.12). The sign of each coefficient for the descriptors in Eq. 1 indicated that the *cLogP*, *MDEN.22* and *DCW* descriptors had synergistic effects (positive coefficients) on the prediction of the retention time property, while the *Eta_D_epsiD* and *Alkyl-Amines* exhibited antagonistic effects (negative coefficients). Consequently, pesticides exhibited high retention when increasing the *cLogP*, *MDEN.22* and *DCW* descriptors. In contrast, the t_R of compounds decreased with increasing values of the *Eta_D_epsiD*

and *Alkyl-Amines* descriptors. Table S3 details the references for each molecular descriptor included in the QSPR model.

The *DCW* flexible descriptor was computed from a hydrogen-filled graph considering as attributes the sum of vertex degrees at topological distance 2 (*S2*) relatively to the *k*th vertex, and the nearest neighbors code (*NNC*) relatively to this *k*th vertex (i.e. the contribution of the total number of atoms, as well as carbon and non-carbon atoms). This flexible descriptor considers these two variable types and 144 active attributes derived from the SR. Thus, the synergistic effect of the *DCW* descriptor in predicting the t_R could be related to the degree of branching and complexity of the pesticide molecules, that is, the *DCW* descriptor may describe compounds exhibiting the highest interaction with the silica-based stationary phase.

The calculated octanol-water partition coefficient (*clogP*) was obtained following the fragmental method proposed by Leo and Hansch, where molecular structures were decomposed into fragments (i.e., atoms or polyatomic functional groups) by means of a unique and simple set of rules in order to obtain a unique solution. Then, diverse correction factors were derived from compounds by considering more than one substituent to better estimate the experimental *logP* values. This descriptor considers proximity effects provided by multiple halogenation and groups with hydrogen donors, intramolecular hydrogen-bonds involving O and N atoms, electronic effects in aromatic systems, unsaturation, branching, chains, and rings. Its positive coefficient could be related, on the one hand, to the solvent strength; that is, the ability of water and methanol to elute polar pesticides from the stationary phase. The solvent strength property is characterized, under normal phase conditions, by the Hildebrand's elution strength scale (E^0), as well as to the solvent polarity (Dong, 2019). In fact, the polarity index (*P'*) of the water and methanol, 10.2 and 5.1, respectively, permit pesticides with low *clogP* to have more affinity to interact with the mobile phases, decreasing the retention time (synergistic effect). On the other hand, formate buffers (max. pH range of 2.8 and 4.8) (Agilent Technologies, 2016), which have been commonly used in LC/MS analysis, increase the affinity of the mobile phases to interact with polar groups that are present in the pesticide scaffold (Dong, 2019). Thus, hydrophilic pesticides (low *clogP*) have strong affinity to the mobile phases (aqueous regions), while hydrophobic pesticides (high *clogP*) exhibit better affinity to the stationary phase (hydrophobic region). The usefulness of *clogP* descriptor in QSPR studies regarding the HPLC retention time was summarized elsewhere (Kaliszan, 2007).

The Extended Topochemical Atom (ETA) indices are topological indices calculated from a H-depleted molecular graph, where a vertex is considered to be comprised of a core and a valence electronic environment. In particular, the electronegativity ETA measure (*Eta_epsi*) combines the core count of an atom with its valence electron number (Z^v). Thus, the ETA measure of the hydrogen

bond donor atoms (*Eta_D_epsiD*) characterizes the capacity of a pesticide to interact with the mobile phases. Thus, water acts as a proton acceptor (i.e. interactions through π-π bonds), while methanol acts as both a proton acceptor and donor with pesticides; consequently, the retention time is decreased (Dong, 2019).

The *Alkyl-Amines* is a functional group count descriptor that quantifies the number of amino groups ($R\text{-}NH_2$) in a molecule, except those attached to an aromatic hydrocarbon (Aryl Amines). Amines had been widely used during pesticide manufacturing, and it had been stated that these compounds were difficult to analyze by gas chromatography due to the basicity and the large dipole induced by the amino group in the molecule. The N atom exhibits a lone electron pair that form the ammonium ion NH_4^+. At low pH, more ammonia fragments are converted into NH_4^+ (positively charged) and the t_R of basic pesticides might have been reduced due to the limited interaction with the silanol groups (Si-OH) of the stationary phase (i.e. low adsorption) (Agilent Technologies, 2016). This phenomenon possibly explained the antagonistic effect of this descriptor in predicting the retention time.

The Molecular Distance Edge (MDE) vector considers the geometrical means of the topological distances between carbon atoms, classified as primary ($-CH_3$), secondary ($>CH_2$), tertiary (>CH-) and quaternary (>C<), to compute a 10-dimensional vector descriptor by considering all the possible pair combinations among these carbon types. A particular variation of the MDE vector is obtained when the nitrogen atom is considered instead of the carbon atom. Thus, the *MDEN.22* descriptor measures the distance edge between all secondary nitrogen atoms. The synergistic effect of this descriptor could have been related to the *silanophile* effect; that is, a high affinity of the strong basic amine (in this case described by the secondary nitrogen atoms) for active or acidic silanol groups on the silica surface (Agilent Technologies, 2016), generating slow elution of polar pesticides through the column (Dong, 2019).

Figure 2 should be inserted around here

The applicability domain of the *in silico* model was defined to provide the theoretical space within the predictions of the t_R of new pesticides to show reliability (i.e., interpolations). The leverage approach defined a threshold or warning leverage $h^* = 0.033$, which indicated that predictions were restricted to only pesticides exhibiting a leverage value below this threshold ($h_i < h^*$); otherwise predictions were the result of a substantial extrapolation of the model (i.e., unreliable). In this work, three pesticides felt outside the AD of the model (Figure 2 a-c): *Ethametsulfuron-methyl* ($h_i = 0.048$,

PubChem CID 91756, CAS 97780-06-8), *Aziprotryne* (h_i = 0.068, PubChem CID 3032472, CAS 4658-28-0) and *Spinetoram* (h_i = 0.070, CAS 935545-74-7). The *Ethametsulfuron-methyl* (HRAC & WSSA CODE number 2) is a selective herbicide from the Sulfonylureas family which inhibits the acetolactate synthase (ALS) enzyme; while the *Aziprotryne* (HRAC & WSSA CODE number 5), another herbicide, belongs to the Triazines families and acts by inhibiting photosynthesis at PSIL - Serine 264 Binders. On the other hand, Spinetoram (IRAC MoA classification number 5) is an insecticide of the Spinosyns group, which primarily acts by disrupting the nicotinic/gamma amino butyric acid (GABA)-gated chloride channels (MacBean, 2012). Therefore, we suspected that the predictive limitations of the model were related to herbicidal or insecticidal compounds having the 1,3,5 triazine fragment in their scaffold, as well as complex structures such as the insecticide *Spinetoram*, a mixture of two naturally-occurring spinosyns with activity against a wide range of common insect pests.

Due to the fact that the majority of the pesticides (270 molecules) fell inside this theoretical chemical space, and the CDB considered diverse heterogenous compounds (complex datasets), the model could be generalized to other kinds of pesticides not considered in this study. This ability of a QSPR model to be generalized was not feasible when dealing with databases of only homogeneous families (Rojas *et al.*, 2019; Rojas *et al.*, 2015).

3.2. Application of the *in silico* model

Since we excluded some ambiguous pesticides during the dataset curation, we used these compounds, as well as some of their metabolites or fragments, to develop an external set of 57 compounds to predict their t_R as a test of the QSPR model of Eq. 1. Table 1 summarizes the information and results for the predicted t_R of these pesticides in the Hypersil Gold stationary phase in UHPLC/ESI Q-Orbitrap technique. According to the results presented in Table 1, 54 pesticides belonged to the AD of the model; i.e., they had leverage values below the warning leverage ($h_i < h^*$) that defined the AD of the model, and consequently their t_R were interpolations of the model (reliable). In contrast, the *Dinotefuran metabolite DN phosphate* (h_i = 0.059 and t_R = 1.56), *Tritosulfuron Metabolite 635M03* (h_i = 0.037 and t_R = 4.22) and *Tritosulfuron Metabolite BH635-4* (h_i = 0.036 and t_R = 4.10) pesticides (refer to Figure 2 d-f) had leverage values higher than the leverage threshold ($h_i > h^*$), and consequently their predicted t_R could be considered as a substantial extrapolation of the model (unreliable).

The absolute difference between the predicted and the experimental retention time (Δt_R) reported by Wang et al. (Wang *et al.*, 2019) was used to verify the reliability of the QSPR model. The lower difference was for the *Aldicarb* pesticide (Δt_R = 0.09 min), while the *N-1-Naphthylacetamide*,

Fentrazamide, *Isoprocarb*, *Methoxyfenozide*, *Dialifos*, and *Demeton-S-sulfone* exhibited a Δt_R below one minute. On the contrary, the largest difference corresponded to the *Chlorpropham* (Δt_R = 1.36 min), followed by the *Dodine* (Δt_R = 1.26 min) and *Fentin* (Δt_R = 1.10 min) pesticides. In addition, the same authors published (between 2014 and 2017) experimental t_R for some of the pesticide residues presented in Table 1, which were obtained under similar UHPLC/ESI Q-Orbitrap conditions. The analysis of these results showed that the predicted t_R for the *Aldicarb* (5.41 min) was closely related to the average experimental t_R = 5.42 min (standard deviation s = 0.09 min); while *Chlorpropham* (average t_R = 7.80 min), *Dialifos* (average t_R = 8.80 min), and *Fentrazamide* (average t_R = 8.47 min) had a standard deviation s = 0.07 min. On the other hand, *Methoxyfenozide* (average t_R = 7.79 min) exhibited the lowest standard deviation (s = 0.06); while *Isoprocarb* (average t_R = 6.99 min) exhibited the largest one (s = 0.13 min). Thus, the negligible difference between the predicted and the experimental t_R confirmed the accuracy of the *in silico* model.

Consequently, the foodinformatic model (Eq. 1) developed in this work provides a useful tool for predicting the t_R of pesticides commonly used in raw foods by means of the UHPLC/ESI Q-Orbitrap in the Hypersil Gold stationary phase. In addition, this model could be useful for food chemistry researchers for the rapid screening of retention times of pesticides not considered in this extensive dataset, for which the experimental t_R property is not yet available. Thus, it is possible to predict the t_R for new potential pesticides obtained by *de novo* design, and for which there is no available standard to be used by chromatographers in UHPLC/ESI Q-Orbitrap. Finally, the use of conformation-independent foodinformatic models emerges as a promising approach when dealing with the retention time or retention index of compounds of interest in the field of food chemistry, as well as an approach for the quality control of both raw food materials and by-products (Rojas *et al.*, 2019; Rojas *et al.*, 2018).

4. Conclusions

In this work we developed a foodinformatic model based on the QSPR approach for the retention times of 823 pesticides present in the Compound DataBase (CDB), which were identified in five fruits and vegetables products. The canonical SMILES was used to calculate conformation-independent molecular descriptors and fingerprints in several available software programs. To deal with the huge number of descriptors, the use of the unsupervised variable reduction V-WSP technique permitted the exclusion of either non-informative descriptors. Subsequently, the supervised Replacement Method variable subset selection method was applied to find four suitable models, which were used to perform an intelligent MLR consensus to improve the quality of the predictions.

The optimal model was extensively validated by applying several internal and external protocols, according to the five OECD principles to make it applicable to predict the retention time of an external set of 57 pesticides or fragments. Thus, this conformation-independent QSPR approach could be implemented for food chemistry researchers, particularly chromatographers, working on the pesticide residue identification in raw or processed foods, based on the retention time measured in the Hypersil Gold stationary phase in the ultrahigh-performance liquid chromatography electrospray ionization quadrupole-Orbitrap (UHPLC/ESI Q-Orbitrap) mass spectrometry technique.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors express their gratitude to the National Scientific and Technical Research Council of Argentina [Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)] for the project grant PIP0311; and the Ministry of Science, Technology and Productive Innovation of Argentina [Ministerio de Ciencia, Tecnología e Innovación Productiva (MINCYT)] for the use of the electronic library facilities. Pablo R. Duchowicz and Eduardo A. Castro are members of the Scientific Researcher Career of CONICET.

Appendix A. Supplementary data

Supporting information for this research is available in the Supplementary data section.

References

Agilent Technologies. (2016). *The LC Handbook. Guide to LC Columns and Method Development*. USA.

Alvascience. (2020a). alvaDesc (software for molecular descriptors calculation) version 1.0.22, https://www.alvascience.com.

Alvascience. (2020b). alvaMolecule (software to view and prepare chemical datasets) version 1.0.4, https://www.alvascience.com.

Anastassiades, M., Lehotay, S. J., Štajnbaher, D., & Schenck, F. J. (2003). Fast and easy multiresidue method employing acetonitrile extraction/partitioning and "dispersive solid-phase extraction" for the determination of pesticide residues in produce. *Journal of AOAC international, 86*(2), 412-431.

Ballabio, D., Consonni, V., Mauri, A., Claeys-Bruno, M., Sergent, M., & Todeschini, R. (2014). A Novel Variable Reduction Method Adapted from Space-Filling Designs. *Chemometrics and Intelligent Laboratory Systems, 136*, 147-154.

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., & Wiswedel, B. (2008). KNIME: The Konstanz Information Miner. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme & R. Decker (Eds.), *Data Analysis, Machine Learning and Applications*, (pp. 319-326): Springer Berlin Heidelberg.

Dashtbozorgi, Z., Golmohammadi, H., & Konoz, E. (2013). Support vector regression based QSPR for the prediction of retention time of pesticide residues in gas chromatography–mass spectroscopy. *Microchemical Journal, 106*, 51-60.

Dearden, J. C. (2016). The History and Development of Quantitative Structure-Activity Relationships (QSARs). *International Journal of Quantitative Structure-Property Relationships, 1*(1), 1-44.

Dong, M. W. (2019). *HPLC and UHPLC for Practicing Scientists* (Second ed.). New Jersey: Hoboken: John Wiley & Sons, Inc.

Duchowicz, P. R., Castro, E. A., & Fernández, F. M. (2006). Alternative Algorithm for the Search of an Optimal Set of Descriptors in QSAR-QSPR Studies. *MATCH Communications in Mathematical and in Computer Chemistry, 55*(1), 179-192.

FAO. (2019). Codex Pesticides Residues in Food Online Database. http://www.fao.org/fao-who-codexalimentarius/codex-texts/dbs/pestres/en/.

Fioressi, S. E., Bacelo, D. E., Rojas, C., Aranda, J. F., & Duchowicz, P. R. (2019). Conformation-independent quantitative structure-property relationships study on water solubility of pesticides. *Ecotoxicology and environmental safety, 171*, 47-53.

Fourches, D., Muratov, E., & Tropsha, A. (2010). Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *Journal of Chemical Information and Modeling, 50*(7), 1189-1204.

Golbraikh, A., & Tropsha, A. (2002). Beware of q2! *Journal of Molecular Graphics and Modelling, 20*(4), 269-276.

Guha, R. (2007). Chemical Informatics Functionality in R. *Journal of Statistical Software, 18*(5), 1-16.

Hoffmann, R., Minkin, V. I., & Carpenter, B. K. (1996). Ockham's razor and chemistry. *Bulletin de la Société chimique de France, 133*(2), 117-130.

http://www.insilico.eu/coral/. CORAL-QSAR/QSPR.

Hypercube, I. HyperChem™ Professional version 8.0. http://www.hyper.com.

Jia, W., Chu, X., Ling, Y., Huang, J., & Chang, J. (2014). High-throughput screening of pesticide and veterinary drug residues in baby food by liquid chromatography coupled to quadrupole Orbitrap mass spectrometry. *Journal of Chromatography A, 1347*, 122-128.

Kaliszan, R. (2007). QSRR: Quantitative Structure-(Chromatographic) Retention Relationships. *Chemical Reviews, 107*(7), 3212-3246.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., & Yu, B. (2019). PubChem 2019 update: improved access to chemical data. *Nucleic acids research, 47*(D1), D1102-D1109.

Lehotay, S. J., Son, K. A., Kwon, H., Koesukwiwat, U., Fu, W., Mastovska, K., Hoh, E., & Leepipatpiboon, N. (2010). Comparison of QuEChERS sample preparation methods for the analysis of pesticide residues in fruits and vegetables. *Journal of Chromatography A, 1217*(16), 2548-2560.

MacBean, C. (2012). *The Pesticide Manual: A World Compendium* (Sixteenth ed.): BCPC (British Crop Production Council).

OECD. (2014). Guidance Document on the Validation of (Quantitative)Structure-Activity Relationships [(Q)SAR] Models: OECD Publishing, Paris.

Poma, G., López-García, M., Romero, R., González, A. G. F., & Covaci, A. (2019). Determination of Pesticide Residues in Food of Animal Origin. In J. L. Tadeo (Ed.), *Analysis of Pesticides in Food and Environmental Samples* Second ed., (pp. 207-243).

Rojas, C., Duchowicz, P. R., Tripaldi, P., & Pis Diez, R. (2015). QSPR analysis for the retention index of flavors and fragrances on a OV-101 column. *Chemometrics and Intelligent Laboratory Systems, 140*(0), 126-132.

Rojas, C., Tripaldi, P., Pérez-González, A., Duchowicz, P. R., & Diez, R. P. (2018). A Retention Index-Based QSPR Model for the Quality Control of Rice. *Journal of Cereal Science, 79*, 303-310.

Rojas, C., Duchowicz, P. R., & Castro, E. A. (2019). Foodinformatics: Quantitative Structure-Property Relationship Modeling of Volatile Organic Compounds in Peppers. *Journal of Food Science, 84*(4), 770-781.

Roy, K., Ambure, P., Kar, S., & Ojha, P. K. (2018). Is it possible to improve the quality of predictions from an "intelligent" use of multiple QSAR/QSPR/QSTR models? *Journal of Chemometrics, 32*(4), e2992.

Rücker, C., Rücker, G., & Meringer, M. (2007). Y-Randomization and its variants in QSPR/QSAR. *Journal of Chemical Information and Modeling, 47*(6), 2345-2357.

Sahigara, F., Mansouri, K., Ballabio, D., Mauri, A., Consonni, V., & Todeschini, R. (2012). Comparison of different approaches to define the applicability domain of QSAR models. *Molecules, 17*(5), 4791-4810.

Sander, T., Freyss, J., von Korff, M., & Rufener, C. (2015). DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis. *Journal of Chemical Information and Modeling, 55*(2), 460-473.

Thoreau, F. (2016). 'A mechanistic interpretation, if possible': How does predictive modelling causality affect the regulation of chemicals? *Big Data & Society*, 1-11.

Todeschini, R., & Consonni, V. (2009). *Molecular Descriptors for Chemoinformatics* (Vol. 1). Weinheim: WILEY-VCH.

Todeschini, R., Ballabio, D., & Grisoni, F. (2016). Beware of unreliable Q^2! A comparative study of regression metrics for predictivity assessment of QSAR models. *Journal of Chemical Information and Modeling, 56*(10), 1905-1913.

Torrens, F., & Castellano, G. (2014). QSPR prediction of chromatographic retention times of pesticides: Partition and fractal indices. *Journal of Environmental Science and Health, Part B, 49*(6), 400-407.

Vu-Duc, N., Nguyen-Quang, T., Le-Minh, T., Nguyen-Thi, X., Tran, T. M., Vu, H. A., Nguyen, L.-A., Doan-Duy, T., Van Hoi, B., & Vu, C.-T. (2019). Multiresidue Pesticides Analysis of Vegetables in Vietnam by Ultrahigh-Performance Liquid Chromatography in Combination with High-Resolution Mass Spectrometry (UPLC-Orbitrap MS). *Journal of analytical methods in chemistry, 2019*, 1-12.

Wang, J., Chow, W., Wong, J. W., Leung, D., Chang, J., & Li, M. (2019). Non-target data acquisition for target analysis (nDATA) of 845 pesticide residues in fruits and vegetables using UHPLC/ESI Q-Orbitrap. *Analytical and Bioanalytical Chemistry, 411*, 1421-1431.

Yap, C. W. (2011). PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. *Journal of Computational Chemistry, 32*(7), 1466-1474.

Zdravković, M., Antović, A., Veselinović, J. B., Sokolović, D., & Veselinović, A. M. (2018). QSPR in forensic analysis-The prediction of retention time of pesticide residues based on the Monte Carlo method. *Talanta, 178*, 656-662.

Table 1. External set of pesticides, their metabolites or fragments: name, PubChem CID, CAS registry number, predicted retention times using Eq. 2 for the external set of pesticides in the Hypersil Gold column in the UHPLC/ESI Q-Orbitrap, and available experimental retention times from literature.

name	PubChem CID	CAS registry number	canonical SMILES	t_R	
				predicted	exp
carb	9570071	116-06-3	CNC(=O)ON=CC(C)(C)SC	5.41	5.3
carb sulfone (aldoxicarb)	9570093	1646-88-4	CNC(=O)ON=CC(C)(C)S(C)(=O)=O	4.12	
carb sulfoxide	9568700	1646-87-3	CNC(=O)ON=CC(C)(C)S(C)=O	4.29	
carb oxime	9570092	1646-75-9	CSC(C)(C)C=NO	4.61	
carb nitrile	119417	10074-86-9	CSC(C)(C)C#N	5.25	
carb oxime sulfoxide	9589350	7635-32-7	CS(=O)C(C)(C)C=NO	3.49	
carb nitrile sulfoxide	12628029	14668-28-1	CS(=O)C(C)(C)C#N	4.54	
carb sulfone oxime	518932	---[c]	CC(C)(C=NO)S(C)(=O)=O	3.32	
carb sulfone nitrile	3014848	14668-29-2	CC(C)(C#N)S(C)(=O)=O	4.00	
nazate-diazene D3598	69250380	149878-40-0	COc1ccc(cc1N=NC(=O)OC(C)C)-c1ccccc1	8.79	
89	11943	613-37-6	COc1ccc(cc1)-c1ccccc1	7.29	
ethoxybiphenyl					
30	7103	92-69-3	Oc1ccc(cc1)-c1ccccc1	6.74	
droxybiphenyl					
30S	177718	16063-85-7	OS(=O)(=O)Oc1ccc(cc1)-c1ccccc1	6.41	
droxybiphenyl sulphate					
63	14386780	37055-80-4	COc1ccc(cc1O)-c1ccccc1	6.72	
droxy-4-methoxybiphenyl					
bamate	---[b]	---[c]	COc1ccc(cc1NC(=O)OC(C)C)-c1ccccc1	7.73	

nazate-carbamate					
HC/DDC	---[b]	---[c]	COc1ccc(cc1N(NC(=O)OC(C)C)c1cc(ccc1OC)-c1ccccc1)-c1ccccc1.COc1ccc(c2c1nc1c(ccc(c21)-c1ccccc1)OC)-c1ccccc1	14.21	
72	7075	92-05-7	Oc1ccc(cc1O)-c1ccccc1	6.21	
I/IBMHC	---[b]	---[c]	COc1ccc(cc1NN(C(=O)OC(C)C)c1cc(ccc1OC)-c1ccccc1)-c1ccccc1	11.23	
IDD	---[b]	---[c]	Oc1cc(c(cc1-c1ccccc1)C1=CC(=C(O)C(=O)C1=O)c1ccccc1)O	8.54	
nazate-diazene oxide	74336768	---[c]	COc1ccc(cc1[N+]([O-])=NC(=O)OC(C)C)-c1ccccc1	7.14	
droxy-4'-methoxybiphenyl	11030839	16881-71-3	COc1ccc(cc1)-c1ccc(cc1)O	6.88	
nazate glucuronide	---[b]	---[c]	COc1ccc(cc1N(NC(=O)OC(C)C)C1OC(C(O)C(O)C1O)C(O)=O)-c1ccccc1	5.69	
droxybiphenyl glucuronide	3084305	19132-91-3	OC1C(O)C(OC(C1O)C(O)=O)Oc1ccc(cc1)-c1ccccc1	5.15	
dihydroxybiphenyl	7112	92-88-6	Oc1ccc(cc1)-c1ccc(cc1)O	6.33	
droxy bifenazate	---[b]	---[c]	COc1ccc(cc1NNC(=O)OC(C)C)-c1ccc(cc1)O	7.20	
droxy bifenazate-diazene	---[b]	---[c]	COc1ccc(cc1N=NC(=O)OC(C)C)-c1ccc(cc1)O	8.48	
orpropham	2728	101-21-3	CC(C)OC(=O)Nc1cccc(c1)Cl	6.52	7.8
droxychloropropham sulfate	125398281	28705-88-6	CC(C)OC(=O)Nc1ccc(c(c1)Cl)OS(O)(=O)=O	5.76	
heton-S-sulfone	17239	2496-91-5	CCOP(=O)(OCC)SCCS(=O)(=O)CC	5.77	
heton-O	9273	298-03-3	CCOP(=S)(OCC)OCCSCC	7.76	

neton-O-methyl	13346	867-27-6	CCSCCOP(=S)(OC)OC	6.64	
ifos	25146	10311-84-9	CCOP(=S)(OCC)SC(CCl)N1C(=O)c2ccccc2C1=O	9.47	8.8
otefuran metabolite DN sphate[a]	---[b]	---[c]	CNC(=N)NCC1CCOC1.OP(O)(O)=O	1.56	
ine	17110	2439-10-3	CCCCCCCCCCCCN=C(N)N.CC(O)=O	7.69	
in	91481	668-34-8	c1ccc(cc1)[Sn+](c1ccccc1)c1ccccc1	8.58	
in hydroxide	6327657	76-87-9	O[Sn](c1ccccc1)(c1ccccc1)c1ccccc1	7.78	
in acetate	16682804	900-95-8	CC(=O)O[Sn](c1ccccc1)(c1ccccc1)c1ccccc1	8.87	
in chloride	12540	639-58-7	Cl[Sn](c1ccccc1)(c1ccccc1)c1ccccc1	9.20	
in Flouride	9786	379-52-2	F[Sn](c1ccccc1)(c1ccccc1)c1ccccc1	8.52	
razamide	3081363	158237-07-1	CCN(C1CCCCC1)C(=O)N1N=NN(C1=O)c1ccccc1Cl	8.91	8.5 8.4
rocarb	17517	2631-40-5	CNC(=O)Oc1ccccc1C(C)C	6.56	7.1
hoxyfenozide	105010	161050-58-4	COc1cccc(c1C)C(=O)NN(C(=O)c1cc(cc(c1)C)C)C(C)(C)C	8.46	7.8 7.7
-Naphthylacetamide	68461	575-36-0	CC(=O)Nc1cccc2ccccc12	5.76	
aphthaleneacetamide	6861	86-86-2	NC(=O)Cc1cccc2ccccc12	5.18	
opropylphenol	6943	88-69-7	CC(C)c1ccccc1O	6.28	
oridazon Metabolite B1	594330	17254-80-7	CN1N=CC(=C(Cl)C1=O)N	2.68	
orothalonil Metabolite R611965	19028628	142733-37-7	NC(=O)c1c(c(cc(c1Cl)C(O)=O)Cl)Cl	4.22	
alaxyl Metabolite CGA 108906	117065479	104390-56-9	COCC(=O)N(C(C)C(O)=O)c1c(cccc1C(O)=O)C	5.05	
alaxyl Metabolite CGA 62826	13073467	87764-37-2	COCC(=O)N(C(C)C(O)=O)c1c(cccc1C)C	6.42	

azachlor Metabolite BH479-4	86290103	1231244-60-2	Cc1cccc(c1N(Cn1cccn1)C(=O)C(O)=O)C	5.90	
azachlor Metabolite BH479-9	139291839	---[c]	Cc1cccc(c1N(Cn1cccn1)C(=O)CS(=O)CC(O)=O)C	5.41	
azachlor Metabolite BH479-11	51071993	1242182-77-9	Cc1cccc(c1N(Cn1cccn1)C(=O)CS(C)=O)C	5.81	
azachlor Metabolite 479M12	139291822	---[c]	Cc1cccc(c1N(Cn1cccn1)C(=O)C(O)=O)C(O)=O	4.54	
outhylazine Metabolite A324007	---[b]	---[c]	CC(C)(C)Nc1nc(nc(n1)O)O	4.99	
outhylazine Metabolite N545666	---[b]	---[c]	CN1C(=NC(=NC1=O)NC(C)(C)C)O	4.96	
osulfuron Metabolite 635M03[a]	---[b]	---[c]	NC(=N)NC(=O)NS(=O)(=O)c1ccccc1C(F)(F)F	4.22	
osulfuron Metabolite BH635-4[a]	139597579	---[c]	NC(=O)NC(=N)NC(=O)NS(=O)(=O)c1ccccc1C(F)(F)F	4.10	

[a] Pesticides falling outside the applicability domain of the QSPR model (h_i > 0.033).

[b] PubChem CID not available.

[c] CAS number not available.

[d] Wang, J., Chow, W., Wong, J. W., Leung, D., Chang, J., & Li, M. (2019). *Analytical and Bioanalytical Chemistry*, 411, 1421-1431.

[e] Wang, J., Chow, W., Chang, J., & Wong, J. W. (2017). *Journal of Agricultural and Food Chemistry*, 65(2), 473-493.

[f] Wang, J., Chow, W., Chang, J., & Wong, J. W. (2014). *Journal of Agricultural and Food Chemistry*, 62(42), 10375-10391.

Table 2. The best foodinformatic models obtained by the replacement method supervised variable selection for predicting the t_R of pesticides in the Hypersil Gold column by means of the UHPLC/ESI Q-Orbitrap

Model	descriptors	R^2_{train}	s_{train}	R^2_{val}	s_{val}	R^2_{loo}	s_{loo}
IM1	*cLogP, DCW*	0.82	0.94	0.73	0.96	0.81	0.95
IM2	*Eta_D_epsiD, cLogP, DCW*	0.83	0.91	0.76	0.91	0.82	0.92
IM3	*Eta_D_epsiD, cLogP, SubFP26, DCW*	0.85	0.85	0.78	0.86	0.84	0.87
IM4	*Eta_D_epsiD, cLogP, Alkyl-Amines, MDEN.22, DCW*	0.87	0.81	0.79	0.82	0.86	0.83

Table 3. Summary of the test set statistical quality for both the individual and consensus models for predicting the retention time of pesticides in the Hypersil Gold column in the UHPLC/ESI Q-Orbitrap. The best model is highlighted in bold.

Model	Q^2_{F1}	$Q^2_{F1\_95\%}$	Q^2_{F2}	$Q^2_{F2\_95\%}$	Q^2_{F3}	$\overline{R^2_m}$	ΔR^2_m	*MAE*	$MAE_{95\%}$	Prediction quality
IM1	0.71	0.78	0.69	0.77	0.83	0.61	0.01	0.70	0.62	good
IM2	0.72	0.78	0.70	0.77	0.83	0.61	0.04	0.68	0.59	good
IM3	0.73	0.79	0.71	0.78	0.84	0.62	0.06	0.66	0.57	good
IM4	**0.75**	**0.81**	**0.74**	**0.80**	**0.85**	**0.63**	**0.15**	**0.62**	**0.53**	**good**
CM0	0.73	0.80	0.72	0.79	0.84	0.63	0.07	0.65	0.57	good
CM1	0.74	0.80	0.72	0.79	0.84	0.63	0.07	0.65	0.57	good
CM2	0.74	0.80	0.73	0.79	0.85	0.63	0.08	0.65	0.56	good
CM3	0.75	0.81	0.74	0.80	0.85	0.63	0.13	0.63	0.54	good

Figure 1. a) Experimental versus predicted retention times for pesticide residues detected in fruits and vegetables using UHPLC/ESI Q-Orbitrap in the Hypersil Gold selectivity column. b) Scatter plot of the standardized residuals versus the predicted retention times for pesticide residues detected in fruits and vegetables using UHPLC/ESI Q-Orbitrap in the Hypersil Gold selectivity column.

Figure 2. Pesticides falling outside the AD of the foodinformatic model (leverage value above > 0.033) for the test set (a-c) and the external set (d-f).

Credit authorship contribution statement

Cristian Rojas: Conceptualization, Methodology, Data curation, Software, Investigation, Validation, Writing - review & editing. José F. Aranda: Conceptualization, Methodology, Software, Writing - review & editing. Elisa Pacheco Jaramillo: Data curation, Methodology. Irene Losilla: Data curation, Methodology. Piercosimo Tripaldi: Data curation, Methodology, Investigation. Pablo R. Duchowicz: Conceptualization, Methodology, Software, Validation, Writing - review & editing. Eduardo A. Castro: Methodology, Investigation, Writing - review & editing.

Research highlights

1. Retention times of a large set of pesticide residues detected in fruits and vegetables using UHPLC/ESI Q-Orbitrap in the Hypersil Gold selectivity column.
2. Filtering and curation of the Compound DataBase (CDB) of pesticides.
3. Establishment of a foodinformatic model for the prediction of retention times by means of unsupervised and supervised machine learning approaches, as well as consensus analysis.
4. Implementation of the *in silico* model as a real task for the prediction of the retention times of an external set of 57 pesticides and their metabolites or fragments.