**Predicting reversed-phase liquid chromatographic retention times of pesticides by deep neural networks**

Julien Parinet\*1

\*corresponding author: julien.parinet@anses.fr

1 Université de Paris-Est, ANSES, Laboratory for Food Safety, 94700, Maisons-Alfort, France

**Abstract**

Diverse deep neural networks (DNN) with various set of inputs were developed and compared in order to predict reversed-phase liquid chromatographic (RPLC) retention times of 792 pesticides. The development of quantitative structure–retention relationship models (QSRR) requires selection of the best molecular descriptors and machine-learning algorithms. Here, we used two main approaches, one based on an extensive literature review to select the best set of molecular descriptors (16), and a second with diverse strategies in order to select among 1545 molecular descriptors (MD), 16 MD. In both cases, DNN were optimized through a gridsearch. The DNN with 16 molecular descriptors inherited from the literature review showed better accuracy than the different DNN with 16 molecular descriptors selected by diverse strategies.

**Keywords:** pesticides, QSRR, molecular descriptors, deep neural network, reversed-phase liquid chromatography, selection of inputs

1. **Introduction**

Pesticides in food are of growing interest as the general public is increasingly concerned about their health effects (Dashtbozorgi et al., 2013). Depending on their concentrations, toxicity, and frequence of detection in food and in the environment, pesticides may lead to health impairment, disease and even death (Colosio et al., 2017). Detecting and quantifying these compounds helps to guarantee compliance of imported goods with the laws and regulations of the importing country (Chiesa et al., 2016).

The high accuracy and mass sensitivity of high-resolution mass spectrometry (HRMS) instruments hyphenated to liquid (LC) or gas (GC) chromatography make it possible to observe thousands of chemical features in food and environment samples. These features include monoisotopic exact mass, chromatographic retention time (RT), abundance, isotope profiles and MS² fragmentations. However, data processing and chemical characterization remain difficult despite recent developments. Chemical reference standards and spectral data enable us to confirm the structure of observed characteristics, but reference standards are rarely available for thousands of characteristics in non-target analysis (NTA) and suspect screening analysis (SSA) (McEachran et al., 2018), and having these thousands of standards can also represent a considerable cost.

Since the appearance of HRMS, the interest in improving confidence in the identification of small molecules increase, such as pesticides, from putative positive samples based on detection to confirmation (Bade et al., 2015a; Schymanski et al., 2014). SSA studies are those in which observed but unknown features are compared against a database of chemical suspects to identify plausible hits. NTA studies are those in which chemical structures of unknown compounds are postulated without the aid of suspect lists (Sobus et al., 2018). In both cases, confirming the identification of a contaminant requires its standard, which may be unavailable, expensive, or time-consuming to obtain in the case of food poisoning. This is especially true for pesticides where there are a few thousand analytes and metabolites. In order to increase confidence in the tentative identification of compounds, especially in SSA, it is conceivable to predict their chromatographic retention time (RT) (Bade et al., 2015b; Barron and McEneff, 2016; Parinet, 2021; Randazzo et al., 2016).

To predict RT, various strategies using diverse molecular descriptor (MD) sets and diverse machine-learning algorithms have been tested and published (Aalizadeh et al., 2019; Bade et al., 2015a; Barron and McEneff, 2016; Goryński et al., 2013; McEachran et al., 2018; Munro et al., 2015; Noreldeen et al., 2018; Parinet, 2021; Randazzo et al., 2016). These strategies run the gamut from the use of logKow models (Bade et al., 2015b) to more complex *in silico* approaches based on quantitative structure–retention relationship (QSRR) modelling, including artificial neural networks (ANNs), support-vector machines (SVMs), random forest (RF), partial least squares regression (PLS-R), and multi-linear regression (MLR) (Ghasemi and Saaidpour, 2009; Munro et al., 2015; Parinet, 2021).

In the first part of this study, two different approaches were tested and compared in order to build an effective QSRR model dedicated specifically to predicting pesticide RTs analyzed by reversed-phase liquid chromatography (RPLC) (C18) in SSA or NTA. The first approach was based on an exhaustive literature review in order to find the best MD set to predict pesticide RTs. The second approach had no preconceived ideas as to which MDs that should be selected among 1545 MDs to feed the QSRR. Indeed, in this second approach, various strategies using the Lasso regression, a Pearson correlation feature selection (Pearson), a recursive feature elimination (RFE) and the use of principal components analysis (PCA) have been used in order to select among the entire MD available, sixteen MD. In both cases, a deep learning algorithm was retained and optimized (a multilayer perceptron (MLP)) in order to predict RTs of pesticides, and a comparison was done between the two approaches in order to select the best one.

1. **Materials and Methods**
   1. Dataset

Initially, the dataset included 843 RTs of pesticides collected from the article of Wang et al. (2019). Ultra-high-performance liquid chromatography (UHPLC) gradient conditions, column temperatures, mobile phases, columns, and instruments used to generate the data are listed in **Table 1** and presented in detail in Wang et al. (2019).

Three free software applications were used to calculate the molecular descriptors of pesticide compounds. These applications are free of charge, can calculate a large number of descriptors, and are widely available. Firstly, the free version of ACD software (Advanced Chemistry Development, Toronto, ON, Canada) was used to calculate *LogP* and *LogD*. Secondly, the Toxicity Estimation Software Tool (TEST, Cincinnati, OH, USA) was used to calculate *Hy, Ui, IB, BEHp1, BEHp2, GATS1m, and GATS2m*.The rest of the molecular descriptors (1834 MD) were calculated using the ChemDes online platform (<http://scbdd.com/chemdes/>).

Once the MDs were calculated, the dataset was cleaned in order to remove constant and missing values (**Figure 1**). At the end of this curation process, 792 pesticides, their RTs, and 1545 MDs remained in the final dataset. The dataset containing the MDs for each pesticide was then ready to build QSRR models (Supplementary Data).

* 1. QSRR model development

The dataset constituted previously and containing the pesticides (792), their MDs (1545), and RTs was used in order to select among them the best MDs inherited from the literature review (*Model 1).* Importantly, in order to find the best set of MDs, a literature review was done by selecting the most recent and pertinent papers with the following criteria: the prediction of retention times measured by RPLC and for pesticides or similar compounds (pharmaceuticals, veterinary drugs). At the end of this literature review, seven articles, their MDs, and models were selected (shown in **Table 2** with their performances) and compared in term of performance measured principally through the *percentage of error*, which is the ratio between the root mean square error (RMSE) divided by the maximum retention time measured on the last eluted compound. In order to pursue the *no a priori approach* on which MD to select (*Model 2 to Model 8*), diverse strategies were used and compared in order to select among the 1545 MD, the best sixteen MD. Sixteen MD were retained in order to be able to compare the performances of the models (*Model 2 to 8*) to the model inherited from the literature review (*Model 1*). Hence, the Lasso regression, a regularized linear regression that aims to constrain the coefficients to be close to 0 or equal to zero, thus allowing an automatic selection of the characteristics/MD, here 16 MD (*ATS8m, ATS5i, iedm, SRW10, ATS5v, VR2\_Dt, VR1\_D, VR1\_Dt, VR2\_D, ATS8i, ATS7i, ATS3i, ATSC3m, ATS0m, ATS0v, ATS4v*). The second strategy was based on the Pearson correlation between the 1545 MD and the output (pesticides RTs), and the larger the relationship and more likely the feature/MD should be selected for modeling, then sixteen MD were selected based on this strategy (*LogP, BEHm4, CrippenLogP, ALOGP2, ALOGP, XLOGP2, XLOGP, ATS6p, ATS5p, ATS4p, ATS3p, ATS1p, ATS6v, BEHm8, BEHm5, BEHm7)*. The third strategy, a recursive feature elimination (RFE), was based on an iterative selection of features/MD made by initially selecting all the MD, then a model is built (here a multi-linear regression), then the least important characteristic is rejected and this process is done until a model with 16 MD is obtained (*maxtsC, MWC2, MWC03, MWC4, MWC5, nN, k2, MDEN-23, MDEN-33, MDEO-11, MDEO-12, MDEC-34, MDEC-44, MAXDP2, MDEN-22, ieadjmm*). Finally, the fourth strategy was based on principal component analysis (PCA) and declined under four sub strategies (*PCA1* to *PCA4*). For the four sub strategies, the same PCA was used. Hence, a PCA was done on the 1545 MD and measured on the 792 pesticides. The MD were normalized (reduced and centered) before doing the PCA and 16 principal components (PC) were retained; *PCA1* strategy was based on the selection of the MD most correlated to each PC, thus 16 MD were selected (*TWC, CIC1, ETA\_Epsilon\_2, AATS1p, icyce, MLFER\_E, MATS2v, nCl, AATSC3p, R, JGI3, StsC, nHCHnX, ATSC6e, MATS6i, MATS6m).* The *PCA2* strategy was based on the selection of the 16 MD most correlated to PC1, as PC1 was the PC the most correlated to RT (*TWC, Zagreb, nBonds, nBO, MWC01, SRW02, MPC01, ZM1, WTPT-1, SRW04, CID, nHeavyAtom, MPC2, nSK, SRW01, BID*). The *PCA3* strategy was based on the selection of the 16 MD most correlated to PC1 (8 MD) and PC4 (8 MD) as PC1 and PC4 were the most correlated to RT (*TWC, Zagreb, nBonds, nBO, MWC01, SRW02, MPC01, ZM1, AATS1p, AATS0p, AATS4p, Mp, ETA\_AlphaP, AATS3p, AATS5p, AATS2p*). Finally, the *PCA4* strategy was based on the selection of the 16 PC and their corresponding scores used as input (PC1 to PC16).

Regardless of the MD dataset used, the following procedure was used. The MD datasets, and the corresponding values of pesticide RTs, were divided into two subsets: an internal dataset and an external dataset (**Figure 1**). The internal dataset was composed of 594 pesticides chosen randomly, their corresponding MD (input) and experimentally measured pesticide RTs (output). The external dataset was composed of 198 randomly chosen pesticides, their corresponding MDs, and experimentally measured pesticide RTs. The internal dataset itself was divided randomly into a training and a test set, with a size ratio of three to one, respectively.

Initially, the internal training dataset was used to train the DNN, here an MLP, by tuning the hyper-parameters through a gridsearch and a cross-validation process, where the internal training dataset was divided in five equal size sub-datasets (cv = 5). The hyper-parameters tuned were:

* Number of hidden layers constituted each by a number of neurons equal to the number of MD used as inputs Geron (2017)
* The activation function among: ReLu, tanh and logistic
* The alpha value: 10 or 1
* The solver function among: Adam, SGD and Lbfgs.

The data were standardized (mean-centered) in order to accelerate and enhance the training and the predictions, and also to simplify interpretation of the importance of the features/MDs.

* 1. Model validation

The validation of QSRR models is probably the most important and crucial part of model evaluation. For this reason, we carried out the validation step using internal and external validation (Noreldeen et al., 2018) (**Figure 1**).

The coefficient of determination (R²) and the RMSE were used to evaluate and compare the models extracted from the literature review and were measured on the test set (**Table 2**). These parameters were also used for the models developed in this study in order to determine the error between the experimental and predicted RTs in the QSRR models, especially in terms of their ability to be generalized to new pesticide substances with unknown RTs. The lower the RMSE and the higher the R² value, the better the model. The R² and RMSE were measured, in the case of the models developed in this present study, on the training set (n = 445 pesticides) coming from the internal set, on the test set of the internal set (n = 148 pesticides), and on the external set (n = 198 pesticides) (**Table 3**).

The percentage of error was used to compare the models. Of note, the gradient durations are not the same between the different studies mentioned in the literature review (**Table 2**), and an RMSE of 1 minute does not have the same meaning for a gradient of 10 minutes or for a gradient of 40 minutes. For this reason, the maximum chromatographic retention time (RT max) was systematic recorded (**Tables 2, 3**). The RT max, displayed in **Table 3**, corresponds to the elution time of the last compound analyzed.

The following statistics were calculated using Python Software (Version 3.8) for model validation and comparison (McEachran et al., 2018):

* The coefficient of determination (R²) between predicted and experimental RTs was calculated as follows (Eq.1):

(1)

where and are the predicted and experimental RTs, respectively, and is the mean experimental RT.

* The root mean square error (RMSE) between predicted and experimental RTs was calculated as follows (Eq.2):

RMSE = (2)

where and are the predicted and experimental responses, respectively.

* The percentage of error (% error) was calculated as follows (Eq.3):

Percentage of error = (3)

* 1. Structure of the DNN

DNN is a computer program inspired by the biological neural network and designed in order to modelize complex, non-linear problems (classification or regression). A typical DNN is composed of a number of neurons from a few to millions, which are arranged in a series of layers (Zhong et al., 2020). A neuron is a computational unit that has one or more weighted input connections, a transfer function that combines the inputs in some way, and an output connection. There are three kind of neurons: input and output neurons, and the neurons of the hidden layers between them. The input neurons in the input layer are designed to receive the external data, such as the MDs used here, and the output neurons in the last layer are the final predictions made by the DNN, which will be used to compare with the true target data, such as RTs of pesticides. Between the input layer and the output layer are hidden layers, often more than one layer (Zhong et al., 2020). The input data go into the DNN through the input layer, are then transformed in the hidden layers, and finally become the predictions in the output layer. The values in all neurons in the hidden and output layers are calculated by the application of an activation function on the sum of the values in the previous neurons×weight+bias calculation, in which weights and biases can be updated based on the errors between the predictions and the target until the errors reach a minimum value. Update of the weights and biases is done through back-propagation of the errors between the target (RT experimental) and the prediction (RT predicted). This process is the “learning” process of DNN. DNNs have two main hyperparameters: the number of neurons per layer, and the number of layers. The number of layers and neurons is also called the “depth” and “width” of DNN, respectively. Larger numbers of layers and neurons mean deeper and wider DNNs, which often have more powerful fitting ability and can achieve better accuracy on the prediction. However, too many layers and neurons can lead to an overfitting problem, which is an accurate prediction on the training set but poorer prediction on the test set. It is crucial for the DNN to be able to generalize on a dataset never seen before. For this last reason, we split the dataset into an internal and an external dataset, in order to evaluate the capacity of the DNN to generalize. The model development process is hence to develop an optimum architecture of the DNN with an appropriate fitting ability. In this study, our DNN was composed of an input layer, several hidden layers, and an output layer. In each layer, there are numerous neurons accepting values from the neurons of the neighboring layer. In the input and hidden layers, the number of neurons was equal to the number of MDs selected. For instance, if the number was 16 MDs, then there were 16 neurons in the input and in each hidden layer, as suggested by Geron (2017). The number of neurons in the output layer was 1 because there was only one RT for each pesticide. The number of neurons in the hidden layers was set manually before the learning process began. Here, we focused on the following hyperparameters: the number of hidden layers, the activation function, the alpha value, and the solver used. We investigated their effects on the performance of the DNN through a gridsearch and a cross-validation (cv=5) process done on the internal training set. The R² and RMSE values were calculated to evaluate the effects of the hyperparameters on the performances of the models developed and on overfitting. A detailed description of the theory behind DNNs has been adequately provided elsewhere (Zhong et al., 2020). Model training was stopped after 1000 epochs (iterations).

1. **Results and discussion**

For a DNN, prediction accuracy is highly related to its structure, the number of layers, neurons, other hyperparameters (activation function, solver for weight optimization, etc.), and even more to the inputs retained, in our case the MDs.

* 1. Comparison of published QSRR models

One of the main bottlenecks in designing QSRR models is selecting the MDs (May et al., 2011; Parinet, 2021; Scotti et al., 2016). The selection of the most suitable MDs, among several thousand, can follow various strategies (May et al., 2011); this step is particularly complicated because there are many molecular descriptors that can be calculated and used (Aalizadeh et al., 2019; Bade et al., 2015b, 2015a; McEachran et al., 2018; Munro et al., 2015; Noreldeen et al., 2018) and many strategies to select the MDs.

Here, to develop the most accurate QSRR dedicated to pesticides, we used two different approaches. The first approach was based on an extensive literature review on the prediction of RPLC retention times of compounds similar in their structures and properties to pesticides, such as pharmaceuticals and veterinary drugs. Based on this literature review, seven articles emerged (**Table 2**). In order to select the best set of MDs among the seven research papers, a study of the QSRR models developed was carried out. In order to do this, the performances of the QSRR models were documented and compared (**Table 2**). The number of contaminants used to build and optimize the QSRR models was found to be between 95 and 1830 compounds, the number of MDs selected was between 5 and 16, and the RT max values measured were between 9.3 and 40.8 min. The machine learning algorithms used were SVM, DNN (MLP and general regression neural networks (GRNN)), and MLR. The performances measured on the test set are for the R² between 0.63 and 0.95, and for the RMSE between 0.62 and 1.42 min. Nevertheless, the gradients are not similar, reflected by the different RT max measurements. The RMSE and the R² alone are not sufficient to determine which MD set and QSRR model is the most efficient. For this reason, we calculated the percentage of error (Eq. 3), which was not done in the recent article of Parinet (2021) where all the references selected, and their corresponding MD datasets were applied directly on the pesticides dataset in order to make the prediction of RT. The percentage of error was between 5.4% and 9.4%. The lowest value for the percentage of error was obtained for the QSRR developed by Bade and colleagues (2015) on 544 emerging contaminants and by the use of 16 MDs (*nDB, nTB, nC, nO, nR04-nR09, UI, Hy, MLogP, ALogP, LogP, LogD*) and a DNN (MLP). Based on these results, we retained for our QSRR development, the Bade and colleagues (2015) MD set and the MLP as the best ML algorithm to use (*model 1*) with a percentage of error equal to 5.4%. Then, we used the MD listed by Bade and colleagues (2015) on our dataset and through a MLP (*Bade-MLP – Model 1*) as described before in the text. By this approach we got a R² on the internal training and test set equal to 0.95 and 0.90, respectively (**Table 3**). The RMSE obtained on the internal training and test set were equal to 0.43 and 0.63. On the external set, never used for the learning and optimizing process, the R² was equal to 0.82 and the RMSE equal to 0.67 (**Table 3**). These past results are similar to those obtained by Parinet (2021) with the McEachran 3 MDs, on the external dataset, and by the use of SVM and MLP as machine learning algorithms where the R² were between 0.85-0.89 and the RMSE between 0.64-0.69, respectively. The percentage of error obtained thanks to these molecular descriptors and with a MLP was around 6%, which is close to the 5.4% got by Bade and colleagues (2015) on their compounds.

* 1. Comparison between QSRR models developed thanks to the literature review and to the no *a priori* approaches

To develop the most efficient QSRR model specifically for pesticides, we compared the performances obtained for *Model 1* (*Bade-MLP*) with those of *Model 2 to 8* (no *a priori* approach).

The performances of *Model 2* (*Lasso-MLP*) applied on our pesticide dataset gave R² on the internal training and test set equal to 0.60 and 0.50, respectively (**Table 3**). The RMSE obtained on the internal training and test set were equal to 1.19 and 1.27. On the external set, the R² was equal to 0.49 and the RMSE equal to 1.36 (**Table 3**). The percentage of error obtained thanks to these molecular descriptors and with a MLP was around 12%, which is twice as much as *Model 1* (*Bade-MLP*) with 6% on the same compounds.

The performances of *Model 3* (*Pearson-MLP*) applied on our pesticide dataset gave R² on the internal training and test set equal to 0.79 and 0.79, respectively (**Table 3**). The RMSE obtained on the internal training and test set were equal to 0.86 and 0.83. On the external set, the R² was equal to 0.78 and the RMSE equal to 0.88 (**Table 3**). The percentage of error obtained thanks to these molecular descriptors and with a MLP was around 8%, which is less good as *Model 1* (*Bade-MLP*) with 6% on the same compounds but much better than *Model 2*.

The performances of *Model 4* (*RFE-MLP*) applied on our pesticide dataset gave R² on the internal training and test set equal to 0.69 and 0.60, respectively (**Table 3**). The RMSE obtained on the internal training and test set were equal to 1.04 and 1.15. On the external set, the R² was equal to 0.63 and the RMSE equal to 1.16 (**Table 3**). The percentage of error obtained thanks to these molecular descriptors and with a MLP was around 10%, which is less good as *Model 1* (*Bade-MLP*) with 6% on the same compounds, and less good as *Model 3*.

The performances of *Model 5* (*PCA1-MLP*) applied on our pesticide dataset gave R² on the internal training and test set equal to 0.75 and 0.61, respectively (**Table 3**). The RMSE obtained on the internal training and test set were equal to 0.94 and 1.12. On the external set, the R² was equal to 0.64 and the RMSE equal to 1.14 (**Table 3**). The percentage of error obtained thanks to these molecular descriptors and with a MLP was around 10%, which is less good as *Model 1* (*Bade-MLP*) with 6% on the same compounds, and quite similar to *Model 4*.

The performances of *Model 6* (*PCA2-MLP*) applied on our pesticide dataset gave R² on the internal training and test set equal to 0.42 and 0.34, respectively (**Table 3**). The RMSE obtained on the internal training and test set were equal to 1.44 and 1.47. On the external set, the R² was equal to 0.38 and the RMSE equal to 1.50 (**Table 3**). The percentage of error obtained thanks to these molecular descriptors and with a MLP was around 13%, which is less good as *Model 1* (*Bade-MLP*) with 6% on the same compounds, and the worst model developed with performances quite similar to *Model 2*.

The performances of *Model 7* (*PCA3-MLP*) applied on our pesticide dataset gave R² on the internal training and test set equal to 0.61 and 0.53, respectively (**Table 3**). The RMSE obtained on the internal training and test set were equal to 1.18 and 1.24. On the external set, the R² was equal to 0.56 and the RMSE equal to 1.26 (**Table 3**). The percentage of error obtained thanks to these molecular descriptors and with a MLP was around 11%, a little better than *Model 5* but which is less good as *Model 1* (*Bade-MLP*) with 6% on the same compounds.

The performances of *Model 8* (*PCA4-MLP*) applied on our pesticide dataset gave R² on the internal training and test set equal to 0.82 and 0.75, respectively (**Table 3**). The RMSE obtained on the internal training and test set were equal to 0.79 and 0.91. On the external set, the R² was equal to 0.76 and the RMSE equal to 0.93 (**Table 3**). The percentage of error obtained thanks to these molecular descriptors and with a MLP was around 8%, better than all the models developed thanks to the PCA approach and similar in term of performances to *Model 3,* but still less good as *Model 1* (*Bade-MLP*).

Whatever the strategy used, the model which offers the best performances, is the *Model 1* (*Bade-MLP*) inherited from the literature review. Nevertheless, the *no a priori* approach offers two models (*Model 3 and Model 8*) with effective performances. Among all the models developed thanks to the PCA approach, the Model 8 offers the best performances, and then comes next the *Model 5* and *7* and finally the *Model 6* that is the worst one.

* 1. Optimization of the hyperparameters

The QSRR models were optimized using an MLP through a gridsearch process. Nevertheless, the number of neurons per hidden layers was set manually and was determined by applying the recommendations of Geron (2017). Importantly, Geron mentions that the common practice of sizing the hidden layers to form a funnel, with an ever-decreasing number of neurons at each layer is no longer as common, and instead we can simply give the same size to all the hidden layers, resulting in only one hyperparameter to adjust instead of one per layer. Nonetheless, it is more useful, still according to Geron (2017), to increase the number of layers rather than the number of neurons per layer. For this reason, the number of hidden layers used by the gridsearch was between 1 to 5 layers, irrespective of the QSRR.

Once the number of neurons per hidden layer and the number of hidden layers are set, there remains a large number of hyperparameters to optimize. Nevertheless, some of them are more important than others, such as the activation function and the solver used. For this reason, the gridsearch for the activation function was done among the following functions: ReLu, tanh, and logistic. A gridsearch was also carried out to select the best solver among three possible choices (Adam, SGD and Lbfgs). The last hyperparameter to optimize through the gridsearch was the alpha value, which is a regularization parameter (L2 regularization); alpha value was comprised between 0.01 and 100 (**Table 3**). All the architecture of DNN and theire hyperparameters retained through the girdsearch for the *models 1 to 8* are listed in **Table 3**. Hence, the number of layers are comprised between 1 to 5, two activation functions among three were used (ReLu and tanh) and the logisitic function was never retained by the gridsearch, two solver (Adam and SGD) among three were used. Finally, despite the amplitude values of alpha, two alpha values were retained: 1 and 10.

1. **Conclusions**

We compared a literature review approach to a no *a priori* approach in order to select, by diverse strategies, the best set of molecular descriptors among 1545 MD in order to predict, through a QSRR model, the RPLC retention times of 792 pesticides. The literature review approach yielded the best results when DNN was used as the ML algorithm, with an R² of 0.82 and an RMSE of 0.67 min (*Model 1*) on the external set. Nevertheless, the *Model 3* (a Pearson selection) and *8* (a PCA approach) gave reasonable results with a percentage of error equal to 8% in both cases. However, it could be useful in future resaerch to test some other *no a priori* selection strategies in order to determine new MD datasets and also to consider reducing the number of MD with the goal to simplify the models while obtaining good predictions.

**Tables**

**Table 1** Ultra-high-performance liquid chromatography conditions used to generate the retention time data

**Table 2** QSRR models selected from the literature review

**Table 3** Performances of QSRR models applied to the pesticide dataset

**Figures**

**Figure 1** QSRR model development and evaluation of performances

**Funding sources**

This work was funded by the French National Research Agency (ANR), AlimOmic project: grant ANR-19-CE21-0002.

**Conflicts of interest**

The authors declare that they have no conflicts of interest.

**Acknowledgements**

The author thanks warmly Jian Wang, Willis Chow, Jon W. Wong, Daniel Leung, James Chang, and Mengmeng Li for agreeing to use of their published results.

**References**

Aalizadeh, R., Nika, M.C., Thomaidis, N.S., 2019. Development and application of retention time prediction models in the suspect and non-target screening of emerging contaminants. Journal of Hazardous Materials 363, 277–285. https://doi.org/10.1016/j.jhazmat.2018.09.047

Bade, R., Bijlsma, L., Miller, T.H., Barron, L.P., Sancho, J.V., Hernández, F., 2015a. 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, 934–941. https://doi.org/10.1016/j.scitotenv.2015.08.078

Bade, R., Bijlsma, L., Sancho, J. V., Hernández, F., 2015b. 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, 143–149. https://doi.org/10.1016/j.talanta.2015.02.055

Barron, L.P., McEneff, G.L., 2016. Gradient liquid chromatographic retention time prediction for suspect screening applications: A critical assessment of a generalised artificial neural network-based approach across 10 multi-residue reversed-phase analytical methods. Talanta 147, 261–270. https://doi.org/10.1016/j.talanta.2015.09.065

Chiesa, L.M., Labella, G.F., Giorgi, A., Panseri, S., Pavlovic, R., Bonacci, S., Arioli, F., 2016. The occurrence of pesticides and persistent organic pollutants in Italian organic honeys from different productive areas in relation to potential environmental pollution. Chemosphere 154, 482–490. https://doi.org/10.1016/j.chemosphere.2016.04.004

Colosio, C., Rubino, F.M., Moretto, A., 2017. Pesticides, in: International Encyclopedia of Public Health. pp. 454–462. https://doi.org/10.1016/B978-0-12-803678-5.00329-5

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. https://doi.org/10.1016/j.microc.2012.05.003

Ghasemi, J., Saaidpour, S., 2009. QSRR prediction of the chromatographic retention behavior of painkiller drugs. Journal of Chromatographic Science 47, 156–163. https://doi.org/10.1093/chromsci/47.2.156

Goryński, K., Bojko, B., Nowaczyk, A., Buciński, A., Pawliszyn, J., Kaliszan, R., 2013. Quantitative structure-retention relationships models for prediction of high performance liquid chromatography retention time of small molecules: Endogenous metabolites and banned compounds. Analytica Chimica Acta 797, 13–19. https://doi.org/10.1016/j.aca.2013.08.025

May, R., Dandy, G., Maier, H., 2011. Review of Input Variable Selection Methods for Artificial Neural Networks. Artificial Neural Networks - Methodological Advances and Biomedical Applications. https://doi.org/10.5772/16004

McEachran, A.D., Mansouri, K., Newton, S.R., Beverly, B.E.J., Sobus, J.R., Williams, A.J., 2018. A comparison of three liquid chromatography (LC) retention time prediction models. Talanta 182, 371–379. https://doi.org/10.1016/j.talanta.2018.01.022

Munro, K., Miller, T.H., Martins, C.P.B., Edge, A.M., Cowan, D.A., Barron, L.P., 2015. Artificial neural network modelling of pharmaceutical residue retention times in wastewater extracts using gradient liquid chromatography-high resolution mass spectrometry data. Journal of Chromatography A 1396, 34–44. https://doi.org/10.1016/j.chroma.2015.03.063

Noreldeen, H.A.A., Liu, X., Wang, X., Fu, Y., Li, Z., Lu, X., Zhao, C., Xu, G., 2018. Quantitative structure-retention relationships model for retention time prediction of veterinary drugs in food matrixes. International Journal of Mass Spectrometry 434, 172–178. https://doi.org/10.1016/j.ijms.2018.09.022

Parinet, J., 2021. Chemosphere Prediction of pesticide retention time in reversed-phase liquid chromatography using quantitative-structure retention relationship models : A comparative study of seven molecular descriptors datasets. Chemosphere 275, 130036. https://doi.org/10.1016/j.chemosphere.2021.130036

Randazzo, G.M., Tonoli, D., Hambye, S., Guillarme, D., Jeanneret, F., Nurisso, A., Goracci, L., Boccard, J., Rudaz, S., 2016. Prediction of retention time in reversed-phase liquid chromatography as a tool for steroid identification. Analytica Chimica Acta 916, 8–16. https://doi.org/10.1016/j.aca.2016.02.014

Schymanski, E.L., Jeon, J., Gulde, R., Fenner, K., Ruff, M., Singer, H.P., Hollender, J., 2014. Identifying small molecules via high resolution mass spectrometry: Communicating confidence. Environmental Science and Technology 48, 2097–2098. https://doi.org/10.1021/es5002105

Scotti, M.T., Scotti, L., Ishiki, H.M., Peron, L.M., de Rezende, L., do Amaral, A.T., 2016. Variable-selection approaches to generate QSAR models for a set of antichagasic semicarbazones and analogues. Chemometrics and Intelligent Laboratory Systems 154, 137–149. https://doi.org/10.1016/j.chemolab.2016.03.023

Sobus, J.R., Wambaugh, J.F., Isaacs, K.K., Williams, A.J., Mceachran, A.D., Richard, A.M., Grulke, C.M., Ulrich, E.M., Rager, J.E., Strynar, M.J., Newton, S.R., 2018. Integrating tools for non-targeted analysis research and chemical safety evaluations at the US EPA. Journal of Exposure Science & Environmental Epidemiology 411–426. https://doi.org/10.1038/s41370-017-0012-y

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. https://doi.org/10.1007/s00216-019-01581-z

Zhong, S., Hu, J., Fan, X., Yu, X., Zhang, H., 2020. A deep neural network combined with molecular fingerprints (DNN-MF) to develop predictive models for hydroxyl radical rate constants of water contaminants. Journal of Hazardous Materials 383, 121141. https://doi.org/10.1016/j.jhazmat.2019.121141