Journal of Mathematical Chemistry (2024) 62:2373–2387
https://doi.org/10.1007/s10910-023-01517-0

ORIGINAL PAPER

The prediction of the retention time of pesticide based on the Monte Carlo method with the use of the vector of the ideality of correlation and correlation weights of local symmetry fragments

Alla P. Toropova[1] · Andrey A. Toropov[1] · Ivan Raska Jr.[2] · Maria Raskova[2] · Ramon Carbó-Dorca[3,4]

Received: 27 June 2023 / Accepted: 3 September 2023 / Published online: 23 September 2023

Abstract

Recently, the retention time of pesticides has been considered an informative indicator of the ecological quality of pesticides. Two new possibilities are proposed for building pesticide retention time models using the CORAL program (http://www.insilico.eu/coral). Firstly, the possibility of being involved in modelling the correlation weights of local symmetry fragments in SMILES. Secondly, using two criteria of predictive potential (correlation ideality index and correlation intensity index) as a vector in Monte Carlo optimization for model building. Building models of the retention time of pesticides using the CORAL software confirms the effectiveness of these innovations.

Keywords Pesticide · Retention time · QSPR/QSAR · Monte Carlo method · CORAL software

✉ Andrey A. Toropov
andrey.toropov@marionegri.it

[1] Laboratory of Environmental Chemistry and Toxicology, Department of Environmental Health Science, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Via Mario Negri 2, 20156 Milan, Italy

[2] 3rd Department of Medicine - Department of Endocrinology and Metabolism, First Faculty of Medicine, Charles University in Prague and General University Hospital in Prague, U Nemocnice 1, 12808 Prague 2, Czech Republic

[3] Ronin Institute, 127 Haddon Pl, Montclair, NJ 07043, USA

[4] Institute of Computational Chemistry and Catalysis, University of Girona, Campus Montilivi, 17071 Girona, Spain

1 Introduction

The prediction of the physicochemical and biochemical behavior of different chemicals is one problem which can be solved via special conceptions of mathematical and theoretical chemistry. It can be related to everyday life [1] or, drug discovery [2], or diverse tasks on the border of mathematics, chemistry, and biology [3]. Pesticides have broad applications. Most of the world's food crops are protected by pesticides, which are important agricultural tools [4–7]. However, as a rule, these substances are dangerous in toxicological and ecological aspects. The retention time is a comprehensive criterion for assessing a substance's toxicity (pesticide). The traditional methods used to analyze pesticide pollution need complex equipment and expensive consultations with experts [4, 5]. This stimulates searching for retention time prediction quantitative structure–property/activity relationships (QSPR/QSAR) for unknown substances based on available databases [6–9].

Monte Carlo methods are one of the directions in searching for algorithms for predicting the retention time of pesticides [6] using a simplified molecular input-line entry system (SMILES) as an interpretation of molecular diversity [10]. It should be noted that SMILES notation has effective applications in building valuable QSAP medicinal chemistry models [11, 12].

Symmetry has many manifestations in reality [13, 14]. In this work, an attempt to improve the model of pesticide retention times was built up using the CORAL software [6] via the local symmetry fragments in SMILES, interpreting these as a certain molecular feature capable of influencing retention times. In addition, to this end, the so-called vector of the ideality of correlation is applied in the Monte Carlo calculations via the CORAL software (http://www.insilico.eu/coral/). The vector has two components: the index of ideality of correlation (*IIC*) and the correlation intensity index (*CII*). These values were studied [15–17], and their relationships with the predictive potential were detected. It should be noted that *IIC* and *CII* are sensitive to completely different features of correlations [17]. In other words, the above components (the symmetry fragments and vector of ideality of correlation) are likely effective tools for improving the predictive potential of QSPR models.

2 Computational details

2.1 Data

Experimental data on the retention time of pesticides taken in the literature [6]. After removing duplicates, the work set of pesticides contains 359 compounds (Table S1 in the *Supplementary material*). The list of six duplicates which were deleted is the following:

ID = 40 SMILES = CCNc1nc(C)c(CCCC)c(OS(=O)(=O)N(C)C)n1.

ID = 68 SMILES = Cl[C@@H]1[C@H](Cl)[C@H](Cl)[C@H](Cl)[C@@H](Cl)[C@H]1Cl.

ID = 98 SMILES = FC(F)(F)c1cccc(c1)N2CC(CCl)C(Cl)C2 = O.

ID = 109 SMILES = ClC1 = C(Cl)[C@]3(Cl)C(Cl)(Cl)[C@@]1(Cl)[C@@H]2CO[S +]([O-])OC[C@@H]23 ID = 209 SMILES = Clc2cc(OP(= S)(OC)c1ccccc1)c(Cl)cc2Br.

ID = 250 SMILES = Clc2c(C(= O)NCc1ccc(cc1)C(C)(C)C)n(C)nc2CC.

These 359 compounds were randomly split in equivalent percentages into the active training set, the passive training set, the calibration set, and the validation set. Five such distributions were considered the basis for developing five versions of the pesticide retention time model (Table S2 in the *Supplementary material*). Table 1 shows that these distributions are quite diverse. These distributions were done using the CORAL software (http://www.insilico.eu/coral).

2.2 The optimal descriptor

The retention time (Rt) models examined here are the following generalized representations via linear regression:

$$\mathrm{Rt} = \mathrm{C}_0 + \mathrm{C}_1 \times DCW(T, N) \tag{1}$$

The SMILES-based optimal descriptor is calculated as follows:

$$DCW(\mathrm{T}, \mathrm{N}) = \sum CW(S_k) + \sum CW(SS_k) \tag{2}$$

$$DCW(\mathrm{T}, \mathrm{N}) = \sum CW(S_k) + \sum CW(SS_k) + \sum CW(Symm) \tag{3}$$

In the above equations, S_k and SS_k are the SMILES attributes applied to building the model for retention time. S_k is a SMILES atom, i.e. one symbol ('C', 'c',' = ') or a group of symbols that cannot be examined separately ('Cl', '@@', %12). SS_k corresponds to consecutive pairs of SMILES-atoms. '*Symm*' are fragments of local symmetry represented by configurations XYX, XYYX, and XYZYX. Table 2 contains an example of the fragments of local symmetry.

CW(x) are correlation weights of *x*, which can be S_k, SS_k, or *Symm*.

Table 1 The matrix of the percentage of identity of the five considered distributions (splits)

i/j[a]	1	2	3	4	5
1	0	44.3	34.6	37.8	44.2
2	33.7	0	37.8	34.4	42.9
3	36.5	38.0	0	38.9	43.2
4	41.3	38.4	34.4	0	44.0
5	45.1	40.0	40.4	38.7	0

[a] The i and j mean the numbering of the 5 examined splits. The matrix element [i, j] if i > j contains the percentage of identity for the active training sets, and if i < j, contains the percentage of identity for the validation sets (external sets)

Table 2 Examples of fragments of local symmetry for "Clc1cc(cc(Cl)c1)C(=O)NC(C)(C)C#C"

Type	Observed fragment of local symmetry	Designation in the CORAL
XYX	c1c c(c C(C C)C C#C	[xyx5]…
XYYX	(cc(	[xyyx1]…
XYZYX	absent	[xyzyx0]…

T is the threshold to distribute SMILES attributes available to analyze in two categories, rare (inactive) and active (non-rare). The correlation weights of inactive attributes fixed zero values (i.e., these were removed from the simulation process). *N* is the number of iterations of the Monte Carlo optimization. Table 3 contains an example of the *DCW(3,15)* calculation for split 1, Eq. 11.

2.3 The Monte Carlo optimization

Equation 1 needs the numerical data on the correlation weights (*CW*). Monte Carlo optimization provides those correlation weights (Table S3 in the *Supplementary material*). Two target functions are examined:

$$TF_1 = \mathrm{r}_{\mathrm{AT}} + \mathrm{r}_{\mathrm{PT}} - \left|\mathrm{r}_{\mathrm{AT}} - \mathrm{r}_{\mathrm{PT}}\right| \times 0.1 \tag{4}$$

$$TF_2 = TF_1 + IIC_C \times 0.3 + CII_C \times 0.3 \tag{5}$$

The r_{AT} and r_{PT} are correlation coefficients between the observed and predicted values of endpoints for the active and passive training sets, respectively.

IIC is the index of the ideality of correlation [15]. IIC_C is calculated with data on the calibration set as follows:

$$IIC_C = r_{CLB}\frac{\min({}^{-}MAE_C, {}^{+}MAE_C)}{max({}^{-}MAE_C, {}^{+}MAE_C)} \tag{6}$$

where the following symbols are used:

$${}^{-}MAE_C = \frac{1}{{}^{-}N}\sum_{k=1}^{{}^{-}N}\left|\Delta_k\right|, \Delta_k < 0; {}^{-}N is\, the\, number\, of\, \Delta_k < 0 \tag{7}$$

$${}^{+}MAE_C = \frac{1}{{}^{+}N}\sum_{k=1}^{{}^{+}N}\left|\Delta_k\right|, \Delta_k \geqslant 0; {}^{+}N is\, the\, number\, of\, \Delta_k \geqslant 0 \tag{8}$$

$$\Delta_k = observed_k - calculated_k \tag{9}$$

Table 3 An example of the calculation retention time for a molecule represented by SMILES = CCc1nc(OCC)cc(OP(= S)(OC)OC)n1 (split 1)

Molecular features extracted from SMILES	Correlation weights (*CW*) of molecular features	Frequency in active training set	Frequency in passive training set	Frequency in calibration set
C……	0.0765	77	80	88
C……	0.0765	77	80	88
c……	0.2037	71	70	68
1……	0.7796	84	78	78
n……	0.1354	17	19	22
c……	0.2037	71	70	68
(……	– 0.0908	90	85	90
O……	0.3240	73	73	74
C……	0.0765	77	80	88
C……	0.0765	77	80	88
(……	– 0.0908	90	85	90
c……	0.2037	71	70	68
c……	0.2037	71	70	68
(……	– 0.0908	90	85	90
O……	0.3240	73	73	74
P……	0.2556	23	23	32
(……	– 0.0908	90	85	90
=……	0.2580	68	64	68
S……	0.4323	32	35	31
(……	– 0.0908	90	85	90
(……	– 0.0908	90	85	90
O……	0.3240	73	73	74
C……	0.0765	77	80	88
(……	-0.0908	90	85	90
O……	0.3240	73	73	74
C……	0.0765	77	80	88
(……	– 0.0908	90	85	90
n……	0.1354	17	19	22
1……	0.7796	84	78	78
C…C……	0.3426	46	53	45
c…C……	0.0967	13	8	7
c…1……	0.5316	65	66	65
n…1……	– 0.0134	10	13	14
n…c……	0.6392	16	19	22
c…(……	0.2507	60	63	61
O…(……	0.4353	56	65	63
O…C……	0.0975	42	49	57
C…C……	0.3426	46	53	45
C…(……	0.0882	76	75	88
c…(……	0.2507	60	63	61

Table 3 (continued)

Molecular features extracted from SMILES	Correlation weights (*CW*) of molecular features	Frequency in active training set	Frequency in passive training set	Frequency in calibration set
c…c……	0.3438	67	64	56
c…(……	0.2507	60	63	61
O…(……	0.4353	56	65	63
P…O……	0.4390	15	16	20
P…(……	0.0440	22	23	32
=…(……	– 0.1924	55	56	58
S…=……	0.4477	16	20	21
S…(……	0.6822	27	30	23
(…(……	0.2683	54	47	59
O…(……	0.4353	56	65	63
O…C……	0.0975	42	49	57
C…(……	0.0882	76	75	88
O…(……	0.4353	56	65	63
O…C……	0.0975	42	49	57
C…(……	0.0882	76	75	88
n…(……	0.5403	9	9	16
n…1……	-0.0134	10	13	14
[xyx0]……	0.3074	7	5	5
[xyyx0]……	0.7754	78	80	83
[xyzyx0]……	– 0.2428	70	72	68
DCW(3,15)=	13.0083			

The *observed* and *calculated* values of the endpoint.

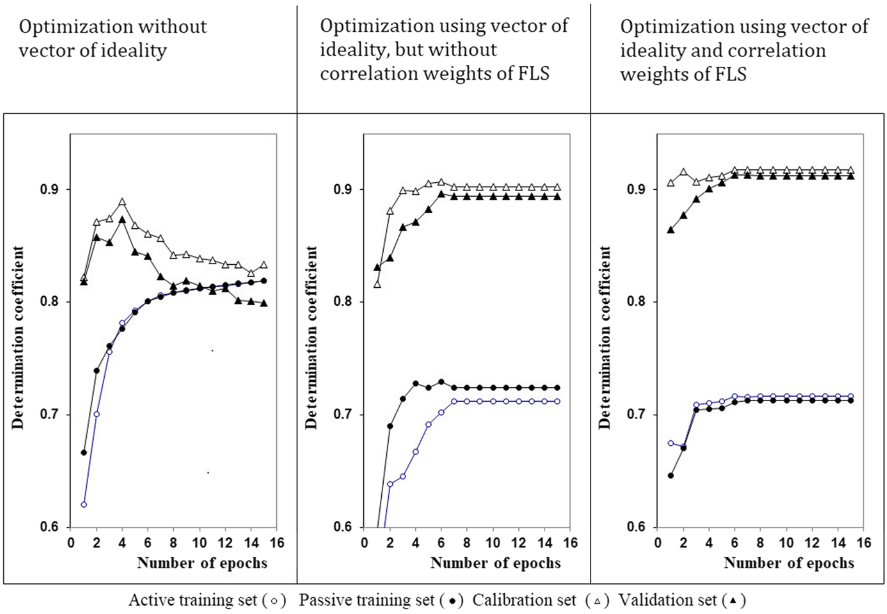
CII is the correlation intensity index [16]. CII_C is calculated as follows

$$CII_C = 1 - \sum \left(\Delta R_j^2 > 0 \right); \text{the} \Delta R_j^2 = R_j^2 - R^2 \tag{10}$$

The R^2 is the correlation coefficient between the experimental and predicted value of the endpoint for the calibration set; the $R^2{}_j$ is the value of the correlation coefficient between the experimental and predicted value of the endpoint for the calibration set if the j-th substance is removed.

It has been shown that these values are completely different [17] but using them together makes it possible to improve the predictive potential of the models [16]. Therefore, here these criteria are used as a vector for the Monte Carlo optimisation of the correlation weights of SMILES fragments.

Figure 1 shows the histories of the Monte Carlo optimization with different target functions. The optimization without the vector of ideality gives the overtraining about 5–7 epochs of the calculation. Using the vector of idealization builds up the model practically without overtraining. The optimization with the vector


Optimization without vector of ideality

Optimization using vector of ideality, but without correlation weights of FLS

Optimization using vector of ideality and correlation weights of FLS

Active training set (○) Passive training set (●) Calibration set (△) Validation set (▲)

Fig. 1 Histories of the Monte Carlo optimization under different conditions

of idealization, together with correlation weights of fragments of local symmetry (FLS), improves the predictive potential of the model.

2.4 Applicability domain

Applicability domain of the CORAL model is defined according to the distribution of SMILES attributes in the active training and calibration sets [15–17].

2.5 Mechanistic interpretation

Mechanistic interpretation of the CORAL-model maybe represented by SMILES attributes, which play roles of promoters of increase and decrease for an endpoint [15–17].

3 Results and discussion

3.1 QSPR models

Table 4 demonstrates the statistical characteristics of models obtained without the vector of ideality of correlation and correlation weights of fragments of local

Table 4 Models built up without the vector of ideality of correlation and correlation weights of fragments of local symmetry: *DCW(3,4)* was used for these models

Split	Set	*N*	R^2	*CCC*	*IIC*	*CII*	Q^2	*RMSE*	*MAE*	*F*
1	A	92	0.8519	0.9200	0.7420	0.9061	0.8455	2.69	1.91	518
	P	87	0.8521	0.8932	0.6895	0.8986	0.8463	2.93	2.36	490
	C	90	0.6945	0.8251	0.6370	0.8400	0.6776	3.09	2.48	200
	V	90	0.6651	–	–	–	–	3.32	–	–
2	A	93	0.8316	0.9081	0.7510	0.8960	0.8249	2.82	2.08	449
	P	90	0.8315	0.8935	0.7426	0.8928	0.8249	2.69	2.11	434
	C	88	0.8352	0.9063	0.7423	0.9147	0.8262	2.68	2.22	436
	V	88	0.6530	–	–	–	–	3.17	–	–
3	A	87	0.8412	0.9138	0.7452	0.8981	0.8346	2.70	1.95	450
	P	91	0.8412	0.9106	0.6891	0.8968	0.8349	2.82	2.18	471
	C	90	0.7664	0.8062	0.6123	0.8708	0.7500	4.01	2.96	289
	V	91	0.7893	–	–	–	–	3.67	–	–
4	A	93	0.8375	0.9115	0.6907	0.8842	0.8319	2.76	1.96	469
	P	88	0.8398	0.9006	0.7201	0.9114	0.8317	2.73	2.11	451
	C	89	0.7857	0.8475	0.5730	0.8916	0.7748	3.45	2.65	319
	V	89	0.7456	–	–	–	–	3.37	–	–
5	A	89	0.8487	0.9181	0.7520	0.9026	0.8422	2.83	2.01	488
	P	89	0.8487	0.9162	0.7930	0.8981	0.8427	2.47	1.85	488
	C	89	0.6459	0.7849	0.7031	0.8012	0.6290	3.71	2.82	159
	V	92	0.7744	–	–	–	–	3.09	–	–

A active training set, *P* passive training set, *C* calibration set, *V* validation set, *n* the number of compounds in a set, R^2 determination coefficient, *CCC* concordance correlation coefficient, *IIC* index of ideality of correlation, *CII* correlation intensity index, Q^2 cross-validated R^2, *RMSE* root mean squared error, *MAE* mean absolute error, *F* Fischer F-ratio

symmetry. The average determination coefficient for the external validation set is 0.7255 ± 0.0562.

Table 5 contains the statistical quality of models obtained with the vector of ideality of correlation and without correlation weights of fragments of local symmetry. The average value of the determination coefficient for the external validation set is 0.8866 ± 0.0068.

Table 6 demonstrates the statistical characteristics of models observed in the case of the use vector of ideality of correlation and correlation weights of fragments of local symmetry. The average determination coefficient for the external validation set is 0.9127 ± 0.0084.

The best approach is represented by the following regression models related to five random splits:

$$Rt = -1.599(\pm 0.155) + 1.675(\pm 0.010) * DCW(3,15) \tag{11}$$

$$Rt = -1.896(\pm 0.171) + 1.143(\pm 0.0080) * DCW(3,15) \tag{12}$$

Table 5 Models were built up using the vector of ideality of correlation but without correlation weights of fragments of local symmetry

Split	Set	*n*	R^2	*CCC*	*IIC*	*CII*	Q^2	*RMSE*	*MAE*	*F*
1	A	92	0.7051	0.8270	0.7697	0.8421	0.6916	3.79	3.21	215
	P	87	0.7326	0.8134	0.8359	0.8323	0.7217	3.75	3.24	233
	C	90	0.9073	0.9492	0.9525	0.9509	0.9027	1.61	1.35	861
	V	90	0.8744	–	–	–	–	1.87	–	–
2	A	93	0.7086	0.8295	0.7892	0.8413	0.6966	3.71	3.13	221
	P	90	0.7217	0.8223	0.6850	0.8369	0.7094	3.45	2.86	228
	C	88	0.9101	0.9501	0.9538	0.9483	0.9046	1.89	1.49	870
	V	88	0.8922	–	–	–	–	1.82	–	–
3	A	87	0.6595	0.7948	0.7580	0.8163	0.6449	3.96	3.32	165
	P	91	0.7202	0.8421	0.7063	0.8417	0.7073	3.43	2.91	229
	C	90	0.8757	0.9264	0.9358	0.9236	0.8707	2.08	1.71	620
	V	91	0.8917	–	–	–	–	1.94	–	–
4	A	93	0.6861	0.8139	0.8107	0.8232	0.6731	3.84	3.18	199
	P	88	0.7387	0.8213	0.6386	0.8601	0.7264	3.56	3.05	243
	C	89	0.9281	0.9615	0.9632	0.9573	0.9248	1.50	1.24	1123
	V	89	0.8838	–	–	–	–	1.78	–	–
5	A	89	0.7157	0.8343	0.6019	0.8307	0.7045	3.89	3.20	219
	P	89	0.7362	0.8378	0.5998	0.8470	0.7245	3.50	2.92	243
	C	89	0.8829	0.9372	0.9396	0.9379	0.8766	1.84	1.41	656
	V	92	0.8911	–	–	–	–	1.86	–	–

A active training set, *P* passive training set, *C* calibration set, *V* validation set, *n* the number of compounds in a set, R^2 determination coefficient, *CCC* concordance correlation coefficient, *IIC* index of ideality of correlation, *CII* correlation intensity index, Q^2 cross-validated R^2, *RMSE* root mean squared error, *MAE* mean absolute error, *F* Fischer F-ratio

$$Rt = -2.962(\pm 0.189) + 1.329(\pm 0.0099) * DCW(3,15) \tag{13}$$

$$Rt = -0.618(\pm 0.145) + 1.850(\pm 0.0111) * DCW(3,15) \tag{14}$$

$$Rt = -3.439(\pm 0.163) + 1.797(\pm 0.0100) * DCW(3,15) \tag{15}$$

The clustering of the correlations of experimental values with calculated retention time values is a curious and non-trivial situation since it does not agree with the traditional conception that the model may be classified as satisfactory if it is good for the training set. The vector of ideality of correlation improves the statistical characteristics of the models for the calibration and validation sets. However, the statistical characteristics of the training set become poorer. One can see the correlation between observed and predicted retention time represented by two separate correlations (Fig. 2). The generalized determination coefficient seems non-attractive despite each separate correlation being quite good. The same situation repeats for all other random splits examined here.

Table 6 Models were built up using the vector of ideality of correlation and correlation weights of fragments of local symmetry

Split	Set	*n*	R^2	*CCC*	*IIC*	*CII*	Q^2	*RMSE*	*MAE*	*F*
1	A	92	0.7429	0.8525	0.8252	0.8565	0.7321	3.54	3.03	260
	P	87	0.7539	0.8191	0.6804	0.8358	0.7435	3.72	3.19	260
	C	90	0.9159	0.9569	0.9569	0.9512	0.9117	1.41	1.14	958
	V	90	0.9134	–	–	–	–	1.52	-	-
2	A	93	0.7081	0.8291	0.8236	0.8523	0.6954	3.71	3.17	221
	P	90	0.7187	0.8147	0.5795	0.8257	0.7070	3.54	2.87	225
	C	88	0.9114	0.9517	0.9547	0.9478	0.9057	1.81	1.43	885
	V	88	0.9022	–	–	–	–	1.68	–	–
3	A	87	0.7039	0.8262	0.6506	0.8313	0.6903	3.69	3.07	202
	P	91	0.7300	0.8507	0.8472	0.8459	0.7187	3.35	2.70	241
	C	90	0.9001	0.9437	0.9487	0.9511	0.8931	1.77	1.37	793
	V	91	0.9227	–	–	–	–	1.66	–	–
4	A	93	0.7181	0.8359	0.7288	0.8338	0.7072	3.64	3.09	232
	P	88	0.7447	0.8259	0.6422	0.8578	0.7334	3.51	2.92	251
	C	89	0.9322	0.9647	0.9653	0.9641	0.9277	1.45	1.16	1196
	V	89	0.9209	–	–	–	–	1.65	–	–
5	A	89	0.7263	0.8415	0.7616	0.8412	0.7161	3.81	3.15	231
	P	89	0.7258	0.8303	0.7090	0.8441	0.7140	3.56	3.04	230
	C	89	0.9171	0.9550	0.9566	0.9574	0.9116	1.56	1.22	962
	V	92	0.9042	–	–	–	–	1.69	–	–

A active training set, *P* passive training set, *C* calibration set, *V* validation set, *n* the number of compounds in a set, R^2 determination coefficient, *CCC* concordance correlation coefficient, *IIC* index of ideality of correlation, *CII* correlation intensity index, Q^2 cross-validated R^2, *RMSE* root mean squared error, *MAE* mean absolute error, *F* Fischer F-ratio

3.2 Applicability domain

The percentage of outliers, accordingly, the values of the statistical defect [15–17], is approximately 10%.

3.3 Mechanistic interpretation

The mechanistic interpretation of models is reduced to identifying lists of molecular features extracted from SMILES with only positive or negative correlation weights. The first with positive correlation weights is increase promoters; the second with negative correlation weights is decrease promoters for endpoint values. Table 7 contains a collection of correlation weights for some SMILES-attributes observed in three runs of the Monte Carlo optimization (split 1). The analysis of these data allows us to suggest that the presence of cycles promotes an increase in the retention time of pesticides (1…); the absence of fragments of local symmetry

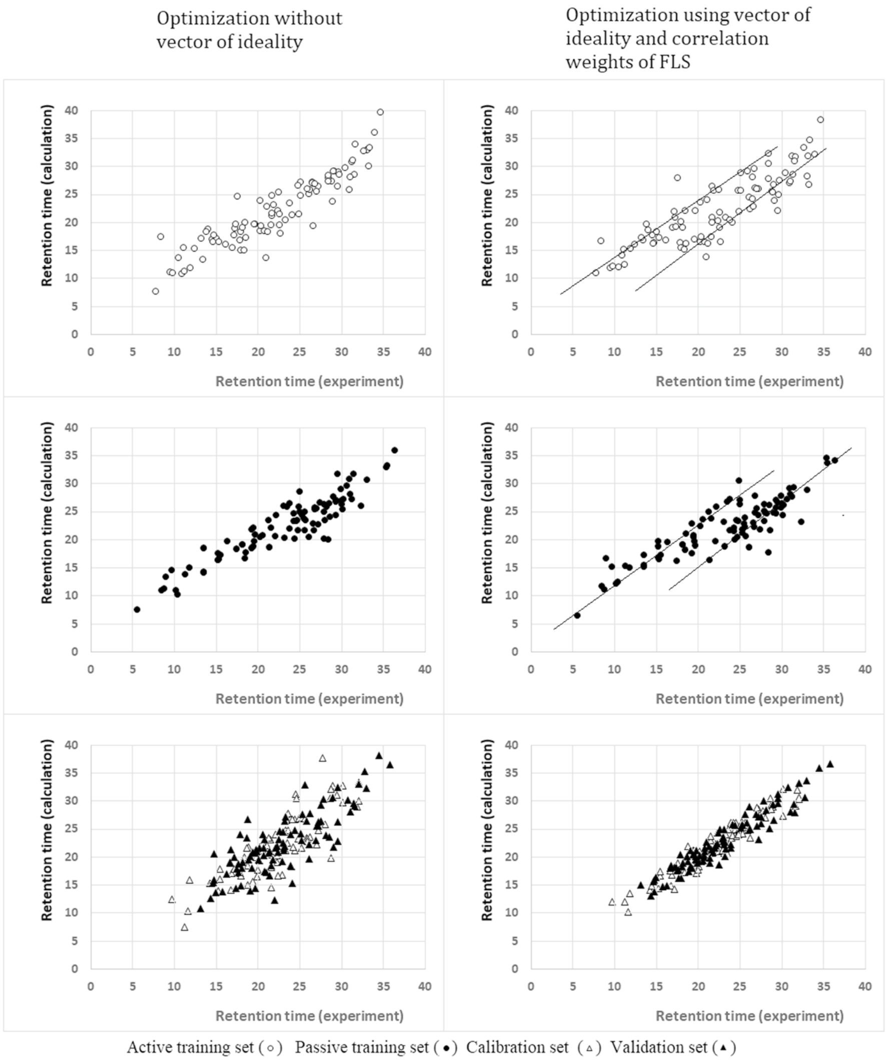


Fig. 2 The model obtained without using the idealization vector and the model obtained using the idealization vector together with the correlation weights of FLS (split 1)

of the type XYYX ([xyyx0]…); the presence of oxygen (O…); the presence of double bonds (=…); and aromaticity (c…). The patrons of the decrease in retention time are the absence of fragments of local symmetry of the XYZYX ([xyzyx0]…) type, as well as the presence of fluorine atoms (F…). Thus, these molecular features are associated with physical reality, but this relationship is statistically probabilistic. Therefore, the circumstances mentioned should be used to formulate hypotheses, but not as verified statements.

Table 7 Promoters of increase and decrease for retention time of pesticides

S_k,SS_k,$Symm$	*CWs* Probe 1	*CWs* Probe 2	*CWs* Probe 3	NA	NP	NC	Statistical Defect
1……	1.11805	0.48815	1.05446	84	78	78	0.0004
[xyyx0]……	0.43803	0.37240	0.08208	78	80	83	0.0006
C……	0.46116	0.07339	0.44016	77	80	88	0.0011
O……	0.18759	0.46434	0.00647	73	73	74	0.0004
c……	0.36265	0.45414	0.13820	71	70	68	0.0005
=……	0.60671	0.14499	0.18356	68	64	68	0.0002
c…c……	0.59407	0.08408	0.22497	67	64	56	0.0012
c…1……	0.24315	0.19000	0.51601	65	66	65	0.0005
c…(……	0.43171	0.19663	0.35003	60	63	61	0.0008
O…=……	0.17710	0.49377	0.66431	58	52	51	0.0008
O…(……	0.47573	0.17773	0.15327	56	65	63	0.0015
N…….	0.53954	0.29197	0.17282	50	43	37	0.0020
2……	0.08708	0.39251	0.43487	48	44	38	0.0015
O…C……	0.19844	0.05761	0.19022	42	49	57	0.0024
N…(……	0.56519	0.30883	0.60473	40	37	34	0.0010
[xyx2]……	0.0803	0.4002	0.3730	17	15	15	0.0008
[xyyx1]….	0.2788	0.2397	0.1816	12	5	7	0.0061
[xyzyx0]…	– 0.02885	– 0.05544	– 0.18670	70	72	68	0.0007
F…(……	– 0.32467	– 0.36936	– 0.19217	8	9	5	0.0044
[xyx5]……	– 0.1540	– 0.0869	– 0.2753	10	9	14	0.0032

NA, NP, and *NC* are the frequencies of SMILES-attribute in the active training, passive training; and calibration sets, respectively

To analyse the promoters of increase or decrease the endpoint under consideration, it is necessary that they be sufficiently common. For fragments of local symmetry, the more, or less common ones are [xyx2], [xyx5], and [xyyx1]. The selection of a collection of molecules containing such fragments showed that for [xyx2] and [xyyx1], the average retention time is 27.7 ± 2.69 and 22.19 ± 3.72, respectively. The average retention time for a collection of molecules containing [xyx5] is 20.25 ± 6.96. Thus, despite the small distribution of these fragments, their influence corresponds to their role in Table 7 ([xyx2] and [xyyx1] are promoters of an increase in retention time, [xyx5] is a promoter of a decrease in retention time. It should be noted that equivalent positions can be indicated for some fragments of local symmetry, for example, COC or NCCN. At the same time, some fragments of local symmetry make it impossible to use the term "equivalent positions", for example, C1C or C(C. In other words, fragments of local symmetry do not always have a "symmetry-like personification" in real molecules.

3.4 Comparison with models from the literature

Table 8 shows the statistical characteristics of models for the retention time of pesticides suggested in the literature. The statistical characteristics models presented here

Table 8 The comparison models for the retention time of pesticides are suggested in the literature

Training set		Validation set		Reference
N	R^2	n	R^2	
275	0.90	90	0.89	[6]
594	0.94	198	0.86	[8]
275	0.87	273	0.79	[9]
269	0.77	90	0.91	In this work (Split 1)

on the validation set are comparable with data on models from the literature [6–9]. Still, the statistical characteristics of the CORAL models on the training set are poorer. This is the literature-described effect of the correlation ideality index (CII), which is that the statistical quality for the external validation set improves, but to the detriment of the statistical characteristics of the model for the training set [15–17].

3.5 Advantages and disadvantages of the CORAL models

The advantage of the models described is their user-friendliness since their implementation requires only SMILES and numerical data for an endpoint. The disadvantage of this approach is the variation of the results depending on the chosen method and split.

All models are wrong, but some are useful [18]. Part of the knowledge can be fuzzy [19]. But this does not mean that such knowledge is useless. The models obtained based on the Monte Carlo method are probabilistic or, in other words, fuzzy. Nevertheless, they can be sources of valuable hints for practice.

The evolution of knowledge in QSPR/QSAR has two components: intensive and extensive [20]. This work mainly takes steps in the direction where "intense" results might be met.

3.6 Reproducibility

Supplementary Materials contain technical details that allow you to reproduce the described models using the CORAL software available on the Internet (http://www.insilico.eu/coral).

4 Conclusions

Using new features of the CORAL program makes it possible to simplify and simultaneously somewhat improve the predictive potential of the retention time models for substances that are potential pesticides. These possibilities are, first, the involvement in the process of modelling the correlation weights of fragments of local symmetry in SMILES, and second, the use of the vector of ideality of correlation, previously described [15–17] as the vector for the optimization for building models for the retention time of pesticides.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10910-023-01517-0.

Acknowledgements A.P.T. and A.A.T. acknowledge EFSA for the financial contribution within the project sOFT-ERA, OC/EFSA/IDATA/2022/02.

Author contributions Conceptualization, APT, AAT, IR, MR, and RC-D; methodology, APT, AAT, IR, MR, and RC-D; software, AAT.; validation, APT, AAT, IR, MR, and RC-D; formal analysis, APT; data curation, APT, AAT; writing—original draft preparation, APT, AAT; writing—review and editing, APT, AAT, IR, MR, and RC-D All authors have read and agreed to the published version of the manuscript.

Data availability The data used in this work and the models developed are freely available in the *Supplementary materials* section and at: http://www.insilico.eu/coral.

Declarations

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

References

1. S.C. Basak, A.K. Bhattacharjee, Curr. Med. Chem. **27**(1), 32–41 (2020). https://doi.org/10.2174/0929867325666181029165413
2. S.C. Basak, M.G. Vracko, Curr. Comput. Aided Drug Des. **16**(1), 1–5 (2020). https://doi.org/10.2174/157340991601200106122854
3. S.C. Basak, Curr. Comput. -Aided Drug Des. **17**(6), 703–707 (2021). https://doi.org/10.2174/1573409917666210907095711
4. M. Sakamoto, T. Tsutsumi, J. Chromatogr. A **1028**(1), 63–74 (2004). https://doi.org/10.1016/j.chroma.2003.11.066
5. F. Hernández, O.J. Pozo, J.V. Sancho, L. Bijlsma, M. Barreda, E. Pitarch, J. Chromatogr. A **1109**(2), 242–252 (2006). https://doi.org/10.1016/j.chroma.2006.01.032
6. M. Zdravković, A. Antović, J.B. Veselinović, D. Soković, A.M. Veselinović, Talanta **178**, 656–662 (2018). https://doi.org/10.1016/j.talanta.2017.09.064
7. C. Feng, Q. Xu, X. Qiu, Y. Jin, J. Ji, Y. Lin, S. Le, J. She, D. Lu, G. Wang, Chemosphere **271**, 129447 (2021). https://doi.org/10.1016/j.chemosphere.2020.129447
8. J. Parinet, Chemosphere. **275**, 130036 (2021). DOI: https://doi.org/10.1016/j.chemosphere.2021.130036
9. C. Rojas, J.F. Aranda, E.P. Jaramillo, I. Losilla, P. Tripaldi, P.R. Duchowicz, E.A. Castro, Food Chem. **342**, 128354 (2021). https://doi.org/10.1016/j.foodchem.2020.128354
10. D. Weininger, J. Chem. Inf. Comput. Sci. **28**(1), 31–36 (1988). https://doi.org/10.1021/ci00057a005
11. M.V. Putz, N.A. Dudaş, Molecules **18**(8), 9061–9116 (2013). https://doi.org/10.3390/molecules18089061
12. M.V. Putz, N.A. Dudaş, Struct. Chem. **24**(6), 1873–1893 (2013). https://doi.org/10.1007/s11224-013-0249-6
13. S. Papuga, M. Djurdjevic, A. Ciccioli, S.V. Ciprioti, Symmetry. **15**(1), 38 (2023). https://doi.org/10.3390/sym15010038
14. P.D. Cruces, A. Toscano, F.J.A. Rodríguez, R. Romo-Vázquez, P.D. Arini, Biomed. Signal. Process Control. **81**, 104493 (2023). https://doi.org/10.1016/j.bspc.2022.104493
15. A.A. Toropov, R. Carbó-Dorca, A.P. Toropova, Struct. Chem. **29**(1), 33–38 (2018). https://doi.org/10.1007/s11224-017-0997-9
16. A.A. Toropov, A.P. Toropova, Toxicol. Lett. **340**, 133–140 (2021). https://doi.org/10.1016/j.toxlet.2021.01.015
17. A.P. Toropova, A.A. Toropov, A. Roncaglioni, E. Benfenati, S.A.R.Q.S.A.R. Environ, Res. **33**(8), 621–630 (2022). https://doi.org/10.1080/1062936X.2022.2104369

18. C.L. Curchoe, J. Assist. Reprod. Genet. **37**(10), 2389–2391 (2020). https://doi.org/10.1007/s10815-020-01895-3
19. L.A. Zadeh, Inf. Control. **8**(3), 338–353 (1965). https://doi.org/10.1016/S0019-9958(65)90241-X
20. A.A. Toropov, A.P. Toropova, Molecules **25**(6), 1292 (2020). https://doi.org/10.3390/molecules25061292

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.