Journal of Molecular Structure 1182 (2019) 141–149

Contents lists available at ScienceDirect

Journal of Molecular Structure

journal homepage: http://www.elsevier.com/locate/molstruc

QSPR and nano-QSPR: What is the difference?

Alla P. Toropova*, Andrey A. Toropov

Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Via Mario Negri 2, 20156 Milan, Italy

ARTICLE INFO

Article history:
Received 23 August 2018
Received in revised form
8 January 2019
Accepted 10 January 2019
Available online 14 January 2019

Keywords:
Fullerenes C60 and C70
Solubility
Quasi-SMILES
Monte Carlo method
CORAL software
Index of ideality of correlation

ABSTRACT

The large number of quantitative structure-property/activity relationships (QSPRs/QSARs) for nanomaterials were published. Majority of these are latent traditional QSPR/QSAR in spite of labels such as "nano-QSPR" or "nano-QSAR". Traditional QSPR/QSAR are calculated with molecular descriptors. In the case of nanomaterials, the molecular descriptors are unavailable or poorly suitable for the QSPR/QSAR analysis. The CORAL software gives possibility to build up QSPR/QSAR models using simplified molecular input-line entry-system (SMILES). Recently, the quasi-SMILES were suggested as an alternative for the traditional SMILES. In this work, quasi-SMILES are used to build up united model for solubility of fullerenes C60 and C70 in organic solvents.

1. Introduction

Physicochemical [1–3] and biochemical [4] endpoints of fullerene are object of studies in the field of mathematical chemistry. Usually, theoretical (computational) prediction of endpoints related to nanomaterials can be carried out via quantitative structure–property/activity relationships (QSPRs/QSARs). However, factually, QSPR/QSAR, which are aimed to predict endpoints related to nanomaterials, often are "latent traditional QSPR/QSAR" [5–8].

Attempts to build up predictive model for solubility of fullerene C60 in organic solvents using optimal descriptors calculated with simplified molecular input-line entry systems (SMILES) [9] are described in the literature [1,2]. There is also similar attempt to build up predictive model for solubility of fullerene C70 [3].

Quasi-SMILES [10–14] are an expansion of traditional SMILES [9]. The expansion is reached via additional symbols, which reflect different conditions and circumstances [10,14].

The CORAL software (http://www.insilico.eu/coral) is a tool to build up predictive models for different endpoint using SMILES as representation of the molecular structure [15–24]. However, after the above expansion, the quasi-SMILES can be used for the CORAL software by the same manner as traditional SMILES.

The aims of this study are (i) building up and estimation of models for solubility of fullerenes C60 and C70; and (ii) estimation of ability of the Index of Ideality of Correlation [25,26] to improve predictive potential of the above models.

2. Method

2.1. Data

The numerical data on the solubility of fullerene C60 and C70 (mole fraction) in different organic solvents are taken in the literature [8]. The models have been built up for solubility transformed into the decimal logarithm (logS). The total number of quasi-SMILES, which are representing solubility of C60 and C70 in organic solvents is 212. These data were randomly distributed into the training (≈30%), invisible training (≈30%), calibration (≈20%), and validation (≈20%) sets. The above sets take part in building up the CORAL model. The essence of tasks for these sets can be described as the following.

The training set is the "builder" of the model. The Monte Carlo optimization of the correlation weights is carrying out for molecular features extracted from quasi-SMILES related to this set.

The invisible training set is the "inspector" of the model. The calculation of descriptor for quasi-SMILES of this set should confirm (or reject) suitability of the model for substances which are not involved directly to the optimization process.

* Corresponding author. Laboratory of Environmental Chemistry and Toxicology, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Via Mario Negri 2, 20156 Milano, Italy.
E-mail address: alla.toropova@marionegri.it (A.P. Toropova).

https://doi.org/10.1016/j.molstruc.2019.01.040
0022-2860/

The calibration set should detect the beginning of overfitting. Computational experiments are indicating, the optimization improves correlation between descriptor and an endpoint for training and invisible training sets, but with increase of the number of epochs of the optimization, the correlation coefficient between descriptor and endpoint for calibration set gradually decrease. The external validation set is the final estimator of the predictive potential of the model.

The traditional SMILES [9], for solvents examined here, were built up with ACD/ChemSketch software [10]. Supplementary materials contain the list of quasi-SMILES used to build up predictive models for the logS. The scheme of translation of traditional SMILES into the quasi-SMILES is described below.

2.2. Hybrid optimal descriptor

The CORAL software at the beginning was aimed to build up QSPR/QSAR models of various endpoints by paradigm

$$\text{Endpoint} = \text{F}(\text{SMILES}) \quad (1)$$

However, further development and checking up of the software have shown that hybrid descriptors [17,25,28] calculated with both SMILES attributes and graph invariants can have higher predictive potential. This can be expressed by paradigm

$$\text{Endpoint} = \text{F}(\text{SMILES},\ \text{Molecular graph}) \quad (2)$$

In the near future, an updates of the CORAL software by the possibility to use combinations of topological parameters of the molecular structure for development of predictive models of various endpoints [29] become available. The updated software is used here to build up predictive model for solubility of fullerene C60 and C70 in form

$$\text{logS} = \text{F}(\text{quasi-SMILES},\ [\text{LI}_1 \pm \text{LI}_2]) \quad (3)$$

where local invariant (LI) of the hydrogen suppressed graph (HSG) is one element of the list of (i) Morgan extended connectivity (e0, e1, e2, e3); (ii) Paths of length 2,3,4 (p2, p3, p4); (iii) Valence shells of second and third orders (s2, s3); and (iv) nearest neighbour code (nn).

Quasi-SMILES for the case of a system "solvent − C60" are represented by SMILES of solvent plus symbol "x", e.g. C1CCC=CC1x, C1CCC2CCCCC2C1x, etc., for the case of a system "solvent − C70" the representation is SMILES of solvent plus symbol "y", e.g. CCCCCy, C1CCCCC1y, etc.

Hybrid descriptor used here is defined as the following:

$$^{Hybrid}DCW(T^*,N^*) = {}^{SMILES}DCW(T^*,N^*) + {}^{GRAPH}DCW(T^*,N^*) \quad (4)$$

$$^{SMILES}DCW(T^*,N^*) = CW(HARD) + CW(Cmax) + \sum CW(S_k) + \sum CW(SS_k) + \sum CW(SSS_k) \quad (5)$$

Table 1
Example of representation for the S_k, SS_k, and SSS_k.in the case SMILES = Cc1ccc(O)c(C)n1.

ID	Comment	1	2	3	4	5	6	7	8	9	10	11	12
1	Representation of S_k	C	.	.	.	.	.	.	.	.	.	.	.
		c	.	.	.	.	.	.	.	.	.	.	.
		1	.	.	.	.	.	.	.	.	.	.	.
		c	.	.	.	.	.	.	.	.	.	.	.
		c	.	.	.	.	.	.	.	.	.	.	.
		c	.	.	.	.	.	.	.	.	.	.	.
		(	.	.	.	.	.	.	.	.	.	.	.
		O	.	.	.	.	.	.	.	.	.	.	.
		(	.	.	.	.	.	.	.	.	.	.	.
		c	.	.	.	.	.	.	.	.	.	.	.
		(	.	.	.	.	.	.	.	.	.	.	.
		C	.	.	.	.	.	.	.	.	.	.	.
		(	.	.	.	.	.	.	.	.	.	.	.
		n	.	.	.	.	.	.	.	.	.	.	.
		1	.	.	.	.	.	.	.	.	.	.	.
2	Representation of SS_k	C	.	.	.	C	.	.	.	.	.	.	.
		C	.	.	.	1	.	.	.	.	.	.	.
		C	.	.	.	1	.	.	.	.	.	.	.
		C	.	.	.	c	.	.	.	.	.	.	.
		C	.	.	.	c	.	.	.	.	.	.	.
		C	.	.	.	(	.	.	.	.	.	.	.
		O	.	.	.	(	.	.	.	.	.	.	.
		O	.	.	.	(	.	.	.	.	.	.	.
		c	.	.	.	(	.	.	.	.	.	.	.
		c	.	.	.	(	.	.	.	.	.	.	.
		C	.	.	.	(	.	.	.	.	.	.	.
		C	.	.	.	(	.	.	.	.	.	.	.
		n	.	.	.	(	.	.	.	.	.	.	.
		n	.	.	.	1	.	.	.	.	.	.	.
3	Representation of SSS_k	C	.	.	.	c	.	.	.	1	.	.	.
		c	.	.	.	1	.	.	.	c	.	.	.
		c	.	.	.	c	.	.	.	1	.	.	.
		c	.	.	.	c	.	.	.	c	.	.	.
		c	.	.	.	c	.	.	.	(	.	.	.
		c	.	.	.	(	.	.	.	O	.	.	.
		(	.	.	.	O	.	.	.	(	.	.	.
		c	.	.	.	(	.	.	.	O	.	.	.
		(	.	.	.	c	.	.	.	(	.	.	.
		c	.	.	.	(	.	.	.	C	.	.	.
		(	.	.	.	C	.	.	.	(	.	.	.
		n	.	.	.	(	.	.	.	C	.	.	.
		1	.	.	.	n	.	.	.	(	.	.	.

Molecular features extracted from SMILES or from graph are represented by sequences of twelve symbols. Table 1 contains an example of representation for the S_k, SS_k, and SSS_k. Fig. 1 contains the elucidation for HARD, Cmax, C5 and C6. Table 2 contains examples of Cmax, C5 and C6 for different molecules.

Fig. 2 contains example of the combination of graph invariant for 3,3-dimethylhexane.

Fig. 3 contains fragment of the user interface of the CORAL software. User can select list of molecular features to define optimal descriptor via interface represented at Fig. 3. Fig. 4 contains the definition of the optimal descriptor that is calculated with Eq. (4).

The general scheme of building up model for the solubility of

$$^{GRAPH}DCW(T^*,N^*) = CW(C5) + CW(C6) + \sum CW(e1_k) + \sum CW(e1_k + e2_k) + \sum CW(|e2_k - e1_k|) + \sum CW(e2_k) + \sum CW(p2_k) + \sum CW(p2_k + p3_k) + \sum CW(|p3_k - p2_k|) + \sum CW(p3_k) + \sum CW(s2_k) + \sum CW(s2_k + s3_k) + \sum CW(|s3_k - s2_k|) + \sum CW(s3_k) \quad (6)$$

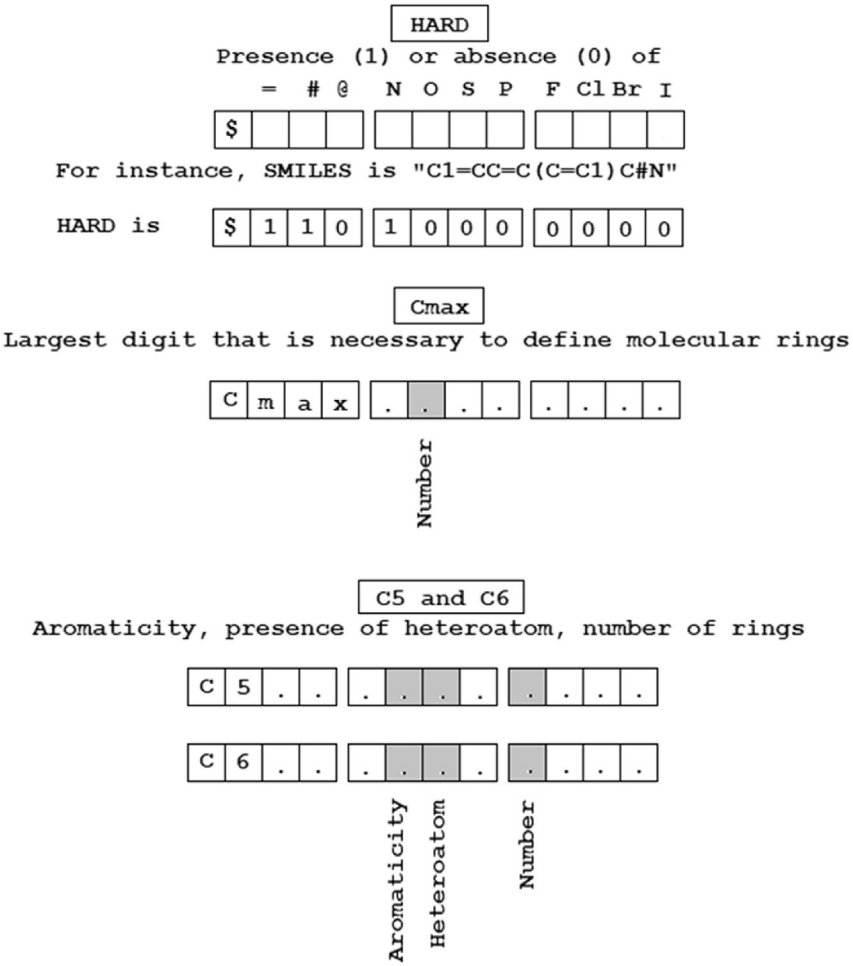


Fig. 1. The representation of molecular features extracted from SMILES and HSG. The HARD is sequence of twelve symbols which encode molecular status: presence/absence of double (=), triple (#), and stereo chemical (@) covalent bonds; presence/absence of nitrogen (N), oxygen (O), sulphur (S), phosphorus (P), fluorine (F), chlorine (Cl), bromine (Br), and iodine (I). Cmax is indicator of number of rings in a molecule (from 0 to 9); C5 and C6 are sequences of twelve symbols which encode specificity of molecular five member and six member rings: "A" indicates presence of aromaticity, absence is indicated by dot "."; "H" indicates presence of heteroatom, absence is indicated by dot ".".

fullerenes C60 and C70 in organic compounds is the following:

1. Definition of the total list of quasi-SMILES available to build up the model;
2. Distribution of all available data into the training, invisible training, calibration, and validation sets;
3. Building up predictive model for solubility of fullerenes C60 and C70 in organic compounds (logS) for the training, invisible training, and calibration sets using the Monte Carlo method;
4. Estimation of the predictive potential of the model with the statistical quality of the model for external validation set. Fig. 5 contains the graphical representation of steps 1–4.

The Monte Carlo method is optimization of the correlation weights (*CWs*) involved in Eq. (5) and Eq. (6) with a target function. Two versions of target function TF_1 and TF_2 are examined here

$$TF_1 = r_{TRN} + r_{iTRN} - |r_{TRN} - r_{iTRN}|*0.1 \quad (7)$$

$$TF_2 = TF_1 + IIC_{CLB}*0.1 \quad (8)$$

The r_{TRN} and r_{iTRN} are correlation coefficient between observed and predicted endpoint for the training and invisible training sets, respectively.

The Index of Ideality of Correlation IIC_{CLB} [26,27] is calculated with data on the calibration (*CLB*) set as the following:

$$IIC_{CLB} = r_{CLB}\frac{\min(^{-}MAE_{CLB}, ^{+}MAE_{CLB})}{max(^{-}MAE_{CLB}, ^{+}MAE_{CLB})} \quad (9)$$

$$^{-}MAE_{CLB} = \frac{1}{^{-}N}\sum_{k=1}^{^{-}N}|\Delta_k|, \quad \Delta_k < 0; \; ^{-}N \text{ is the number of } \Delta_k < 0 \quad (10)$$

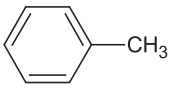
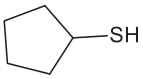
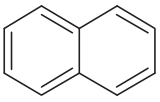
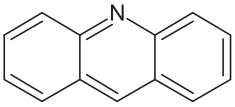
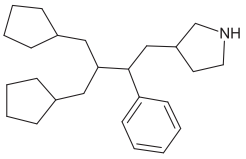
$$^{+}MAE_{CLB} = \frac{1}{^{+}N}\sum_{k=1}^{^{+}N}|\Delta_k|, \quad \Delta_k \geq 0; \; ^{+}N \text{ is the number of } \Delta_k \geq 0 \quad (11)$$

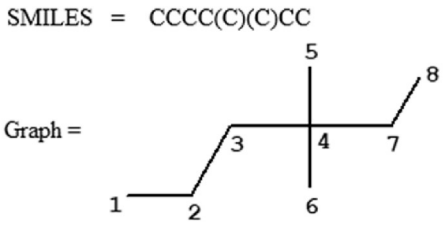
$$\Delta_k = observed_k - calculated_k \quad (12)$$

The observed and calculated are corresponding values of the endpoint.

The model for solubility C60 and C70 fullerenes in organic solvents is the following:

Table 2
Clarification of molecular features, which are related to quality and quantity of rings.

Structure	SMILES	C5	C6	Cmax
	Cc1ccccc1	C5 ……0 …	C6…A..1 …	Cmax.1 ……
	SC1CCCC1	C5 ……1…	C6 ……0…	Cmax.1 ……
	c1ccccn1	C5 ……0 …	C6 … AH.1 …	Cmax.1……
	c1cccc2ccccc12	C5 ……0 …	C6 … A..2 …	Cmax.2 ……
	c1c3ccccc3nc2ccccc12	C5 ……0 …	C6 … AH.3 …	Cmax.3 ……
	C(C(CC1CCCC1)C(CC2CCNC2)c3ccccc3)C4CCCC4	C5 ….H.3 …	C6 … A..1 …	Cmax.4 ……



Adjacency Matrix

	1	2	3	4	5	6	7	8
1	0	1	0	0	0	0	0	0
2	1	0	1	0	0	0	0	0
3	0	1	0	1	0	0	0	0
4	0	0	1	0	1	1	1	0
5	0	0	0	1	0	0	0	0
6	0	0	0	1	0	0	0	0
7	0	0	0	1	0	0	0	1
8	0	0	0	0	0	0	1	0

Graph invariants

e0	e1	e2	e3	p2	p3	p4	s2	s3	nn
1	2	3	8	1	1	3	2	4	110
2	3	8	12	1	3	1	4	4	220
2	6	9	27	4	1	0	5	1	220
4	6	19	29	2	1	0	3	1	440
1	4	6	19	3	2	1	5	3	110
1	4	6	19	3	2	1	5	3	110
2	5	8	24	3	1	1	4	2	220
1	2	5	8	1	3	1	4	4	110

$e1_k$ = 2,3,6,6,4,4,5,2
$e2_k$ = 3,8,9,19,6,6,8,5
$e1_k + e2_k$ = 5,11,15,25,10,10,13,7
$e2_k - e1_k$ = 1,5,3,13,2,2,3,3

$p2_k$ = 1, 1, 4, 2, 3,3,3,1
$p3_k$ = 1, 3, 1, 1, 2, 2, 1, 3
$p2_k + p3_k$ = 2,4,5,3,5,5,4,4
$p3_k - p2_k$ = 0,2,3,1,1,1,2,2

$s2_k$ = 2,4,5,3,5,5,4,4
$s3_k$ = 4,4,1,1,3,3,2,4
$s3_k + s2_k$ = 6,8,6,4,8,8,6,8
$s3_k - s2_k$ = 2,0,4,2,2,2,2,0

Fig. 2. Example of calculation of graph invariants and their combinations for the optimal descriptor in the case of 3,3-dimethylhexane. It should be noted combinations such as e2±p3, p3±s2, s3±nn, etc. are also possible.

$$logS = C_0 + C_1{}^{Hybrid} DCW(T^*, N^*) \quad (13)$$

Having the results of several runs of the optimization one can extract statistically significant molecular features with positive value of the correlation weight for all runs, which are promoters of solubility increase, and vice versa, the features with negative value of the correlation weight for all runs, which are promoters of

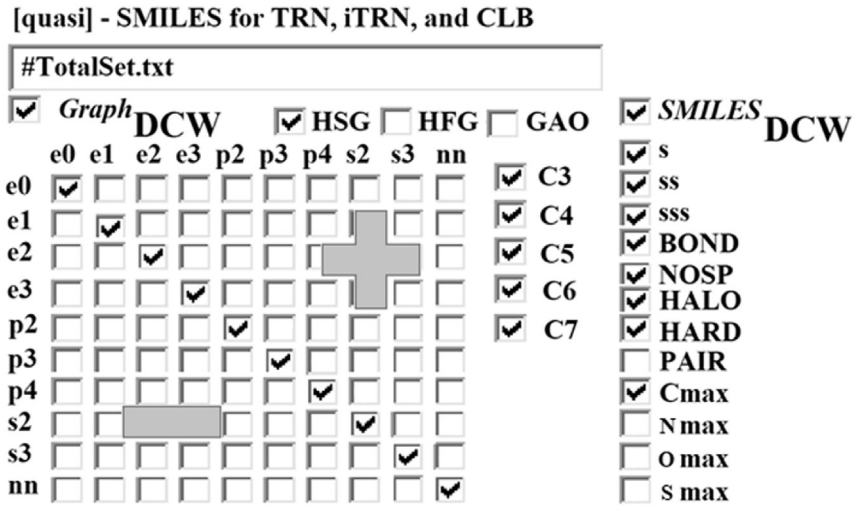


Fig. 3. Interface for definition of the optimal descriptor: the diagonal is place for traditional graph invariants: extended connectivity of zero (e0), first (e1), second (e2), third (e3) orders; paths of length two (p2), three (p3), four (p4), valence shells of second (s2) and third orders; and nearest neighbour code (nn). The selection of pair e1 and e2 (upper triangle) means absolute value of the arithmetic operation "e1 plus e2". The selection of pair e2 and e1 (low triangle) means absolute value for arithmetic operation "e1 minus e2".

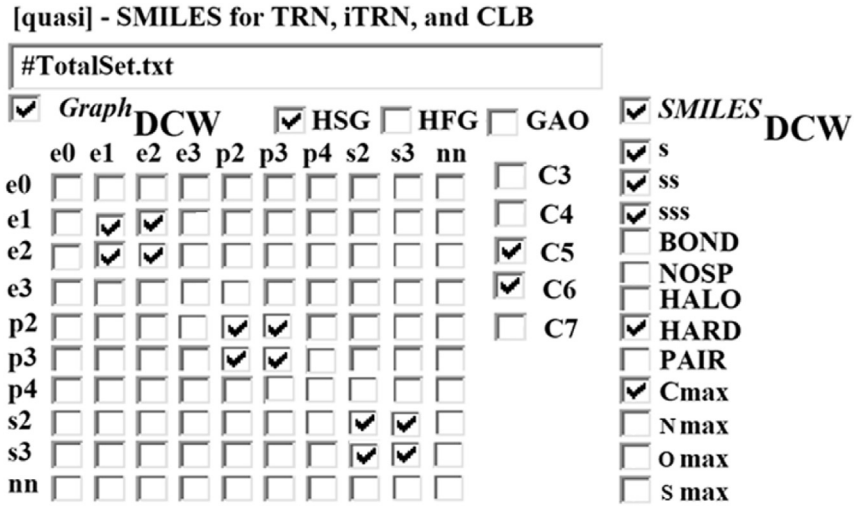


Fig. 4. The scheme of definition for descriptor calculated with Eq. (4).

decrease for solubility. Table 3 contains examples of promoters for increase of solubility fullerenes C60 and C70, which have been detected in this study.

Estimation of the quality of distribution into the above-mentioned sets via (i) defects of molecular features extracted from quasi-SMILES and from HSG; (ii) defects of quasi-SMILES; and (iii) defect of distribution. The definition of statistical defect for each molecular feature (F_k) which is involved (non blocked) to build up the model is the following:

$$d_k = \frac{|P(F_k) - P'(F_k)|}{N(F) + N(F_k)} \quad (14)$$

where $P(F_k)$ and $P'(F_k)$ are probability of F_k in the training and calibration sets, respectively;

$N(A_k)$ and $N'(A_k)$ are frequencies of F_k in the training and calibration sets, respectively.

The calculation for all substances the statistical defect of quasi-SMILES and HSG-defect (D_j) is the following

$$D_j = \sum_{k=1}^{NF} d_k \quad (15)$$

where NF is the number of non-blocked molecular features extracted from SMILES or HSG

A j-th substance falls in the domain of applicability if

$$Dj < 2*\overline{D} \quad (16)$$

where $\overline{D}$ is average of the statistical defect of quasi-SMILES and HSG for the training set.

3. Results and discussion

The CORAL models for solubility of fullerenes C60 and C70 in organic solvents are the following:

Split 1, TF_1

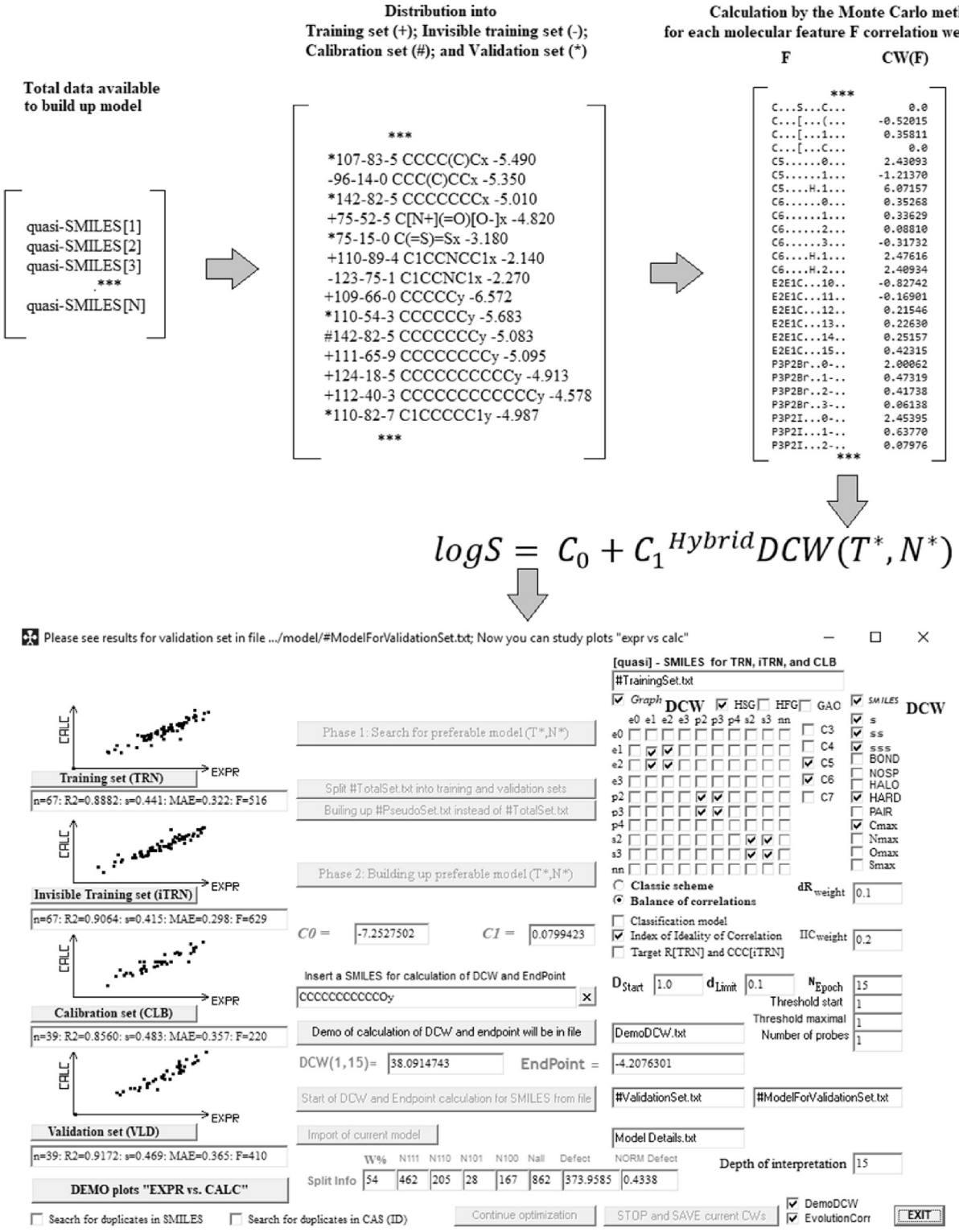


Fig. 5. Graphical representation of the general scheme of building up a CORAL model.

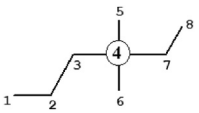
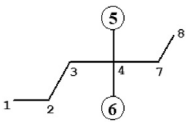
$$\text{Endpoint} = -7.7530513(\pm 0.0112580) + 0.1127878(\pm 0.0003266)*\text{DCW}(1,1) \quad (17)$$

Split 1, TF_2

$$\text{Endpoint} = -7.2527502(\pm 0.0220563) + 0.0799423(\pm 0.0004969)*\text{DCW}(1,15) \quad (18)$$

Split 2, TF_1

Table 3
A collection of molecular features, which are promoters of increase for solubility of C60 and C70 fullerenes.

Split	Molecular features, F_k	CW(F_k) Probe 1	CW(F_k) Probe 2	CW(F_k) Probe 3	N1[a]	N2	N3	Examples
1	x …. …. …	0.29092	0.20955	0.15487	58	57	35	
2	x …. …. …	0.05410	0.08207	0.67886	63	63	27	
3	x …. …. …	0.28179	0.43144	0.41658	61	56	35	
1	p2p3C … 3+..	0.44287	0.04536	0.04385	36	29	23	
2	p2p3C … 3+..	0.03826	0.30465	0.16629	37	34	15	
3	p2p3C … 3+..	0.12807	0.38603	0.32834	36	29	20	
1	EC2-C … 6 …	0.02235	0.20091	0.48699	30	31	18	
2	EC2-C … 6 …	0.19209	0.00721	0.46186	33	33	13	
3	EC2-C … 6 …	0.22087	0.17727	0.08613	33	33	13	

[a] N1 = number of quasi-SMILES in the training set; N2 = number of quasi-SMILES in the invisible training set; N3 = number of quasi-SMILES in the calibration set.

$$\text{Endpoint} = -7.5092233(\pm 0.0093407) + 0.1079987(\pm 0.0002635)*\text{DCW}(1,2) \quad (19)$$

Split 2, TF_2

$$\text{Endpoint} = -7.2588610(\pm 0.0207275) + 0.0943503(\pm 0.0005490)*\text{DCW}(1,15) \quad (20)$$

Split 3, TF_1

$$\text{Endpoint} = -7.2023643(\pm 0.0082713) + 0.1576026(\pm 0.0003704)*\text{DCW}(1,1) \quad (21)$$

Split 3, TF_2

$$\text{Endpoint} = -7.2321532(\pm 0.0241920) + 0.0771553(\pm 0.0005233)*\text{DCW}(1,15) \quad (22)$$

Table 4 contains the statistical characteristics of these models. The Monte Carlo optimization with target function TF_2 gives models with better predictive potential than the optimization with TF_1. In other words, the Index of Ideality of Correlation gives possibility to improve predictive potential of described models.

At present, there is trend to establishing predictive models for physicochemical properties of mixtures [30–32]. Factually, the number of tasks related to prediction of endpoints related to mixtures is larger than the number of tasks related to pure substances. Described quasi-SMILES is an approach able to be used to solve the above task.

In addition, model for solubility of fullerene C60 based on

Table 4
The statistical characteristics of the CORAL models for solubility of C60 and C70 fullerenes in organic solvents.

Split[a]	TF	Set	n	r^2	CCC	*IIC*	q^2	$Q2_{F1}$	$Q2_{F2}$	$Q2_{F3}$	s	MAE	F
1	*1*	TRN	67	0.9583							0.270	0.191	1494
		iTRN	67	0.9575							0.317	0.244	1464
		CLB	39	0.8050	0.8802	0.5485	0.7831	0.7370	0.7296	0.7542	0.659	0.550	153
		VLD	39	0.8479							0.561	0.415	
	2	TRN	67	0.8882							0.441	0.322	516
		iTRN	67	0.9064							0.415	0.298	629
		CLB	39	0.8560	**0.9189**	**0.9252**	**0.8438**	**0.8586**	**0.8547**	**0.8679**	0.483	0.357	220
		VLD	39	**0.9172**							**0.470**	**0.365**	
2	*1*	TRN	72	0.9692							0.234	0.158	2200
		iTRN	71	0.9713							0.332	0.279	2332
		CLB	34	0.8526	0.9185	0.7622	0.8272	0.8409	0.8407	0.8765	0.497	0.364	185
		VLD	35	0.6760							0.649	0.402	
	2	TRN	72	0.8674							0.485	0.404	458
		iTRN	71	0.8673							0.545	0.447	451
		CLB	34	**0.9336**	**0.9641**	**0.9662**	**0.9266**	**0.9324**	**0.9323**	**0.9475**	0.324	0.253	450
		VLD	35	**0.8991**							**0.351**	**0.271**	
3	*1*	TRN	67	0.9675							0.248	0.171	1935
		iTRN	67	0.9671							0.266	0.185	1910
		CLB	39	0.6052	0.7710	0.6142	0.5646	0.5612	0.5161	0.6950	0.782	0.591	57
		VLD	39	0.6548							0.694	0.532	
	2	TRN	67	0.8645							0.506	0.380	415
		iTRN	67	0.8880							0.490	0.382	515
		CLB	39	**0.8619**	**0.9273**	**0.9284**	**0.8446**	**0.8718**	**0.8587**	**0.9109**	0.423	0.329	231
		VLD	39	**0.9099**							**0.357**	**0.273**	

[a] Split = number of split; TF = target function; Set: TRN = training; iTRN = invisible training; CLB = calibration; VLD = validation sets; n = number of solvents in set; r^2 = correlation coefficient; CCC = concordance correlation coefficient; *IIC* = index of ideality of correlation; q^2 = cross-validated correlation coefficient; the best statistical characteristics indicated by bold.

Table 5
Comparison of predictive models for solubility of C60 fullerene in organic compounds.

Method	Set	n	r^2	References
MLR[a]	Training set	92	0.861	[33]
	Validation set	30	0.903	
PLS	Training set	80	0.674	[34]
	Validation set	28	0.692	
SVM	Training set	92	0.871	[35]
	Validation set	30	0.940	
DTB	Training set	145	0.970	[8]
	Validation set	36	0.964	
Monte Carlo	Training set	55	0.947	This work
	Invisible training set	55	0.939	
	Calibration set	36	0.918	
	Validation set	35	0.915	

[a] MLR multiple linear regression; PLS partial least square regression; SVM support vector machine; DTB decision tree boost.

traditional SMILES is examined here (Table S7 and Table S8 in Supplementary materials). Table 5 contains the comparison of models for solubility of fullerene C60 in organic solvent suggested in the literature. One can see (Table 5) the statistical quality of the CORAL model is comparable with models from the literature.

Thus, the CORAL models examined here are associated with the following information:

1. A defined endpoint is solubility of fullerenes C60 and C70 in organic solvent.
2. An unambiguous algorithm is the Monte Carlo optimization using the CORAL software.
3. A defined domain of applicability is defined via inequality 16.
4. Appropriate measures of goodness-of-fit, robustness and predictivity are estimated by the correlation coefficient and root mean squared error for the external validation set.
5. A mechanistic interpretation is available via described stable promoters of increase for the solubility (Table 3).

The application of traditional QSPR/QSAR in order to build up models, which are more or less related to phenomena of physico-chemical or biochemical impact of various nanomaterials is described in the literature [5–8]. However, are there works where the QSPR/QSAR analysis of nanomaterials contains ideas and techniques not used for traditional QSPR/QSAR?

The difference between traditional QSPR and nano-QSPR is the following: the latter should be sensitive to presence of nanomaterials. Factually, quasi-SMILES are a tool conceptually inequivalent to traditional approaches of the QSPR/QSAR analyses, because quasi-SMILES are able to be sensitive to presence of nanomaterials and to factors, which influence nanomaterials.

4. Conclusions

The described approach based on quasi-SMILES gives good model of solubility for fullerenes C60 and C70 in organic solvents. The index of ideality of correlation improves the predictive potential of the CORAL models. The CORAL models are comparable with similar models for solubility of fullerenes suggested in the literature (Table 5). The described approach build up models in accordance with the OECD principles [28]. The proposed approach can be useful to build up predictive models of the physico-chemical or biochemical behavior of multicomponent systems subjected to various impacts. For example, it can be systems of nanomaterials, peptides, and multicomponent mixtures.

Acknowledgement

Authors thank the project LIFE-CONCERT contract (LIFE17 GIE/IT/000461) for financial support.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.molstruc.2019.01.040.

Disclosure

The authors confirm that this article content has no conflict of interest.

References

[1] A.A. Toropov, B.F. Rasulev, D. Leszczynska, J. Leszczynski, Additive SMILES based optimal descriptors: QSPR modeling of fullerene C60 solubility in organic solvents, Chem. Phys. Lett. 444 (1–3) (2007) 209–214.
[2] A.A. Toropov, B.F. Rasulev, D. Leszczynska, J. Leszczynski, Multiplicative SMILES-based optimal descriptors: QSPR modeling of fullerene C60 solubility in organic solvents, Chem. Phys. Lett. 457 (4–6) (2008) 332–336.
[3] A.P. Toropova, A.A. Toropov, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, CORAL: QSPR models for solubility of [C60] and [C70] fullerene derivatives, Mol. Divers. 15 (1) (2011) 249–256.
[4] A.A. Toropov, A.P. Toropova, Quasi-SMILES and nano-QFAR: united model for mutagenicity of fullerene and MWCNT under different conditions, Chemosphere 139 (2015) 18–22.
[5] D. Fourches, D. Pu, C. Tassa, R. Weissleder, S.Y. Shaw, R.J. Mumper, A. Tropsha, Quantitative nanostructure - activity relationship modelling, ACS Nano 4 (10) (2010) 5703–5712.
[6] A.A. Toropov, A.P. Toropova, T. Puzyn, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, QSAR as a random event: modeling of nanoparticles uptake in PaCa2 cancer cells, Chemosphere 92 (2013) 31–37.
[7] M. González-Durruthy, L.C. Alberici, C. Curti, Z. Naal, D.T. Atique-Sawazaki, J.M. Vázquez-Naya, H. González-Díaz, C.R. Munteanu, Experimental-computational study of carbon nanotube effects on mitochondrial respiration: in silico nano-QSPR machine learning models based on new Raman spectra transform with markov-shannon entropy invariants, J. Chem. Inf. Model. 57 (5) (2017) 1029–1044.
[8] S. Gupta, N. Basant, Predictive modeling: solubility of C60 and C70 fullerenes in diverse solvents, Chemosphere 201 (2018) 361–369.
[9] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci. 28 (1988) 31–36.
[10] ACD/ChemSketch Freeware Version 10.00, Adv. Chem. Dev. Inc. Toronto, ON, Canada, 2006. www.acdlabs.com.
[11] A.A. Toropov, A.P. Toropova, Quasi-QSAR for mutagenic potential of multi-walled carbon-nanotubes, Chemosphere 124 (1) (2015) 40–46.
[12] A.P. Toropova, A.A. Toropov, R. Rallo, D. Leszczynska, J. Leszczynski, Optimal descriptor as a translator of eclectic data into prediction of cytotoxicity for metal oxide nanoparticles under different conditions, Ecotoxicol. Environ. Saf. 112 (2015) 39–45.
[13] A.A. Toropov, P.G.R. Achary, A.P. Toropova, Quasi-SMILES and nano-QFPR: the predictive model for zeta potentials of metal oxide nanoparticles, Chem. Phys. Lett. 660 (2016) 107–110.
[14] A.P. Toropova, A.A. Toropov, Nano-QSAR in cell biology: model of cell viability as a mathematical function of available eclectic data, J. Theor. Biol. 416 (2017) 113–118.
[15] T.X. Trinh, J.-S. Choi, H. Jeon, H.-G. Byun, T.-H. Yoon, J. Kim, Quasi-smiles-based nano-quantitative structure-activity relationship model to predict the cytotoxicity of multiwalled carbon nanotubes to human lung cells, Chem. Res. Toxicol. 31 (3) (2018) 183–190.
[16] P.G.R. Achary, QSPR modelling of dielectric constants of π-conjugated organic compounds by means of the CORAL software, SAR QSAR Environ. Res. 25 (6) (2014) 507–526.
[17] E. Amata, A. Marrazzo, M. Dichiara, M.N. Modica, L. Salerno, O. Prezzavento, G. Nastasi, A. Rescifina, G. Romeo, V. Pittalà, Comprehensive data on a 2D-QSAR model for Heme Oxygenase isoform 1 inhibitors, Data Brief 15 (2017) 281–299.
[18] A. Ghaedi, Predicting the cytotoxicity of ionic liquids using QSAR model based on SMILES optimal descriptors, J. Mol. Liq. 208 (2015) 269–279.
[19] A. Worachartcheewan, C. Nantasenamat, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, QSAR study of H1N1 neuraminidase inhibitors from influenza a virus, Lett. Drug Des. Discov. 11 (4) (2014) 420–427.
[20] J.B. Veselinović, V. Đorđević, M. Bogdanović, I. Morić, A.M. Veselinović, QSAR modeling of dihydrofolate reductase inhibitors as a therapeutic target for multiresistant bacteria, Struct. Chem. 29 (2) (2018) 541–551.
[21] Q. Li, X. Ding, H. Si, H. Gao, QSAR model based on SMILES of inhibitory rate of

2, 3-diarylpropenoic acids on AKR1C3, Chemometr. Intell. Lab. Syst. 139 (2014) 132–138.
[22] M.H. Fatemi, H. Malekzadeh, CORAL: predictions of retention indices of volatiles in cooking rice using representation of the molecular structure obtained by combination of SMILES and graph approaches, J. Iran. Chem. Soc. 12 (3) (2015) 405–412.
[23] M.A. Islam, T.S. Pillay, Simplified molecular input line entry system-based descriptors in QSAR modeling for HIV-protease inhibitors, Chemometr. Intell. Lab. Syst. 153 (2016) 67–74.
[24] A. Kumar, S. Chauhan, QSAR differential model for prediction of SIRT1 modulation using Monte Carlo method, Drug Res. 67 (3) (2017) 156–162.
[25] A. Rescifina, G. Floresta, A. Marrazzo, C. Parenti, O. Prezzavento, G. Nastasi, M. Dichiara, E. Amata, Sigma-2 receptor ligands QSAR model dataset, Data Brief 13 (2017) 514–535.
[26] A.A. Toropov, A.P. Toropova, The index of ideality of correlation: a criterion of predictive potential of QSPR/QSAR models? Mutat. Res. Genet. Toxicol. Environ. Mutagen 819 (2017) 31–37.
[27] A.A. Toropov, R. Carbó-Dorca, A.P. Toropova, Index of Ideality of Correlation: new possibilities to validate QSAR: a case study, Struct. Chem. 29 (1) (2018) 33–38.
[28] A.P. Toropova, A.A. Toropov, Hybrid optimal descriptors as a tool to predict skin sensitization in accordance to OECD principles, Toxicol. Lett. 275 (2017) 57–66.
[29] A.A. Toropov, E. Benfenati, QSAR models for Daphnia toxicity of pesticides based on combinations of topological parameters of molecular structures, Bioorg. Med. Chem. 14 (8) (2006) 2779–2788.
[30] A.P. Toropova, A.A. Toropov, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, CORAL: models of toxicity of binary mixtures, Chemometr. Intell. Lab. Syst. 119 (2012) 39–43.
[31] F. Luan, X. Xu, H. Liu, M.N.D.S. Cordeiro, Prediction of the baseline toxicity of non-polar narcotic chemical mixtures by QSAR approach, Chemosphere 90 (6) (2013) 1980–1986.
[32] L. Jiao, X. Zhang, Y. Qin, X. Wang, H. Li, QSPR study on the flash point of organic binary mixtures by using electrotopological state index, Chemometr. Intell. Lab. Syst. 156 (2015) 211–216.
[33] T. Petrova, B.F. Rasulev, A.A. Toropov, D. Leszczynska, J. Leszczynski, Improved model for fullerene C60 solubility in organic solvents based on quantum-chemical and topological descriptors, J. Nanopart. Res. 13 (2011) 3235–3247.
[34] J.B. Ghasemi, M. Salahinejad, M.K. Rofouei, Alignment independent 3DQSAR modeling of fullerene (C60) solubility in different organic solvents, Fullerenes, Nanotub. Carbon Nanostruct. 21 (2013) 367–380.
[35] W.-D. Cheng, C.-Z. Cai, Accurate model to predict the solubility of fullerene C60 in organic solvents by using support vector regression, Fullerenes, Nanotub. Carbon Nanostruct. 25 (2017) 58–64.