



HAL
open science

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

Julien Parinet

► **To cite this version:**

Julien Parinet. Predicting reversed-phase liquid chromatographic retention times of pesticides by deep neural networks. *Heliyon*, 2021, 7 (12), pp.e08563. 10.1016/j.heliyon.2021.e08563 . anses-03509974

HAL Id: anses-03509974

<https://anses.hal.science/anses-03509974>

Submitted on 8 Jan 2024

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

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



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

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

3 Julien Parinet^{*1}

4

5 ^{*}corresponding author: julien.parinet@anses.fr

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

8

9 **Abstract**

10 To be able to predict reversed phase liquid chromatographic (RPLC) retention times of
11 contaminants is an asset in order to solve food contamination issues. The development of
12 quantitative structure–retention relationship models (QSRR) requires selection of the best
13 molecular descriptors and machine-learning algorithms. In the present work, two main
14 approaches have been tested and compared, one based on an extensive literature review to
15 select the best set of molecular descriptors (16), and a second with diverse strategies in order
16 to select among 1545 molecular descriptors (MD), 16 MD. In both cases, a deep neural
17 network (DNN) were optimized through a gridsearch.

18

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

21

22 **1. Introduction**

23 Contaminants and especially pesticides in food are of growing concern as the general
24 public is increasingly aware about their health effects (Dashtbozorgi et al., 2013). Depending
25 on their concentrations, toxicity, and frequency of detection in food and in the environment,

26 pesticides may lead to health impairment, disease and even death (Colosio et al., 2017).
27 Detecting and quantifying these compounds helps to guarantee compliance of imported goods
28 with the laws and regulations of the importing country (Chiesa et al., 2016).
29 The high accuracy and mass sensitivity of high-resolution mass spectrometry (HRMS)
30 instruments hyphenated to liquid (LC) or gas (GC) chromatography make it possible to
31 observe thousands of chemical features in food and environment samples. These features
32 include monoisotopic exact mass, chromatographic retention time (RT), abundance, isotope
33 profiles and MS² fragmentations. However, data processing and chemical characterization
34 remain difficult despite recent developments. Chemical reference standards and spectral data
35 enable us to confirm the structure of observed characteristics, but reference standards,
36 especially metabolites and by-products, are rarely available for thousands of characteristics in
37 non-target analysis (NTA) and suspect screening analysis (SSA) (McEachran et al., 2018),
38 and having these thousands of standards can also represent a considerable cost.
39 Since the appearance of HRMS, the interest in improving confidence in the identification of
40 small molecules increase, such as pesticides, from putative positive samples based on
41 detection to confirmation (Bade et al., 2015a; Schymanski et al., 2014). SSA studies are those
42 in which observed but unknown features are compared against a database of chemical
43 suspects to identify plausible hits. NTA studies are those in which chemical structures of
44 unknown compounds are postulated without the aid of suspect lists (Sobus et al., 2018). In
45 both cases, confirming the identification of a contaminant requires its standard, which may be
46 unavailable, expensive, or time-consuming to obtain in the case of food poisoning. This is
47 especially true for pesticides where there are a few thousand analytes, metabolites and by-
48 products. In order to increase confidence in the tentative identification of compounds,
49 especially in SSA, it is conceivable to predict their chromatographic retention time (RT)
50 (Bade et al., 2015b; Barron and McEneff, 2016; Parinet, 2021; Randazzo et al., 2016).

51 To predict RT, different strategies using various molecular descriptor (MD) sets and multiple
52 machine-learning algorithms have been tested and published (Aalizadeh et al., 2019; Bade et
53 al., 2015a; Barron and McEneff, 2016; Goryński et al., 2013; McEachran et al., 2018; Munro
54 et al., 2015; Noreldeen et al., 2018; Parinet, 2021; Randazzo et al., 2016). These strategies
55 range from the use of logKow models (Bade et al., 2015b) to more complex in silico
56 approaches based on quantitative structure-retention relationship (QSRR) modeling, including
57 artificial neural networks (ANNs), support vector machines (SVMs), random forest (RF),
58 partial least squares regression (PLS-R), and multilinear regression (MLR) (Ghasemi and
59 Saaidpour, 2009; Munro et al., 2015; Parinet, 2021).

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

72

73

74 **2. Materials and Methods**

75 2.1 Dataset

76 Initially, the dataset included 843 RTs of pesticides collected from the article of Wang et al.
77 (2019). Ultra-high-performance liquid chromatography (UHPLC) gradient conditions, column
78 temperatures, mobile phases, columns, and instruments used to generate the data presented in
79 detail in Wang et al. (2019).

80 Three free software applications have been used in order to compute the pesticide's MD.
81 These applications are free, can calculate a large number of descriptors and are widely
82 available. The ACD software (Advanced Chemistry Development, Toronto, ON, Canada) was
83 used to calculate *LogP* and *LogD*. The Toxicity Estimation Software Tool (TEST, Cincinnati,
84 OH, USA) was used to compute *Hy*, *Ui*, *IB*, *BEHp1*, *BEHp2*, *GATS1m*, and *GATS2m*. The
85 rest of the molecular descriptors (1834 MD) were calculated using the ChemDes online
86 platform (<http://scbdd.com/chemdes/>).

87 Once the MDs were computed, the dataset was cleaned in order to remove constant and
88 missing values (**Figure 1**). Indeed, constant values are useless in order to develop QSRR
89 models and missing values make learning and prediction impossible. The missing values are
90 due to the softwares and their inability to generate, depending on the molecules, the MD. At
91 the end of this curation process, 792 pesticides, their RTs, and 1545 MDs remained in the
92 final dataset. The dataset containing the MDs for each pesticide was then ready to build
93 QSRR models (**Table S1**).

94

95 2.2 QSRR model development

96 The dataset constituted previously and containing the pesticides (792), their MDs (1545), and
97 RTs was used in order to select among them the best MDs inherited from the literature review
98 (*Model 1*). Importantly, in order to find the best set of MDs, a literature review was done by
99 selecting the most recent and pertinent papers with the following criteria: the prediction of
100 retention times measured by RPLC and for pesticides or similar compounds (pharmaceuticals,

101 veterinary drugs). At the end of this literature review, seven articles, their MDs, and models
102 were selected (shown in **Table 1** with their performances) and compared in term of
103 performance measured principally through the *percentage of error*, which is the ratio between
104 the root mean square error (RMSE) divided by the maximum retention time measured on the
105 last eluted compound. In order to pursue the *no a priori approach* on which MD to select
106 (*Model 2 to Model 8*), diverse strategies were used and compared in order to select among the
107 1545 MD, the best sixteen MD. Sixteen MD were retained in order to be able to compare the
108 performances of the models (*Model 2 to 8*) to the model inherited from the literature review
109 (*Model 1*). Hence, the Lasso regression, a regularized linear regression that aims to constrain
110 the coefficients to be close to 0 or equal to zero, thus allowing an automatic selection of the
111 characteristics/MD, here 16 MD (*ATS8m, ATS5i, iedm, SRW10, ATS5v, VR2_Dt, VR1_D,*
112 *VR1_Dt, VR2_D, ATS8i, ATS7i, ATS3i, ATSC3m, ATS0m, ATS0v, ATS4v*). The second
113 strategy was based on the Pearson correlation between the 1545 MD and the output
114 (pesticides RTs), and the larger the relationship and more likely the feature/MD should be
115 selected for modeling, then sixteen MD were selected based on this strategy (*LogP, BEHm4,*
116 *CrippenLogP, ALOGP2, ALOGP, XLOGP2, XLOGP, ATS6p, ATS5p, ATS4p, ATS3p, ATS1p,*
117 *ATS6v, BEHm8, BEHm5, BEHm7*). The third strategy, a recursive feature elimination (RFE),
118 was based on an iterative selection of features/MD made by initially selecting all the MD,
119 then a model is built (here a multi-linear regression), then the least important characteristic is
120 rejected and this process is done until a model with 16 MD is obtained (*maxtsC, MWC2,*
121 *MWC03, MWC4, MWC5, nN, k2, MDEN-23, MDEN-33, MDEO-11, MDEO-12, MDEC-34,*
122 *MDEC-44, MAXDP2, MDEN-22, ieadjmm*). Finally, the fourth strategy was based on
123 principal component analysis (PCA) and declined under four sub strategies (*PCA1 to PCA4*).
124 For the four sub strategies, the same PCA was used. Hence, a PCA was done on the 1545 MD
125 and measured on the 792 pesticides. The MD were normalized (reduced and centered) before

126 doing the PCA and 16 principal components (PC) were retained; *PCA1* strategy was based on
127 the selection of the MD most correlated to each PC, thus 16 MD were selected (*TWC, CICI,*
128 *ETA_Epsilon_2, AATS1p, icyce, MLFER_E, MATS2v, nCl, AATSC3p, R, JGI3, StsC,*
129 *nHCHnX, ATSC6e, MATS6i, MATS6m*). The *PCA2* strategy was based on the selection of the
130 16 MD most correlated to PC1, as PC1 was the PC the most correlated to RT (*TWC, Zagreb,*
131 *nBonds, nBO, MWC01, SRW02, MPC01, ZM1, WTPT-1, SRW04, CID, nHeavyAtom, MPC2,*
132 *nSK, SRW01, BID*). The *PCA3* strategy was based on the selection of the 16 MD most
133 correlated to PC1 (8 MD) and PC4 (8 MD) as PC1 and PC4 were the most correlated to RT
134 (*TWC, Zagreb, nBonds, nBO, MWC01, SRW02, MPC01, ZM1, AATS1p, AATS0p, AATS4p,*
135 *Mp, ETA_AlphaP, AATS3p, AATS5p, AATS2p*). Finally, the *PCA4* strategy was based on the
136 selection of the 16 PC and their corresponding scores used as input (PC1 to PC16).

137 Regardless of the MD dataset used, the following procedure was used. The MD datasets, and
138 the corresponding values of pesticide RTs, were divided into three subsets: a training, a test
139 and a validation dataset (**Figure 1**). The training dataset was composed of 445 pesticides
140 chosen randomly, their corresponding MD (input) and experimentally measured pesticide RTs
141 (output). The test dataset was composed of 148 pesticides chosen randomly, their
142 corresponding MD (input) and experimentally measured pesticide RTs (output). The training
143 and a test set have a size ratio of three to one, respectively. The validation dataset was
144 composed of 198 randomly chosen pesticides never used before, their corresponding MDs,
145 and experimentally measured pesticide RTs.

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

- 149 - Number of hidden layers constituted each by a number of neurons equal to the number
- 150 of MD used as inputs Geron (2017): from 1 to 5 hidden layers constituted each by 16
- 151 neurons
- 152 - The activation function among: ReLu, tanh and logistic
- 153 - The alpha value: 10 or 1
- 154 - The solver function among: Adam, SGD and Lbfgs.

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

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

161

162

163 2.3 Model validation

164 The validation of QSRR models is probably the most significant and critical part of model
165 evaluation in order to prevent overfitting in particular. For this reason, we carried out the
166 validation step using the validation dataset never used for the training and testing parts
167 (Noreldeen et al., 2018) (**Figure 1**).

168 The coefficient of determination (R^2) and the RMSE were used to evaluate and compare the
169 models extracted from the literature review and were measured on the test set (**Table 1**).

170 These parameters were also used for the models developed in this study in order to determine
171 the error between the experimental and predicted RTs in the QSRR models, especially in
172 terms of their ability to be generalized to new pesticide substances with unknown RTs. The
173 lower the RMSE and the higher the R^2 value, the better the model. The R^2 and RMSE were

174 measured, in the case of the models developed in this present study, on the training set (n =
175 445 pesticides), on the test set (n = 148 pesticides), and on the validation set (n = 198
176 pesticides) (**Table 2**).

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

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

- 185 • The coefficient of determination (R^2) between predicted and experimental RTs was
186 calculated as follows (Eq.1):

$$187 \quad R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} \quad (1)$$

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

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

$$192 \quad \text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (2)$$

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

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

$$195 \quad \text{Percentage of error} = (\text{RMSE validation set} \div \text{RT max measured}) \times 100$$

196 (3)

197

198 2.4 Structure of the DNN

199 DNN is a computer program inspired by the biological neural network and designed in order
200 to modelize complex, non-linear problems (classification or regression). A typical DNN is
201 composed of a number of neurons from a few to millions, which are arranged in a series of
202 layers (Zhong et al., 2020). A neuron is a computational unit that has one or more weighted
203 input connections, a transfer function that combines the inputs in some way, and an output
204 connection. The input neurons in the input layer are designed to receive the data, such as the
205 MDs used here, and the output neurons in the last layer are the final predictions made by the
206 DNN, which will be used to compare with the true target data, such as RTs of pesticides.
207 Between the input layer and the output layer are hidden layers, often more than one layer
208 (Zhong et al., 2020) in case of DNN. The input data go into the DNN through the input layer,
209 are then transformed in the hidden layers, and finally become the predictions in the output
210 layer. The values in all neurons in the hidden and output layers are calculated by the
211 application of an activation function on the sum of the values in the previous
212 neurons \times weight+bias calculation, in which weights and biases can be updated based on the
213 errors between the predictions and the target until the errors reach a minimum value. Update
214 of the weights and biases is done through back-propagation of the errors between the target
215 (RT experimental) and the prediction (RT predicted). This process is the “learning” process of
216 DNN. DNNs have two main hyperparameters: the number of neurons per layer, and the
217 number of layers. The number of layers and neurons is also called the “depth” and “width” of
218 DNN, respectively. Larger numbers of layers and neurons mean deeper and wider DNNs,
219 which often have more powerful fitting ability and can achieve better accuracy on the
220 prediction. However, too many layers and neurons can lead to an overfitting problem, which
221 is an accurate prediction on the training set but poorer prediction on the test set. It is crucial
222 for the DNN to be able to generalize on a dataset never seen before. For this last reason, we

223 split the dataset into a training, test and validation datasets, in order to evaluate the capacity of
224 the DNN to generalize. The model development process is hence to develop an optimum
225 architecture of the DNN with an appropriate fitting ability. In this study, our DNN was
226 composed of an input layer, several hidden layers, and an output layer. In each layer, there are
227 numerous neurons accepting values from the neurons of the neighboring layer. In the input
228 and hidden layers, the number of neurons was equal to the number of MDs selected. For
229 instance, if the number was 16 MDs, then there were 16 neurons in the input and in each
230 hidden layer, as suggested by Geron (2017). The number of neurons in the output layer was 1
231 because there was only one RT for each pesticide. The number of neurons in the hidden layers
232 was set manually before the learning process began. Here, we focused on the following
233 hyperparameters: the number of hidden layers, the activation function, the alpha value, and
234 the solver used. We investigated their effects on the performance of the DNN through a
235 gridsearch and a cross-validation (cv=5) process done on the training set. The R^2 and RMSE
236 values were calculated to evaluate the effects of the hyperparameters on the performances of
237 the models developed and on overfitting. A detailed description of the theory behind DNNs
238 has been adequately provided elsewhere (Zhong et al., 2020). Model training was stopped
239 after 1000 epochs (iterations).

240

241 **3. Results and discussion**

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

245 3.1 Comparison of published QSRR models

246 One of the main bottlenecks in designing QSRR models is selecting the MDs (May et al.,
247 2011; Parinet, 2021; Scotti et al., 2016). The selection of the most suitable MDs, among

248 several thousand, can follow various strategies (May et al., 2011); this step is particularly
249 complicated because there are many molecular descriptors that can be calculated and used
250 (Aalizadeh et al., 2019; Bade et al., 2015b, 2015a; McEachran et al., 2018; Munro et al.,
251 2015; Noreldeen et al., 2018) and many strategies to select the MDs.

252 Here, to develop the most accurate QSRR dedicated to pesticides, we used two different
253 approaches. The first approach was based on an extensive literature review on the prediction
254 of RPLC retention times of compounds similar in their structures and properties to pesticides,
255 such as pharmaceuticals and veterinary drugs. Based on this literature review, seven articles
256 emerged (**Table 1**). In order to select the best set of MDs among the seven research papers, a
257 study of the QSRR models developed was carried out. In order to do this, the performances of
258 the QSRR models were documented and compared (**Table 1**). The number of contaminants
259 used to build and optimize the QSRR models was found to be between 95 and 1830
260 compounds, the number of MDs selected was between 5 and 16, and the RT max values
261 measured were between 9.3 and 40.8 min. The machine learning algorithms used were SVM,
262 DNN (MLP and general regression neural networks (GRNN)), and MLR. The performances
263 measured on the test set are for the R^2 between 0.63 and 0.95, and for the RMSE between 0.62
264 and 1.42 min. Nevertheless, the gradients are not similar, reflected by the different RT max
265 measurements. The RMSE and the R^2 alone are not sufficient to determine which MD set and
266 QSRR model is the most efficient. For this reason, we calculated the percentage of error (Eq.
267 3), which was not done in the recent article of Parinet (2021) where all the references
268 selected, and their corresponding MD datasets were applied directly on the pesticides dataset
269 in order to make the prediction of RT. The percentage of error was between 5.4% and 9.4%.
270 The lowest value for the percentage of error was obtained for the QSRR developed by Bade
271 and colleagues (2015) on 544 emerging contaminants and by the use of 16 MDs (*nDB*, *nTB*,
272 *nC*, *nO*, *nR04-nR09*, *UI*, *Hy*, *MLogP*, *ALogP*, *LogP*, *LogD*) and a DNN (MLP). Based on

273 these results, we retained for our QSRR development, the Bade and colleagues (2015) MD set
274 and the MLP as the best ML algorithm to use (*model 1*) with a percentage of error equal to
275 5.4%. Then, we used the MD listed by Bade and colleagues (2015) on our dataset and through
276 a MLP (*Bade-MLP – Model 1*) as described before in the text. By this approach we got a R²
277 on the training and test set equal to 0.95 and 0.90, respectively (**Table 2, Figure S1A & S1B**).
278 The RMSE obtained on the training and test set were equal to 0.43 and 0.63. On the validation
279 set, never used for the learning and optimizing process, the R² was equal to 0.82 and the
280 RMSE equal to 0.67 (**Table 2, Figure S1C**). These past results are similar to those obtained
281 by Parinet (2021) with the McEachran 3 MDs, on the validation dataset, and by the use of
282 SVM and MLP as machine learning algorithms where the R² were between 0.85-0.89 and the
283 RMSE between 0.64-0.69, respectively. The percentage of error obtained thanks to these
284 molecular descriptors and with a MLP was around 6%, which is close to the 5.4% got by
285 Bade and colleagues (2015) on their compounds.

286 3.2 Comparison between QSRR models developed thanks to the literature review and to the 287 *no a priori* approaches

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

291 The performances of *Model 2 (Lasso-MLP)* applied on our pesticide dataset gave R² on the
292 training and test set equal to 0.60 and 0.50, respectively (**Table 2, Figure S2A & S2B**). The
293 RMSE obtained on the training and test set were equal to 1.19 and 1.27. On the validation set,
294 the R² was equal to 0.49 and the RMSE equal to 1.36 (**Table 2, Figure S2C**). The percentage
295 of error obtained thanks to these molecular descriptors and with a MLP was around 12%,
296 which is twice as much as *Model 1 (Bade-MLP)* with 6% on the same compounds.

297 The performances of *Model 3 (Pearson-MLP)* applied on our pesticide dataset gave R^2 on the
298 training and test set equal to 0.79 and 0.79, respectively (**Table 2, Figure S3A & S3B**). The
299 RMSE obtained on the training and test set were equal to 0.86 and 0.83. On the validation set,
300 the R^2 was equal to 0.78 and the RMSE equal to 0.88 (**Table 2, Figure S3C**). The percentage
301 of error obtained thanks to these molecular descriptors and with a MLP was around 8%,
302 which is less good as *Model 1 (Bade-MLP)* with 6% on the same compounds but much better
303 than *Model 2*.

304 The performances of *Model 4 (RFE-MLP)* applied on our pesticide dataset gave R^2 on the
305 training and test set equal to 0.69 and 0.60, respectively (**Table 2, Figure S4A & S4B**). The
306 RMSE obtained on the training and test set were equal to 1.04 and 1.15. On the validation
307 set, the R^2 was equal to 0.63 and the RMSE equal to 1.16 (**Table 2, Figure S4C**). The
308 percentage of error obtained thanks to these molecular descriptors and with a MLP was
309 around 10%, which is less good as *Model 1 (Bade-MLP)* with 6% on the same compounds,
310 and less good as *Model 3*.

311 The performances of *Model 5 (PCA1-MLP)* applied on our pesticide dataset gave R^2 on the
312 training and test set equal to 0.75 and 0.61, respectively (**Table 2, Figure S5A & S5B**). The
313 RMSE obtained on the training and test set were equal to 0.94 and 1.12. On the validation set,
314 the R^2 was equal to 0.64 and the RMSE equal to 1.14 (**Table 2, Figure S5C**). The percentage
315 of error obtained thanks to these molecular descriptors and with a MLP was around 10%,
316 which is less good as *Model 1 (Bade-MLP)* with 6% on the same compounds, and quite
317 similar to *Model 4*.

318 The performances of *Model 6 (PCA2-MLP)* applied on our pesticide dataset gave R^2 on the
319 training and test set equal to 0.42 and 0.34, respectively (**Table 2, Figure S6A & S6B**). The
320 RMSE obtained on the training and test set were equal to 1.44 and 1.47. On the validation
321 set, the R^2 was equal to 0.38 and the RMSE equal to 1.50 (**Table 2, Figure S6C**). The

322 percentage of error obtained thanks to these molecular descriptors and with a MLP was
323 around 13%, which is less good as *Model 1 (Bade-MLP)* with 6% on the same compounds,
324 and the worst model developed with performances quite similar to *Model 2*.

325 The performances of *Model 7 (PCA3-MLP)* applied on our pesticide dataset gave R^2 on the
326 training and test set equal to 0.61 and 0.53, respectively (**Table 2, Figure S7A & S7B**). The
327 RMSE obtained on the training and test set were equal to 1.18 and 1.24. On the validation
328 set, the R^2 was equal to 0.56 and the RMSE equal to 1.26 (**Table 2, Figure S7C**). The
329 percentage of error obtained thanks to these molecular descriptors and with a MLP was
330 around 11%, a little better than *Model 5* but which is less good as *Model 1 (Bade-MLP)* with
331 6% on the same compounds.

332 The performances of *Model 8 (PCA4-MLP)* applied on our pesticide dataset gave R^2 on the
333 training and test set equal to 0.82 and 0.75, respectively (**Table 2, Figure S8A & S8B**). The
334 RMSE obtained on the training and test set were equal to 0.79 and 0.91. On the validation set,
335 the R^2 was equal to 0.76 and the RMSE equal to 0.93 (**Table 2, Figure S8C**). The percentage
336 of error obtained thanks to these molecular descriptors and with a MLP was around 8%, better
337 than all the models developed thanks to the PCA approach and similar in term of
338 performances to *Model 3*, but still less good as *Model 1 (Bade-MLP)*.

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

344 3.3 Optimization of the hyperparameters

345 The QSRR models were optimized using an MLP through a gridsearch process. Nevertheless,
346 the number of neurons per hidden layers was set manually and was determined by applying

347 the recommendations of Geron (2017). Importantly, Geron mentions that the common
348 practice of sizing the hidden layers to form a funnel, with an ever-decreasing number of
349 neurons at each layer is no longer as common, and instead we can simply give the same size
350 to all the hidden layers, resulting in only one hyperparameter to adjust instead of one per
351 layer. Nonetheless, it is more useful, still according to Geron (2017), to increase the number
352 of layers rather than the number of neurons per layer. For this reason, the number of hidden
353 layers used by the gridsearch was between 1 to 5 layers, irrespective of the QSRR.
354 Once the number of neurons per hidden layer and the number of hidden layers are set, there
355 remains a large number of hyperparameters to optimize. Nevertheless, some of them are more
356 important than others, such as the activation function and the solver used. For this reason, the
357 gridsearch for the activation function was done among the following functions: ReLu, tanh,
358 and logistic. A gridsearch was also carried out to select the best solver among three possible
359 choices (Adam, SGD and Lbfgs). The last hyperparameter to optimize through the gridsearch
360 was the alpha value, which is a regularization parameter (L2 regularization); alpha value was
361 comprised between 0.01 and 100 (**Table 2**). All the architecture of DNN and their
362 hyperparameters retained through the gridsearch for the *models 1 to 8* are listed in **Table 2**.
363 Hence, the number of layers are comprised between 1 to 5, two activation functions among
364 three were used (ReLu and tanh) and the logistic function was never retained by the
365 gridsearch, two solver (Adam and SGD) among three were used. Finally, despite the
366 amplitude values of alpha, two alpha values were retained: 1 and 10.

367

368 **4. Conclusions**

369 We compared a literature review approach to a no *a priori* approach in order to select, by
370 diverse strategies, the best set of molecular descriptors among 1545 MD in order to predict,
371 through a QSRR model, the RPLC retention times of 792 pesticides. The literature review

372 approach yielded the best results when DNN was used as the ML algorithm, with an R² of
373 0.82 and an RMSE of 0.67 min (*Model 1*) on the validation set. However, it could be useful in
374 future resaerch to test some other *no a priori* selection strategies in order to determine new
375 MD datasets and also to consider reducing the number of MD with the goal to simplify the
376 models while obtaining good predictions.

377

378

379 **Tables**

380 **Table 1** QSRR models selected from the literature review

381 **Table 2** Performances of QSRR models applied to the pesticide dataset

382

383 **Figures**

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

385

386 **Funding sources**

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

389

390 **Conflicts of interest**

391 The authors declare that they have no conflicts of interest.

392

393 **Acknowledgements**

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

396

397

398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422

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>

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

425 Dashtbozorgi, Z., Golmohammadi, H., Konozi, E., 2013. Support vector regression based
426 QSPR for the prediction of retention time of pesticide residues in gas chromatography–
427 mass spectroscopy. *Microchemical Journal* 106, 51–60.
428 <https://doi.org/10.1016/j.microc.2012.05.003>

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

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

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

440 McEachran, A.D., Mansouri, K., Newton, S.R., Beverly, B.E.J., Sobus, J.R., Williams, A.J.,
441 2018. A comparison of three liquid chromatography (LC) retention time prediction
442 models. *Talanta* 182, 371–379. <https://doi.org/10.1016/j.talanta.2018.01.022>

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

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

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

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

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

464 Scotti, M.T., Scotti, L., Ishiki, H.M., Peron, L.M., de Rezende, L., do Amaral, A.T., 2016.
465 Variable-selection approaches to generate QSAR models for a set of antichagasic
466 semicarbazones and analogues. *Chemometrics and Intelligent Laboratory Systems* 154,
467 137–149. <https://doi.org/10.1016/j.chemolab.2016.03.023>

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

473 Wang, J., Chow, W., Wong, J.W., Leung, D., Chang, J., Li, M., 2019. Non-target data
474 acquisition for target analysis (nDATA) of 845 pesticide residues in fruits and vegetables
475 using UHPLC/ESI Q-Orbitrap. *Analytical and Bioanalytical Chemistry* 411, 1421–1431.
476 <https://doi.org/10.1007/s00216-019-01581-z>

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

481

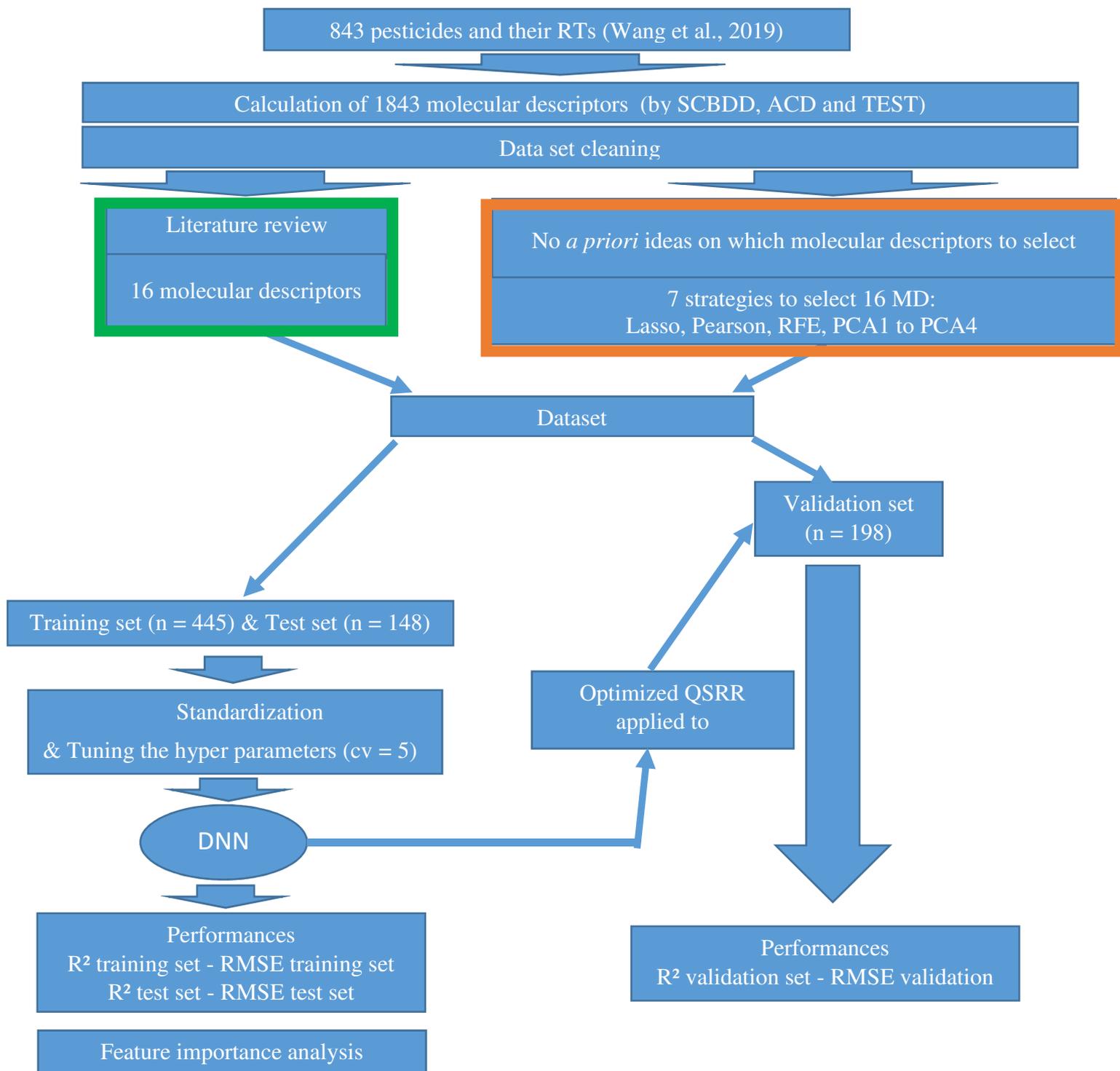


Figure 1 QSRR model development and evaluation of performances

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., 2015	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., in press	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

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	Percentage of error	Number of neurons per hidden layers	Activation function	Solver	Alpha
			R ²	RMSE	R ²	RMSE							
1	16	<i>Bade-MLP</i>	0.95	0.43	0.90	0.63	0.82	0.67	6%	16-16-16-16-16	ReLu	Adam	10
2	16	<i>Lasso-MLP</i>	0.60	1.19	0.50	1.27	0.49	1.36	12%	16	tanh	SGD	1
3	16	<i>Pearson-MLP</i>	0.79	0.86	0.79	0.83	0.78	0.88	8%	16-16	ReLu	SGD	10
4	16	<i>RFE-MLP</i>	0.69	1.04	0.60	1.15	0.63	1.16	10%	16-16-16-16-16	ReLu	SGD	10
5	16	<i>PCA1-MLP</i>	0.75	0.94	0.61	1.12	0.64	1.14	10%	16	tanh	Adam	1
6	16	<i>PCA2-MLP</i>	0.42	1.44	0.34	1.47	0.38	1.50	13%	16	tanh	Adam	1
7	16	<i>PCA3-MLP</i>	0.61	1.18	0.53	1.24	0.56	1.26	11%	16-16-16	ReLu	SGD	10
8	16	<i>PCA4-MLP</i>	0.82	0.79	0.75	0.91	0.76	0.93	8%	16-16-16-16	ReLu	SGD	10