Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

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

Julien Parinet [*]

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

ARTICLE INFO

Keywords:
Pesticides
QSRR
Molecular descriptors
Deep neural network
Reversed-phase liquid chromatography
Selection of inputs

ABSTRACT

To be able to predict reversed phase liquid chromatographic (RPLC) retention times of contaminants is an asset in order to solve food contamination issues. The development of quantitative structure–retention relationship models (QSRR) requires selection of the best molecular descriptors and machine-learning algorithms. In the present work, two main approaches have been tested and compared, 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, a deep neural network (DNN) were optimized through a gridsearch.

1. Introduction

Contaminants and especially pesticides in food are of growing concern as the general public is increasingly aware 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 MS^2 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, especially metabolites and by-products, 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, metabolites and by-products. 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, different strategies using various molecular descriptor (MD) sets and multiple 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 range from the use of logKow models (Bade et al., 2015b) to more complex in silico approaches based on quantitative structure-retention relationship (QSRR) modeling, including artificial neural networks (ANNs), support vector machines (SVMs), random forest (RF), partial least squares regression (PLS-R), and multilinear regression (MLR) (Ghasemi and Saaidpour, 2009; Munro et al., 2015; Parinet, 2021).

[*] Corresponding author.
E-mail address: julien.parinet@anses.fr.

https://doi.org/10.1016/j.heliyon.2021.e08563
Received 15 April 2021; Received in revised form 26 August 2021; Accepted 3 December 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.

2. Materials and methods

2.1. 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 presented in detail in Wang et al. (2019).

Three free software applications have been used in order to compute the pesticide's MD. These applications are free, can calculate a large number of descriptors and are widely available. The ACD software (Advanced Chemistry Development, Toronto, ON, Canada) was used to calculate *LogP* and *LogD*. The Toxicity Estimation Software Tool (TEST, Cincinnati, OH, USA) was used to compute *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 computed, the dataset was cleaned in order to remove constant and missing values (Figure 1). Indeed, constant values are useless in order to develop QSRR models and missing values make learning and prediction impossible. The missing values are due to the softwares and their inability to generate, depending on the molecules, the MD. 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 (Table S1).

2.2. 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 1 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

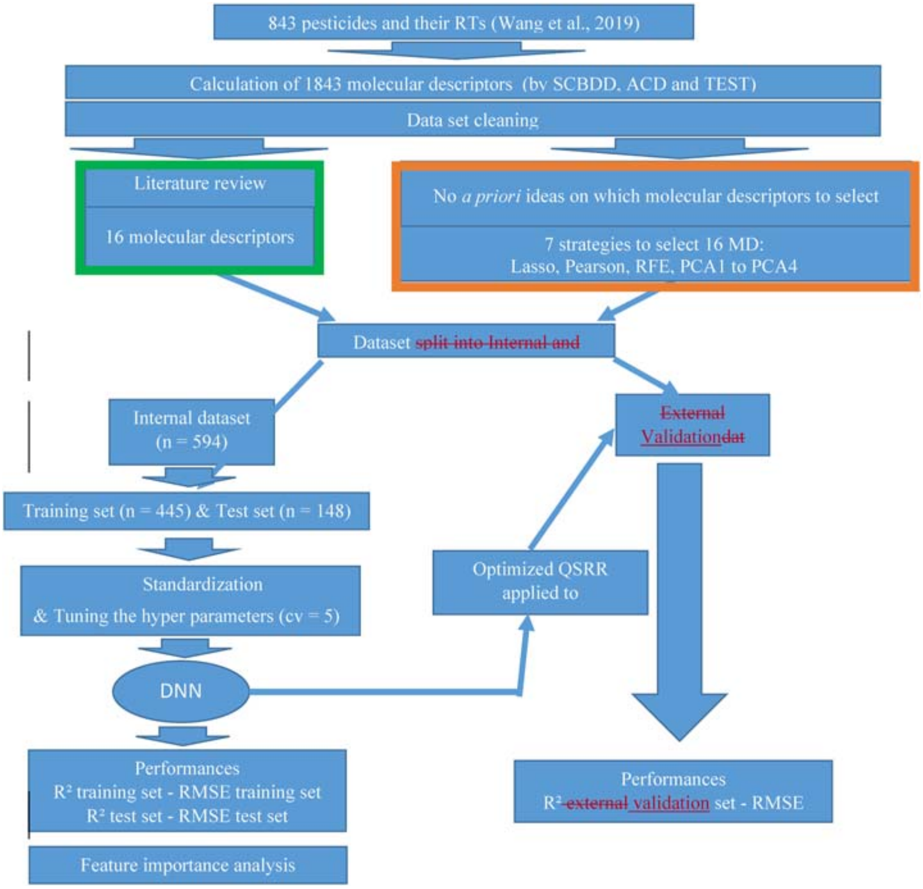
843 pesticides and their RTs (Wang et al., 2019)
Calculation of 1843 molecular descriptors (by SCBDD, ACD and TEST)
Data set cleaning
Literature review
16 molecular descriptors
No *a priori* ideas on which molecular descriptors to select
7 strategies to select 16 MD:
Lasso, Pearson, RFE, PCA1 to PCA4
Dataset ~~split into Internal and~~
Internal dataset
(n = 594)
~~External~~
Validation~~dat~~
Training set (n = 445) & Test set (n = 148)
Standardization
& Tuning the hyper parameters (cv = 5)
Optimized QSRR
applied to
DNN
Performances
R^2 training set - RMSE training set
R^2 test set - RMSE test set
Feature importance analysis
Performances
R^2 ~~external~~ validation set - RMSE

Figure 1. QSRR model development and evaluation of performances.

Table 1. QSRR models selected from the literature review.

References	Type of contaminant	Number of contaminants	MDs selected	Best machine learning algorithms used	RT max measured (min)	R^2 test set	RMSE test set (min)	Percentage of error
Aalizadeh et al. (2019)	Emerging contaminants	1830	LogD[a], CIC1[b], SeigZ[c], RDF020p[d], AlogP[e]	SVM	14.4	0.88	1.04	7%
McEachran et al. (2018)	Environmental contaminants	97	LogP[f], LogD, molecular weight, molecular volume, polar surface area[g], molar refractivity[h], H_donors[i], H_acceptors[j]	ACD /ChromGenius®	40.8	0.92	2.66	6.5%
Bade et al., 2015a, b	Emerging contaminants	544	nDB[k], nTB[l], nC[m], nO[n], nR04-nR09[o], UI[p], Hy[q], MlogP[r], AlogP, logP, logD	MLP	16.5	0.91	0.89	5.4%
Munro et al. (2015)	Pharmaceuticals	166	nDB or nTB, nC or nO, nR04-nR09, UI, Hy, MlogP, AlogP, LogD, nBnz[s], pKa[t]	GRNN	23.2	0.88	1.39	5.9%
Noreldeen et al. (2018)	Veterinary drugs	95	ACDlogP[u], ALOGP, ALOGP2[v], Hy, Ui, ib[w], BEHp1[x], BEHp2[y],GATS1m[z], GATS2m[a2].	MLR	9.3	0.95	0.62	6.6%
Bride et al., 2021	Environmental contaminants	274	logD, DBE[a3], nO, nC, nH, molecular weight, H_donors, logSw[a4]	MLR	14.7	0.76	1.36	9.2%
Yang et al., 2020	Pharmaceuticals	133	XlogP[a5], BCUTp.1h[a6], AATS1i[a7], AATS3i[a8], GATS1e[a9], ALogP, AATSC0p[a10], ETA_EtaP_B[a11], AATS4i[a12], AATS5i[a13]	MLR	15.0	0.63	1.42	9.4%

[a] logD is the measure of hydrophobicity for the ionizable compounds.

[b] CIC1 is the Complementary Information Content index (neighborhood symmetry).

[c] SeigZ is the eigenvalue sum from a Z weighted distance matrix of a Hydrogen-depleted Molecular Graph.

[d] RDF020p is radial distribution function weighted by atomic polarizabilities.

[e] AlogP is logP estimated by the Ghose–Crippen method.

[f] LogP or LogKow, LogP is equal to the logarithm of the ratio of the concentrations of the test substance in octanol and water. This value allows apprehending the hydrophilic or hydrophobic (lipophilic) character of a molecule.

[g] defined as the surface sum over all polar atoms or molecules, primarily oxygen and nitrogen, also including their attached hydrogen atoms.

[h] is a measure of the total polarizability of a mole of a substance.

[i] the number of H-bond donor as descriptors of the H-bonding property.

[j] the number of H-bond acceptor groups as descriptors of the H-bonding property.

[k] number of double bonds.

[l] number of triple bonds.

[m] number of Carbon.

[n] number of Oxygen.

[o] the number of 4–9 membered rings.

[p] unsaturation index.

[q] hydrophilic factor.

[r] Moriguchi logP.

[s] number of benzen groups.

[t] equilibrium constant of the dissociation reaction of an acid species in acid-base reactions.

[u] ACDlogPa molecular properties octanol-water partitioning coefficients.

[v] ALOGP2 molecular properties Ghose-Crippen octanol water coefficient squared.

[w] Ib information indices information bond index.

[x] BEHp1 burden eigenvalue descriptors highest eigenvalue n. 1 of burden matrix/weighted by atomic polarizabilities.

[y] BEHp2 burden eigenvalue descriptors highest eigenvalue n. 2 of burden matrix/weighted by atomic polarizabilities.

[z] GATS1mb 2D autocorrelation descriptors Geary autocorrelation-lag 1/weighted by atomic masses.

[a2] GATS2mb 2D autocorrelation descriptors Geary autocorrelation-lag 2/weighted by atomic masses.

[a3] the double-bond equivalent descriptor is the number of unsaturations present in a organic molecule.

[a4] the water solubility described by the logarithm of water solubility in mg/L at 25 °C.

[a5] XlogP is the constitutional descriptors-describe hydrophobic/hydrophilic properties.

[a6] BCUTp.1h is the BCUT descriptor/nlow highest polarizability weighted BCUTS.

[a7] AATS1i is the autocorrelation descriptor/average Broto-Moreau autocorrelation - lag 1/weighted by first ionization potential.

[a8] AATS3i is the autocorrelation descriptor/average Broto-Moreau autocorrelation - lag 3/weighted by first ionization potential.

[a9] GATS1e is the autocorrelation descriptor/Geary autocorrelation - lag 1/weighted by Sanderson electronegativities.

[a10] AATSC0p is the autocorrelation descriptor/average centered Broto-Moreau autocorrelation - lag 0/weighted by first ionization potential.

[a11] ETA_EtaP_B is the extended topochemical atom descriptor/branching index EtaB relative to molecular size.

[a12] AATS4i is the autocorrelation descriptor/average Broto-Moreau autocorrelation - lag 4/weighted by first ionization potential.

[a13] AATS5i is the autocorrelation descriptor/average Broto-Moreau autocorrelation - lag 5/weighted by first ionization potential.

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 three subsets: a training, a test and a validation dataset (Figure 1). The training dataset was composed of 445 pesticides chosen randomly, their corresponding MD (input) and experimentally measured pesticide RTs (output). The test dataset was composed of 148 pesticides chosen randomly, their corresponding MD (input) and experimentally measured pesticide RTs (output). The training and a test set have a size ratio of three to one, respectively. The validation dataset was composed of 198 randomly chosen pesticides never used before, their corresponding MDs, and experimentally measured pesticide RTs.

Initially, the 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 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): from 1 to 5 hidden layers constituted each by 16 neurons
- 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.

All the models were developed with Python 3.8 from the Python Software Foundation and available at http://www.python.org. In order to optimize and develop the DNN, the Scikit-learn library (https://scikit-learn.org) was used and in particular the sklearn.neural_network module.

2.3. Model validation

The validation of QSRR models is probably the most significant and critical part of model evaluation in order to prevent overfitting in particular. For this reason, we carried out the validation step using the validation dataset never used for the training and testing parts (Nor-eldeen et al., 2018) (Figure 1).

The coefficient of determination (R^2) and the RMSE were used to evaluate and compare the models extracted from the literature review and were measured on the test set (Table 1). 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 R^2 value, the better the model. The R^2 and RMSE were measured, in the case of the models developed in this present study, on the training set (n = 445 pesticides), on the test set (n = 148 pesticides), and on the validation set (n = 198 pesticides) (Table 2).

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 1), and an RMSE of 1 min does not have the same meaning for a gradient of 10 min or for a gradient of 40 min. For this reason, the maximum chromatographic retention time (RT max) was systematic recorded (Tables 1 and 2). The RT max, displayed in Table 2, 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 (R^2) between predicted and experimental RTs was calculated as follows (Eq. (1)):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2} \tag{1}$$

where $\hat{y}_i$ and y_i are the predicted and experimental RTs, respectively, and $\overline{y}_i$ is the mean experimental RT.

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

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{2}$$

where $\hat{y}_i$ and y_i are the predicted and experimental responses, respectively.

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

$$\text{Percentageof error} = (\text{RMSE validation set } \div RT\text{max}\textit{measured}) \times 100 \tag{3}$$

2.4. 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. The input neurons in the input layer are designed to receive the 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) in case of DNN. 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

Table 2. Performances of QSRR models applied to the pesticide dataset.

N° Model	Number of molecular descriptors	Name of the Model	Internal set				Validation set			DNN Optimized			
			Training set		Test set								
			R^2	RMSE	R^2	RMSE	R^2	RMSE	Percentage of error	Number of neurons per hidden layers	Activation function	Solver	Alpha
1	16	*Bade-MLP*	0.95	0.43	0.90	0.63	0.82	0.67	6%	16-16-16-16-16	ReLu	Adam	10
2	16	*Lasso-MLP*	0.60	1.19	0.50	1.27	0.49	1.36	12%	16	tanh	SGD	1
3	16	*Pearson-MLP*	0.79	0.86	0.79	0.83	0.78	0.88	8%	16–16	ReLu	SGD	10
4	16	*RFE-MLP*	0.69	1.04	0.60	1.15	0.63	1.16	10%	16-16-16-16-16	ReLu	SGD	10
5	16	*PCA1-MLP*	0.75	0.94	0.61	1.12	0.64	1.14	10%	16	tanh	Adam	1
6	16	*PCA2-MLP*	0.42	1.44	0.34	1.47	0.38	1.50	13%	16	tanh	Adam	1
7	16	*PCA3-MLP*	0.61	1.18	0.53	1.24	0.56	1.26	11%	16-16-16	ReLu	SGD	10
8	16	*PCA4-MLP*	0.82	0.79	0.75	0.91	0.76	0.93	8%	16-16-16-16	ReLu	SGD	10

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 a training, test and validation datasets, 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 training set. The R^2 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).

3. 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.

3.1. 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., 2015a, 2015b; 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 1). 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 1). 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 R^2 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 R^2 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 R^2 on the training and test set equal to 0.95 and 0.90, respectively (Table 2, Figure S1A & S1B). The RMSE obtained on the training and test set were equal to 0.43 and 0.63. On the validation set, never used for the learning and optimizing process, the R^2 was equal to 0.82 and the RMSE equal to 0.67 (Table 2, Figure S1C). These past results are similar to those obtained by Parinet (2021) with the McEachran 3 MDs, on the validation dataset, and by the use of SVM and MLP as machine learning algorithms where the R^2 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.

3.2. 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 R^2 on the training and test set equal to 0.60 and 0.50, respectively (Table 2, Figure S2A & S2B). The RMSE obtained on the training and test set were equal to 1.19 and 1.27. On the validation set, the R^2 was equal to 0.49 and the RMSE equal to 1.36 (Table 2, Figure S2C). 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 R^2 on the training and test set equal to 0.79 and 0.79, respectively (Table 2, Figure S3A & S3B). The RMSE obtained on the training and test set were equal to 0.86 and 0.83. On the validation set, the R^2 was equal to 0.78 and the RMSE equal to 0.88 (Table 2, Figure S3C). 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 R^2 on the training and test set equal to 0.69 and 0.60, respectively (Table 2, Figure S4A & S4B). The RMSE obtained on the training and test set were equal to 1.04 and 1.15. On the validation set, the R^2 was equal to 0.63 and the RMSE equal to 1.16 (Table 2, Figure S4C). 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 R^2 on the training and test set equal to 0.75 and 0.61, respectively (Table 2, Figure S5A & S5B). The RMSE obtained on the training and test set were equal to 0.94 and 1.12. On the validation set, the R^2 was equal to 0.64 and the RMSE equal to 1.14 (Table 2, Figure S5C). 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 R^2 on the training and test set equal to 0.42 and 0.34, respectively (Table 2, Figure S6A & S6B). The RMSE obtained on the training and test set were equal to 1.44 and 1.47. On the validation set, the R^2 was equal to 0.38 and the RMSE equal to 1.50 (Table 2, Figure S6C). 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 R^2 on the training and test set equal to 0.61 and 0.53, respectively (Table 2, Figure S7A & S7B). The RMSE obtained on the training and test set were equal to 1.18 and 1.24. On the validation set, the R^2 was equal to 0.56 and the RMSE equal to 1.26 (Table 2, Figure S7C). 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 R^2 on the training and test set equal to 0.82 and 0.75, respectively (Table 2, Figure S8A & S8B). The RMSE obtained on the training and test set were equal to 0.79 and 0.91. On the validation set, the R^2 was equal to 0.76 and the RMSE equal to 0.93 (Table 2, Figure S8C). 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.

3.3. 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 2). All the architecture of DNN and theire hyperparameters retained through the gird search for the *models 1 to 8* are listed in Table 2. 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.

4. 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 R^2 of 0.82 and an RMSE of 0.67 min (*Model 1*) on the validation set. However, it could be useful in future research 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.

Declarations

Author contribution statement

Julien Parinet: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This work was supported by Agence Nationale de la Recherche (ANR-19-CE21-0002).

Data availability statement

Data included in article/supplementary material/referenced in article.

Declaration of interests statement

The authors declare no conflict of interest.

Additional information

Supplementary content related to this article has been published online at doi: https://10.1016/j.heliyon.2021.e08563.

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. J. Hazard Mater. 363, 277–285.

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. Sci. Total Environ. 538, 934–941.

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.

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.

Bride, Eloi, Heinisch, Sabine, Bonnefille, Bénilde, Guillemain, Céline, Margoum, Christelle, et al., 2021. Suspect screening of environmental contaminants by UHPLC-HRMS and transposable Quantitative Structure-Retention Relationship modelling. J. Hazardous Mater. 409, 124652. In this issue.

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.

Colosio, C., Rubino, F.M., Moretto, A., 2017. Pesticides. In: International Encyclopedia of Public Health, pp. 454–462.

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. Microchem. J. 106, 51–60.

Ghasemi, J., Saaidpour, S., 2009. QSRR prediction of the chromatographic retention behavior of painkiller drugs. J. Chromatogr. Sci. 47, 156–163.

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. Anal. Chim. Acta 797, 13–19.

May, R., Dandy, G., Maier, H., 2011. Review of input variable selection methods for artificial neural networks. Artif. Neural Network-Method. Adv. Biomed. Appl.

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.

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. J. Chromatogr. A 1396, 34–44.

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. Int. J. Mass Spectrom. 434, 172–178.

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.

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. Anal. Chim. Acta 916, 8–16.

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. Environ. Sci. Technol. 48, 2097–2098.

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. Chemometr. Intell. Lab. Syst. 154, 137–149.

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. J. Expo. Sci. Environ. Epidemiol. 411–426.

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. Anal. Bioanal. Chem. 411, 1421–1431.

Yang, J.J., Han, Y., Mah, C.H., Wanjaya, E., Peng, B., Xu, T.F., Liu, M., Huan, T., Fang, M.L., 2020. Streamlined MRM method transfer between instruments assisted with HRMS matching and retention-time prediction. Anal. Chimica Acta 1100, 88–96.

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. J. Hazard Mater. 383, 121141.