Chapter 8
Quasi-SMILES-Based QSPR/QSAR Modeling

Shahin Ahmadi and Neda Azimi

Abstract Quantitative structure–property/activity relationships (QSPRs/QSARs) have been used to predict the physicochemical property and biological activity of different substances, considering that the physicochemical property/biological activity of a new or untested substance can be inferred from the molecular structure or other properties of similar compounds whose properties/activities have already been assessed. Traditional QSPR/QSAR models based on physicochemical properties and molecular information are not so successful in predicting endpoint of substances such as nanomaterials due to scarcity of available dataset in same conditions. A new approach using eclectic information as descriptors to predict the endpoint of substance materials was developed in CORAL software (http://www.insilico.eu/coral). In this approach, physicochemical properties and the experimental conditions of substance are represented by so-called quasi-SMILES, which are character-based representations derived from traditional Simplified Molecular Input Line Entry System (SMILES). Thus, a main advantage of the quasi-SMILES is to increase the number of available datasets by using the eclectic data in developing quasi-SMILES-based QSPRs/QSARs models. This chapter provides instructions on how to use CORAL software for building QSPR/QSAR models based on quasi-SMILES.

Keywords QSPR · QSAR · Eclectic information · Quasi-SMILES · CORAL software

S. Ahmadi (✉)
Department of Chemistry, Faculty of Pharmaceutical Chemistry, Tehran Medical Sciences, Islamic Azad University, Tehran, Iran
e-mail: s.ahmadi@iautmu.ac.ir; ahmadi.chemometrics@gmail.com

N. Azimi
Advanced Chemical Engineering Research Center, Razi University, Kermanshah, Iran

A. P. Toropova and A. A. Toropov (eds.), *QSPR/QSAR Analysis Using SMILES and Quasi-SMILES*, Challenges and Advances in Computational Chemistry and Physics 33, https://doi.org/10.1007/978-3-031-28401-4_8

Abbreviations

AD	Applicability Domain
CCC	Concordance Correlation Coefficient
CORAL	CORrelation And Logic
CII	Correlation Intensity Index
EP	Endpoint
F	Fischer ratio
IIC	Index of Ideality Correlation
MAE	Mean Absolute Error
NPs	Nanoparticles
OECD	Organization of Economic Co-operation and Development
QSAR	Quantitative Structure–Activity Relationship
QSPR	Quantitative Structure–Property Relationship
RMSE	Root-Mean-Square Error
SMILES	Simplified Molecular Input Line Entry System
TF	Target Function

8.1 Introduction

Quantitative structure–activity/property relationship (QSAR/QSPR) approach is indubitably of considerable importance in food chemistry [1, 2], environmental chemistry [3], modern chemistry [4–6], biochemistry [7], nanotechnology [8, 9], and drug design [10, 11]. The QSAR/QSPR approach is the mathematical and computerized search for compounds with desired activities/properties using chemical intuition and experience. Once a structure–activity/property correlation has been established, any number of compounds, including those not yet synthesized, can be easily screened on a computer to select structures with the desired activity/properties. Then the most promising compounds can be found for synthesis and experimental testing [12]. Therefore, QSAR/QSPR study saves cost and time for the development process of new molecules as drugs, materials, additives, or any other purpose. While finding successful structure–activity models is not an easy task, the recent increase in the number of papers in QSPR/QSAR research clearly indicates the rapid evolution in this area. To obtain a significant correlation, it is very important to use appropriate descriptors, whether they are theoretical, empirical, or derived from easily empirical properties of the constructs [12]. A group of descriptors shows simple molecular properties and therefore can give insight into the physicochemical nature of the activity/property under consideration.

Considering the growth of nanotechnology, modeling the properties or toxicity of nanoparticles (NPs) on living organisms is very important [13–15]. Although it is difficult to conduct toxicological experiments or obtain physical properties of NPs on a case-by-case basis, QSPR/QSAR is a computationally efficient technique because

it saves time, cost, and animal sacrifice. The first part of nano-QSPR/QSAR model implementation includes data collection (including descriptors and endpoints) and data processing. The dataset can be obtained from the literature, databases, experiments, or integrated multiple sources. Therefore, to construct nano-QSPR/QSAR models, it is important to identify a new set of descriptors that can accurately represent the properties of NPs as well as the experimental conditions.

During recent years, the Simplified Molecular Input Line Entry System (SMILES) and quasi-SMILES descriptors have been examined by some researchers for QSPR/QSAR modeling [16–19]. The SMILES can reveal molecular structures, and quasi-SMILES can represent molecular structure and physicochemical properties and exposure conditions [8, 20, 21]. SMILES of a molecule is based on a set of rules that allow a molecular structure to be represented as a sequence of atom and bond symbols, but quasi-SMILES imports the physicochemical properties and experimental conditions as a string of characters after SMILES symbol.

8.2 Principals of QSPR/QSAR Models

Although QSPR/QSAR modeling has been used for over five decades, many studies still do not follow the Organization of Economic Co-operation and Development (OECD) guidelines. Figure 8.1 summarizes the best practices for each step of QSPR/QSAR approach using models in peer reviewed literature. Dearden et al. have reported a detailed description of common errors in QSPR/QSAR research [22].

According to OECD guidelines, if a QSPR/QSAR study is to be reliable, the following five principles must be met: (i) a well-defined endpoint, (ii) an unambiguous algorithm, (iii) a defined applicability domain (AD), (iv) appropriate measures of goodness-of-fit, robustness, and predictivity, and (v) a mechanistic interpretation, if possible.

8.3 Monte Carlo Technique for Nano-QSPR/QSAR

8.3.1 SMILES and Quasi-SMILES

SMILES is a chemical notation system designed by Weininger et al. [23, 24]. According to the principles of molecular graph theory, SMILES uses a very small, natural grammar to specify precise structural features. The SMILES symbol system is also suitable for fast machine processing. Quasi-SMILES is an alternative to SMILES, which is used for substances considering physicochemical properties and experimental conditions.

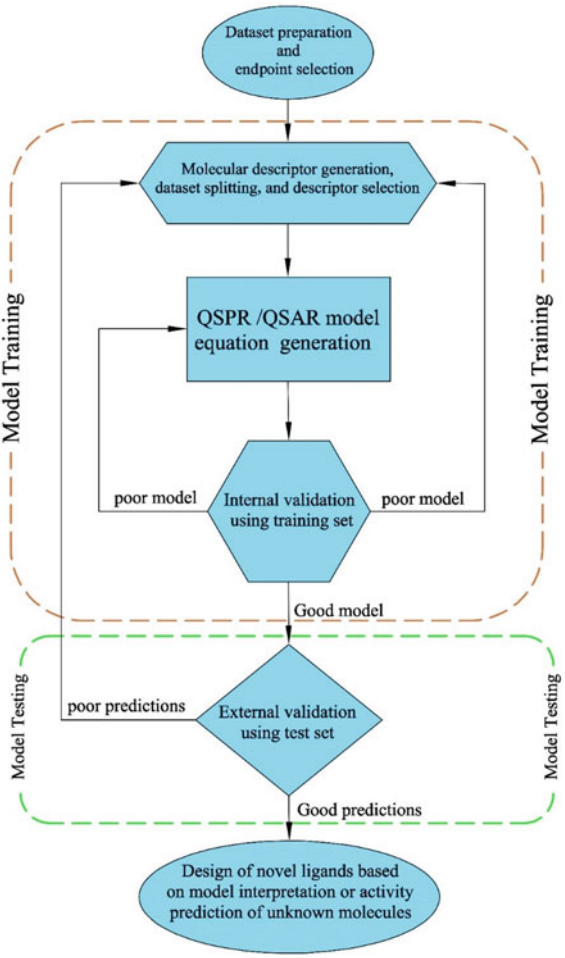

Fig. 8.1 General flowchart for QSPR/QSAR modeling

8.3.2 The Main Step for QSPR/QSAR Modeling by SMILES or Quasi-SMILES

CORrelation And Logic (CORAL) software (http://www.insilico.eu/coral) has two possibilities for building QSPR/QSAR models based on SMILES or quasi-SMILES. In the following, the method of preparing the input data for the CORAL software is described.

(a)

Set	ID	SMILES	EC_{50}
+	1	c1([N+](=O)[O-])ccc(cc1)N	3.51
+	2	c1(ccccc1)C#N	2.90
+	3	c1(Nc2c(cccc2)Cl)nc(nc(n1)Cl)Cl	5.19
#	4	c1(Oc2ccccc2)ccc(cc1)Br	5.93
-	5	c1(Oc2ccc(cc2)Cl)ccc(cc1)N	5.55
+	6	C1(NC2CCCCC2)CCCCC1	4.52
#	7	c1(Oc2ccccc2)ccccc1	5.14
#	8	N(CCCC)(CCCC)CCO	3.29
-	9	N(CCCC)(CCCC)CCCC	3.47
*	10	C(CCl)CCl	4.14
#	11	c1(CNC(C)C)ccccc1	4.53
#	12	c1(/N=N/c2ccccc2)ccccc1	5.04
+	13	C1(CO1)CCC=C	3.37
*	14	c1(/C=C/COC(=O)C)ccccc1	4.13
+	15	c1(ccccc1)N=C=S	5.64
+	16	c1(ccccc1)NC(=O)C	2.66
+	17	c1(ccc(cc1)OC)CCC(=O)C	3.64
*	18	c1(ccc(cc1)N)CCCCCCCCCCCC	6.48

(b)

Set	ID	Quasi-SMILES	Bandgap
-	3	O=[Al]O[Al]=OAZ8	3.1
+	4	O=[Al]O[Al]=OBN5	4.6
+	5	O=[Al]O[Al]=OBN4	4.2
-	7	O=[Al]O[Al]=OAZ1	5.8
+	10	O=[Al]O[Al]=OIZ4	4.06
#	11	O=[Al]O[Al]=OIZ4	4
#	12	O=[Al]O[Al]=OJZ4	4.2
-	13	O=[Al]O[Al]=OJZ6	4.15
-	14	O=[Al]O[Al]=OJZ7	4.09
-	15	O=[Al]O[Al]=OIN1	3.88
#	16	O=[Al]O[Al]=OAN1	3.6
-	17	O=[Ce]=OGN1	3.44
*	18	O=[Ce]=OGN1	3.38
+	19	O=[Ce]=OEQ2	2.78
+	20	O=[Ce]=OAR1	3.33
-	21	O=[Ce]=OAN1	3.49
*	22	O=[Ce]=OAN1	3.38
#	23	O=[Ce]=OKN1	3.03

Fig. 8.2 Sample of data based on **a** SMILES, and **b** quasi-SMILES as input for CORAL

8.3.2.1 Dataset Preparation for Models Based on SMILES

The SMILES string is a procedure for representing a two-dimensional molecular graph as a one-dimensional string that can show the connectivity and chirality of a molecule. In most cases, there are too many SMILES strings for a structure. Canonical SMILES gives a single 'canonical' form for any particular molecule. Molecular structures of desired compounds were transformed to canonical SMILES using different software such as Open Babel and ACD/ChemSketch program. Figure 8.2a, b indicates the sample of data based on SMILES, and quasi-SMILES as input for CORAL software, respectively. The first column indicates set, the second is compound ID, the third is SMILES/quasi-SMILES, and the last column is desired property/activity.

8.3.2.2 Dataset Preparation for Models Based on Quasi-SMILES

For building of QSPR/QSAR in different physicochemical properties and/or the experimental conditions of substance, one can use quasi-SMILES instead of SMILES of molecules. Dataset preparation for quasi-SMILES is same as SMILES, only SMILES is replaced by quasi-SMILES.

8.3.2.3 Quasi-SMILES Definition for Various Datasets/Endpoints

Quasi-SMILES is a sequence of symbols that not only represents the molecular structure but also the different conditions that can affect the endpoint under investigation. Eclectic data can include: different physical properties such as temperature,

Table 8.1 Distinction of standardized physiochemical features into classes 1–9 according to its value

Normalized value	Class
Norm(E) > 0.9	9
0.8 < Norm(E) < 0.9	8
0.7 < Norm(E) < 0.8	7
0.7 < Norm(E) < 0.6	6
0.6 < Norm(E) < 0.5	5
0.5 < Norm(E) < 0.4	4
0.4 < Norm(E) < 0.3	3
0.3 < Norm(E) < 0.2	2
0.2 < Norm(E) < 0.1	1
Norm(E) < 0.1	0

pressure, and assay of experiment to obtain an endpoint, or cell line type, time exposition, concentration, etc. to obtain an activity. The type and number of eclectic data can be different in various datasets.

Quasi-SMILES may be made by eclectic condition, only [4, 13] or combination of SMILES and eclectic conditions [5, 8]. The continuous eclectic conditions can be normalized by the following equation for assigning codes:

$$\text{Norm}(E_i) = \frac{\min(E_i) + E_i}{\min(E_i) + \max(E_i)} \tag{8.1}$$

E_i is its value of physicochemical parameter E, $\min(E_i)$ is minimum value of E, and $\max(E_i)$ indicates maximum value of E.

According to Table 8.1, the number of unique values in each parameter was less than 10; therefore, the quasi-SMILES descriptors representations could be coded by assigning a number between zero and nine in a single character.

A further development of the CORAL software (CORAL-2020) allows the display of experimental conditions through groups of symbols enclosed in parentheses. Table 8.2 shows the comparison codes in the last version (CORAL-2020) and old version of CORAL for creating quasi-SMILES in recently proposed models for cytotoxicity of metal oxide NPs [4]. One can see codes-2020 are quite transparent and consequently are more convenient for a user. As is clearly evident, CORAL-2020 codes being quite transparent and thus more user-friendly. Table 8.2 indicates codes used for the cell line, method, time exposition, concentration, nanoparticle size, and metal oxide type. Table 8.3 indicates the examples of quasi-SMILES obtained based on these codes.

Toropov and Toropova developed a QSAR model based on the new version of CORAL for the toxicity of ZnO NPs [14]. Experimental data from the literature are toxicity assessment of ZnO NPs and ZnO NPs coated with polyethylene glycol (PEG), which are investigated by intraperitoneal injections in the rat (50, 100, 200 mg/kg) for one month. Measurement of the toxic effects of renal factors

Table 8.2 Codes used for the cell line, method, time exposition, concentration, nanoparticle size, and metal oxide type to convert various information of the experimental data to quasi-SMILES [4]

Feature	Value or type	Code	Code 2020	Feature	Value or type	Code	Code 2020
Cell line	MCF-7	H	[MCF-7]	Normalized NPs size	$0.2 < \text{Norm(size)} \leq 0.3$	P	$[0.2 < \text{Norm(size)} \leq 0.3]$
	HT-1080	I	[HT-1080]		$0.3 < \text{Norm(size)} \leq 0.4$	Q	$[0.3 < \text{Norm(size)} \leq 0.4]$
	HepG-2	J	[HepG-2]		$0.4 < \text{Norm(size)} \leq 0.5$	R	$[0.4 < \text{Norm(size)} \leq 0.5]$
	HT-29	K	[HT-29]		$0.5 < \text{Norm(size)} \leq 0.6$	S	$[0.5 < \text{Norm(size)} \leq 0.6]$
	PC-12	L	[PC-12]		$0.9 < \text{Norm(size)} \leq 1.0$	T	$[0.9 < \text{Norm(size)} \leq 1.0]$
Method	MTT	M	[MTT]	Metal oxide type	SnO_2	1	$[SnO_2]$
	NRU	N	[NRU]		MnO_2	2	$[MnO_2]$
Time exposition	24	X	[T24]		ZnO	3	[ZnO]
	48	Y	[T48]		Bi_2O_3	4	$[Bi_2O_3]$
	72	Z	[T72]		NiO	5	[NiO]
Concentration ($\mu g\ mL^{-1}$)	5	A	[C5]		CeO_2	6	$[CeO_2]$
	10	B	[C10]		SiO_2	7	$[SiO_2]$
	25	C	[C25]		TiO_2	8	$[TiO_2]$
	50	D	[C50]				
	100	E	[C100]				
	200	F	[C200]				

Table 8.3 Some examples for quasi-SMILES extracted by codes indicated in Table 8.2

Cell line	Method	Time exposition (h)	Concentration ($\mu g\ mL^{-1}$)	Normalized NPs size	Metal oxide type	Quasi-SMILES	Quasi-SMILES (2020)	Cell viability (%)
MCF-7	MTT	24	5	0.2 < Norm(size) < 0.3	SnO_2	HMXAP1	[MTT][T24] [C5][0.2 < Norm(size) < 0.3][SnO_2]	97.0
MCF-7	MTT	24	25	0.2 < Norm(size) < 0.3	MnO_2	HMXAP2	[MTT][T24] [C25][0.2 < Norm(size) < 0.3][MnO_2]	81.0
HT-1080	NRU	24	10	0.9 < Norm(size) < 0.1	MnO_2	INXBP2	[NRU][T24] [C10][0.9 < Norm(size) < 0.1][MnO_2]	94.0
MCF-7	MTT	48	100	0.2 < Norm(size) < 0.3	ZnO	HMYEP3	[MTT][T48] [C100][0.2 < Norm(size) < 0.3][ZnO]	4.1
HepG2	MTT	72	200	0.2 < Norm(size) < 0.3	SiO_2	JMZFP7	[MTT][T72] [C200][0.2 < Norm(size) < 0.3][SiO_2]	57.0
HepG2	NRU	72	5	0.2 < Norm(size) < 0.3	SiO_2	JNZAP7	[NRU][T72] [C5][0.2 < Norm(size) < 0.3][SiO_2]	95.7
PC12	MTT	48	50	0.3 < Norm(size) < 0.4	TiO_2	LMYDR8	[MTT][T48] [C50][0.3 < Norm(size) < 0.4][TiO_2]	59.0
HepG2	MTT	24	5	0.5 < Norm(size) < 0.6	NiO	JMXAS5	[MTT][T24] [C5][0.5 < Norm(size) < 0.6][NiO]	99.5
HT-29	MTT	24	50	0.3 < Norm(size) < 0.4	CeO_2	KMXDQ6	[MTT][T24] [C50][0.3 < Norm(size) < 0.4][CeO_2]	91.0
MCF-7	MTT	24	200	0.9 < Norm(size) < 1.0	Bi_2O_3	HMXFT4	[MTT][T24] [C200][0.9 < Norm(size) < 1.0][Bi_2O_3]	51.7

Table 8.4 Codes used as fragments of quasi-SMILEs and their meaning

Code	Meaning
[15d]	Renal factor measured after fifteen days post-injection
[30d]	Renal factor measured after thirty days post-injection
[RF1]	Variation in creatinine as renal factor
[RF2]	Variation in uric acid as renal factor
[RF3]	Variation in blood urea nitrogen as renal factor
[50]	50 mg per kg of body weight
[100]	100 mg per kg of body weight
[200]	200 mg per kg of body weight
[ZnO]	Uncoated ZnO NPs is injected
[ZnO][peg]	ZnO coated by PEG NPs is injected

including creatinine, uric acid, and blood urea nitrogen was measured after 15 and 30 days after injection. Table 8.4 shows the quasi-SMILES attributes together with experimental conditions. Table 8.5 represents examples of available quasi-SMILES obtained based on this condition and related activity.

Toropova et al. developed new nano-QSAR model for predicting toxicity of nano-mixtures to *Daphnia magna* based on quasi-SMILES [25]. The binary mixtures of TiO_2 NPs and with of one of the second component including $AgNO_3$, $Cd(NO_3)_2$, $Cu(NO_3)_2$, $CuSO_4$, Na_2HAsO_4, $NaAsO_2$, benzylparaben, and benzophenone-3 have been investigated. Quasi-SMILES contain the following information: (1) Second

Table 8.5 Some examples for quasi-SMILES extracted by codes presented in Table 8.4

Time exposition (days)	Renal factor type	NPs (mg/kg)	NPs type	Quasi-SMILES	Experimental renal factor
15	Creatinine	50	ZnO	[15d][RF1][50][ZnO]	0.79
15	Creatinine	100	ZnO	[15d][RF1][50][ZnO]	0.87
15	Creatinine	100	ZnO-peg	[15d][RF1][50][ZnO][peg]	0.50
15	Uric acid	100	ZnO-peg	[15d][RF2][200][ZnO][peg]	1.37
15	Blood urea nitrogen	100	ZnO-peg	[15d][RF3][100][ZnO][peg]	62.30
30	Creatinine	100	ZnO	[30d][RF1][100][ZnO]	0.72
30	Uric acid	50	ZnO-peg	[30d][RF2][50][ZnO][peg]	1.30
30	Blood urea nitrogen	50	ZnO-peg	[30d][RF3][50][ZnO][peg]	50.33
30	Blood urea nitrogen	200	ZnO-peg	[30d][RF3][200][ZnO][peg]	49.0

Mixed substance	Core diameter of TiO_2 NPs	Zeta potential of TiO_2 NPs	Mole fraction of TiO_2 NPs	Mol fraction of mixed substance	Exposure time (h)
	C	Z	F1	F2	E
$Cd(NO_3)_2$	30	-16.3	0.948	0.052	48
Ag_2NO_3	20	-1.8	0.999	0.001	48
Ag_2NO_3	20	-1.8	0.999	0.001	7
Na_2HAsO_4	20	-1.8	0.982	0.018	48
Na_2HAsO_4	20	-1.8	0.980	0.020	48

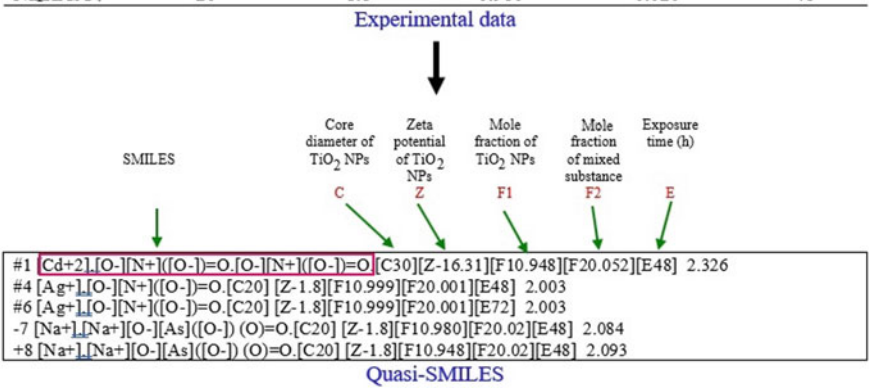

Fig. 8.3 Transfer of experimental data into quasi-SMILES [25]

component of mixture represented by SMILES; (2) core diameter of TiO_2 NPs; (3) Zeta potential of TiO_2 NPs; (4) mole fraction of TiO_2 NPs; (5) mole fraction of mixed substance; and (6) exposure time. Figure 8.3 shows the transformation of the experimental condition and substance into the quasi-SMILES.

8.3.2.4 Model Development

Model development has several steps that can be organized in CORAL software and does not require any software for data partitioning, descriptor generation, and model validation. In the following sections, the main step for QSPR/QSAR modeling using CORAL software is described.

8.3.2.5 Dataset Splitting

After the preparation and curation of dataset, the next step of building a QSAR/QSPR model for an endpoint by CORAL software (http://www.insilico.eu/coral) is loading an array of lines. Each line consists of four components.

The first column is the types of set which '+', '−', '#', and '*' indicate the active training, passive training, calibration, and validation, respectively (Fig. 8.2).

- The second column without space with type of set is number or ID of compound.
- The third column is quasi-SMILES.
- The last column is endpoint value.

After the preparation of input file, the dataset was splitted into training, passive training, calibration, and validation sets using CORAL software, randomly with desired present for each set.

8.3.2.6 Monte Carlo Optimization Process

Quasi-SMILES is a group of attributes where each attribute group is converted into a group of coefficients called correlation weights. Monte Carlo optimization refines the correlation weights that provide numerical data on them, which maximizes the predictive potential of a model as much as possible. Figure 8.4 shows the flowchart of one cycle of Monte Carlo optimization of correlation weights (*n* is the number of correlation weights that contribute to model construction).

There are different target functions (TFs) in CORAL software for Monte Carlo optimization [25–29], which are introduced below four TFs:

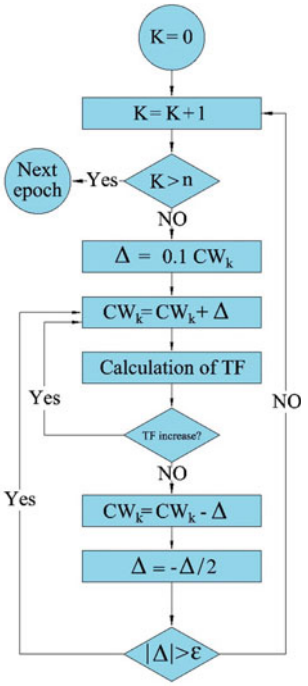

Fig. 8.4 Flowchart of one cycle of the Monte Carlo optimization for finding correct correlation weights (*n* is the number of correlation weights that contribute to model construction)

$$TF_0 = r_{AT} + r_{PT} - |r_{AT} - r_{PT}| \times C \tag{8.2}$$

$$TF_1 = TF_1 + IIC_C \times W_{IIC} \tag{8.3}$$

$$TF_2 = TF_1 + CII_C \times W_{CII} \tag{8.4}$$

$$TF_3 = TF_1 + IIC_C \times W_{IIC} + CII_C \times W_{CII} \tag{8.5}$$

r_{AT} and r_{PT} represent the correlation coefficient between the experimental and predicted endpoints for active and passive training sets, respectively. Empirical constant (C), W_{IIC}, and W_{CII} have a defined numerical value [1, 18, 30–33].

IIC_C is the index of ideality correlation. IIC_C is obtained based on the calibration set as follows:

$$CII_C = r_C \frac{\min\left({}^{-}MAE_C, {}^{+}MAE_C\right)}{\max({}^{-}MAE_C, {}^{+}MAE_C)} \tag{8.6}$$

$${}^{-}MAE_C = \frac{1}{{}^{-}N} \sum |\Delta_i|, \; {}^{-}N \text{ is the number of } \Delta_i < 0 \tag{8.7}$$

$${}^{+}MAE_C = \frac{1}{{}^{-}N} \sum |\Delta_i|, \; {}^{+}N \text{ is the number of } \Delta_i \geq 0 \tag{8.8}$$

$$\Delta_i = Obs_i - Calc_i \tag{8.9}$$

The Obs_i and $Calc_i$ are the experimental and predicted endpoint for ith compound.

The correlation intensity index (CII), like IIC criteria, was developed to modify the quality of the Monte Carlo optimization used to build the QSPR/QSAR models. CII is formulated as follows:

$$CII = 1 - \sum \Delta R_i^2 > 0, \text{ If } \Delta R_i^2 < 0 \text{ then } \Delta R_i^2 = 0 \tag{8.10}$$

$$\Delta R_i^2 = R_i^2 - R^2 \tag{8.11}$$

where R^2 is the coefficient of determination for all endpoints and R_i^2 is the coefficient of determination for all endpoints in the absence of ith compound. Therefore, if ΔR_i^2 is greater than zero, the meaning of ith is an 'opposite' for the correlation between the experimental and calculated values of the set.

A small sum of ΔR_i^2 means a more 'intensive' correlation.

The CORAL model for an endpoint (EP) is defined by the below equation:

$$EP = C_0 + C_1 \times DW(T, N) \tag{8.12}$$

C_0 and C_1 represent regression coefficients, T is a threshold, and N is the number of optimization cycles. The DCW(T, N) is defined as the below equation:

$$\mathrm{DCW}(T, N) = \sum \mathrm{CW}(S_k) \tag{8.13}$$

where S_k represents the symbol of a quasi-SMILES line; the CW(S_k) shows the correlation weights of S_k.

8.3.2.7 Applicability Domain

The AD of QSAR/QSAR models for CORAL software is determined in two steps based on the distribution of SMILES or quasi-SMILES features in the training and calibration sets:

Step 1: the statistical defect (d_k) is calculated for each involved (unblocked) SMILES or quasi-SMILES feature (S_k) to build the model with the following equation:

$$d_k = \frac{\left|P(S_k) - P'(S_k)\right|}{N(S_k) + N'(S_k)} \tag{8.14}$$

here, $P(S_k)$ and $P'(S_k)$ represent the probability of S_k in the active training set and calibration sets, respectively; $N(S_k)$ and $N'(S_k)$ denote the frequencies of S_k in the active training and calibration sets, respectively.

Step 2: the quasi-SMILES (D_i) statistical defect of all compounds is defined according to the following equation:

$$D_i = \sum_{k=1}^{N_\mathrm{A}} d_k \tag{8.15}$$

here N_A denotes the number of non-blocked quasi-SMILES features in the quasi-SMILES.

Quasi-SMILES falls in the AD if:

$$D_i < 2 \times \overline{D} \tag{8.16}$$

where $\overline{D}$ represents average statistical defect of the training set.

8.3.2.8 Model Validation

Validation, as the fourth principle of OECD, is recognized as an intrinsic component to check the robustness, predictability, and reliability of any QSPR/QSAR models. There are three approaches to examine the robustness, reliability, and predictive potential of the QSPR/QSAR models in CORAL software, including:

- Internal validation
- External validation
- Y-scrambling or data randomization.

Various statistical criteria such as determination coefficient (R^2), concordance correlation coefficient (CCC), cross-validated correlation coefficient (Q^2), Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, standard error of estimation (s), mean absolute error (MAE), Fischer ratio (F) and root-mean-square error (RMSE), R^2_{m}, and average of R^2_{m} metric ($\overline{R^2_{\mathrm{m}}}$) are calculated to authenticate the QSPR/QSAR models constructed based on the Monte Carlo optimization by the CORAL software. Table 8.6 indicates the mathematical equation of diverse statistical benchmark of the predictive potential for CORAL models.

Table 8.6 Mathematical formulation of different statistical benchmark of the predictive potential for CORAL models

Criterion of the predictive potential	Description	References
$Q^2 = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \overline{y})^2}$	Leave-one-out cross-validated correlation coefficient	[34]
$Q^2_{F1} = 1 - \frac{\sum_{i=1}^{N_{\mathrm{EXT}}}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N_{\mathrm{EXT}}}(\hat{y}_i - \overline{y}_{\mathrm{TR}})^2}$	Criteria of predictability	[35]
$Q^2_{F2} = 1 - \frac{\sum_{i=1}^{N_{\mathrm{EXT}}}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N_{\mathrm{EXT}}}(\hat{y}_i - \overline{y}_{\mathrm{EXT}})^2}$	Criteria of predictability	[35]
$Q^2_{F3} = 1 - \frac{\left[\sum_{i=1}^{N_{\mathrm{EXT}}}(\hat{y}_i - y_i)^2\right]/N_{\mathrm{EXT}}}{\left[\sum_{i=1}^{N_{\mathrm{EXT}}}(\hat{y}_i - \overline{y}_{\mathrm{EXT}})^2\right]/N_{\mathrm{TR}}}$	Criteria of predictability	[36]
$R^2_{\mathrm{m}} = R^2 \times \left(1 - \sqrt{R^2 - R^2_0}\right)$		[36]
$\overline{R^2_{\mathrm{m}}} = \frac{R^2_{\mathrm{m}}(x,y) - R^2_{\mathrm{m}}(y,x)}{2}$	Average of R^2_{m} metric	[36]
$\mathrm{CCC} = \frac{2\sum(x-\overline{x})(y-\overline{y})}{\sum(x-\overline{x})^2 + \sum(y-\overline{y})^2 + n(\overline{x}-\overline{y})^2}$	Concordance correlation coefficient	[37]
$C_{R^2_{\mathrm{p}}} = R\sqrt{\left(R^2 - R^2_{\mathrm{r}}\right)}$	Coefficient of determination for Y-randomization	[38]

8.3.2.9 Mechanistic Interpretation

The 5th OECD principle focuses on mechanistic interpretation of the QSPR/QSAR model if possible. The model interpretation is used to examine the critical and responsible attributes that influence the endpoint. Finally, the new compounds are designed based on these attributes. In the QSPR/QSAR modeling based on the CORAL software, the same structural attributes (S_k) collected from three or more different splits are used to perform the mechanistic interpretation [39–42]. These structural attributes (S_k) are divided into three categories according to previous studies:

- Increasing factor if the CW(S_k) is positive in all splits and in three attempts,
- Decreasing factor if the CW(S_k) is negative in all splits and in three attempts,
- Undefined attributes if the CW(S_k) is both positive and negative [43–45].

8.4 Examples of Quasi-SMILES-Based QSPR/QSAR Models

Some examples of QSAR/QSPR models base on quasi-SMILES with CORAL software using different TFs are presented in Table 8.7.

8.5 Conclusion and Future Direction

QSPR/QSAR modeling based on SMILES and quasi-SMILES by CORAL software is useful for big dataset. In CORAL software, QSPR/QSAR generally follows the five OECD principles. In addition, additional principles may be defined practically for nano-QSPR/QSAR that reflect the nature of the nanomaterial under investigation. For example, the new principles should take into account the test conditions and the quality of the applied equipment.

The use of CORAL software in building QSPR/QSAR models for nanomaterials in different conditions is simple, and the models can be easily predicted and interpreted. There are very good TFs (TF_0–TF_3) to find reliable correlation weights and this is one of the important capabilities of CORAL for building excellent QSAR/QSAR models. The type and number of input features can change the performance of a QSAR/QSPR model. But there is one of a shortcoming for CORAL software, the user can use only CORAL software descriptors, and it is impossible to add the other descriptors produced by other descriptor generators.

In CORAL software, there is only Monte Carlo algorithm to find correlation weights. The use of various algorithms can increase the quasi-SMILES QSPR/QSAR performance. Data splitting in CORAL software is done randomly; the possibility of using different methods of data splitting can increase the validity of the models.

Table 8.7 Some examples of QSAR/QSPR models base on quasi-SMILES with CORAL software using different TFs

Compound	Endpoint	No. of quasi-SMILES	Eclectic data								
Nano-mixtures	EC50 for *Daphnia magna*	67	Second component of mixture	Core diameter of TiO_2 NPs	Zeta potential of TiO_2 NPs			Mole fraction of TiO_2 NPs			Mole fraction of substance
Nano-mixtures	EC50 for *Daphnia magna*	67	Second component of mixture	Core diameter of TiO_2 NPs	Zeta potential of TiO_2 NPs		Mole fraction of TiO_2 NPs	Mole fraction of substance			Exposure time
Ag NPs	pLC_{50} for *Daphnia magna* and zebrafish	170	Status of NPs (bare, coat, cons)		Organisms (Daphnia or zebrafish)			Zeta potential			–
Metal oxide NPs	Cell viability (%)	83	Cell line	Assay	Time exposition		Concentration	NP size	Metal oxide type		–
Metal–organic frameworks	Log(CO_2 uptake)	260	BET	Specific surface area	Pore volume		Pressure	Temperature	–		–
Metal oxide NPs	Band gap	198	Synthesis method		Annealing temperature			Crystalline size			–
Cadmium containing quantum dots	Hepatic cell viability (%)	115	Core	Norm diameter	Charge	Surface modification	Assay type	Exposure time	Delivery type		QD concentration
CdSe quantum dots with ZnS shell	HeLa cell viability (%)	61	Size	Surface ligand	Surface charge		Surface modification	Assay type	Exposure time		QD concentration

(continued)

Table 8.7 (continued)

Compound	Endpoint	No. of quasi-SMILES	Eclectic data						
Metal oxide NPs	Cell viability (%)	83	Cell line	Assay	Time exposition	Concentration	NPs size	Metal oxide type	–
Multiwalled carbon nanotubes	Cell viability (%) of human lung cells	255	Mean diameter	Mean length	Surface area	Toxic assay method	Cell line	Exposure time	Dose
Metal oxide NPs	Cell viability (%) of human lung and skin cells	336	Core size	Hydrodynamic size	Surface charge	Dose	Toxic assay method	Cell line	–
Nanofluids	Viscosity ratio	100	Size	Volume fraction (%)	–	–	–	–	–
Nanofluids	Viscosity ratio	100	Size	Volume fraction (%)	–	–	–	–	–
Nanofluids	Viscosity ratio	100	Shape	Volume fraction (%)	–	–	–	–	–
Nanofluids	Viscosity ratio	100	Shape	Volume fraction (%)	–	–	–	–	–
Nanozeolites	Cell viability (%)	120	Time exposition	Concentration	Cell type	Sample	–	–	–
Metal oxide NPs	Cell membrane damage	137	Chemical element	Concentration	Time of exposure	–	–	–	–

(continued)

Table 8.7 (continued)

Compound	TF	Statistical indicator value				References
		R^2_{Train}	R^2_{Val}	S_{Train}	S_{Val}	
Nano-mixtures	TF_1	0.64–0.90	0.32–0.80	0.34–0.72	0.48–0.95	[25]
Nano-mixtures	TF_2					
Ag NPs	TF_3		0.62–0.71	0.34–0.55	0.58–0.60	[46]
Metal oxide NPs	TF_1	0.92–0.93	0.87–0.94	7.8–9.0	8.6–13.1	[4]
Metal–organic frameworks	TF_1	0.74–0.77	0.74–0.77	0.25–0.29	0.20–0.26	[5]
Metal oxide NPs	TF_1	0.85–0.88	0.82–0.90	0.33–0.38	0.32–0.38	[8]
Cadmium containing quantum dots	TF_1	0.70–0.90	0.63–0.81	8.7–14.5	–	[9]
CdSe quantum dots with ZnS shell	TF_1	0.84–0.96	0.59–0.93	0.11–0.35	0.27–0.51	[3]
Metal oxide NPs	TF_2	0.95–0.97	0.87–0.94	5.21–6.53	6.2–11.0	[13]
Multiwalled carbon nanotubes	TF_0	0.60–0.80	0.81–0.88	13.7–15.2	8.0–16.4	[47]
Metal oxide NPs	TF_0	0.71–0.73	0.70–0.76	–	–	[20]
Nanofluids	TF_1	0.84–0.92	0.73–0.90	0.75–0.11	0.05–0.08	[18]
Nanofluids	TF_2	0.85–0.91	0.90–0.94	0.08–0.10	0.04–0.09	
Nanofluids	TF_1	0.74–0.88	0.82–0.92	0.25–0.57	0.26–0.42	
Nanofluids	TF_2	0.73–0.85	0.90–0.94	0.40–0.51	0.20–0.51	
Nanozeolites	TF_0	0.83–0.89	0.72–0.81	0.02–0.04	0.02–0.03	[48]
Metal oxide NPs	TF_0	0.50–0.54	0.67–0.93	0.38–0.39	0.25–0.40	[49]

Since the correlation weight of the descriptors in this software is calculated through Monte Carlo approach, the use of consensus modeling can dramatically increase the prediction results.

References

1. Ahmadi S, Ghanbari H, Lotfi S, Azimi N (2021) Mol Divers 25(1):87–97. https://doi.org/10.1007/s11030-019-10026-9
2. Achary PGR, Toropova AP, Toropov AA (2019) Food Res Int 122:40–46. https://doi.org/10.1016/j.foodres.2019.03.067
3. Kumar A, Kumar P (2021) J Hazard Mater 402:123777. https://doi.org/10.1016/j.jhazmat.2020.123777
4. Ahmadi S (2020) Chemosphere 242:125192. https://doi.org/10.1016/j.chemosphere.2019.125192
5. Ahmadi S, Ketabi S, Qomi M (2022) New J Chem 46:8827–8837. https://doi.org/10.1039/D2NJ00596D
6. Lotfi S, Ahmadi S, Kumar P (2021) RSC Adv 11:33849–33857. https://doi.org/10.1039/D1RA06861J
7. Ahmadi S, Khazaei MR, Abdolmaleki A (2014) Med Chem Res 23:1148–1161. https://doi.org/10.1007/s00044-013-0716-z
8. Ahmadi S, Aghabeygi S, Farahmandjou M, Azimi N (2021) Struct Chem 32:1893–1905. https://doi.org/10.1007/s11224-021-01748-4
9. Kumar P, Kumar A (2021) Nanotoxicology 15:1199–1214. https://doi.org/10.1080/17435390.2021.2008039
10. Ghasedi N, Ahmadi S, Ketabi S, Almasirad A (2022) J Recept Signal Transduct 42:418–428. https://doi.org/10.1080/10799893.2021.1988971
11. Ahmadi S, Moradi Z, Kumar A, Almasirad A (2022) J Recept Signal Transduct 42:361–372. https://doi.org/10.1080/10799893.2021.1957932
12. Karelson M, Lobanov VS, Katritzky AR (1996) Chem Rev 96:1027–1044. https://doi.org/10.1021/cr950202r
13. Ahmadi S, Toropova AP, Toropov AA (2020) Nanotoxicology 14:1118–1126. https://doi.org/10.1080/17435390.2020.1808252
14. Toropov AA, Toropova AP (2021) Sci Total Environ 772:145532. https://doi.org/10.1016/j.scitotenv.2021.145532
15. Toropov AA, Toropova AP (2020) Sci Total Environ 737:139720. https://doi.org/10.1016/j.scitotenv.2020.139720
16. Ahmadi S, Akbari A (2018) Environ Res 29:895–909. https://doi.org/10.1080/1062936X.2018.1526821
17. Lotfi S, Ahmadi S, Kumar P (2022) RSC Adv 12:24988–24997. https://doi.org/10.1039/D2RA03936B
18. Jafari K, Fatemi MH, Toropova AP, Toropov AA (2022) Chemom Intell Lab Syst 222:104500. https://doi.org/10.1016/j.chemolab.2022.104500
19. Toropov A, Toropova A, Lombardo A, Roncaglioni A, Lavado G, Benfenati E (2021) Environ Res 32:463–471. https://doi.org/10.1080/1062936X.2021.1914156
20. Choi J-S, Trinh TX, Yoon T-H, Kim J, Byun H-G (2019) Chemosphere 217:243–249. https://doi.org/10.1016/j.chemosphere.2018.11.014
21. Lotfi S, Ahmadi S, Zohrabi P (2020) Struct Chem 31:2257–2270. https://doi.org/10.1007/s11224-020-01568-y
22. Dearden JC, Cronin MTD, Kaiser KLE (2009) Environ Res 20:241–266. https://doi.org/10.1080/10629360902949567

23. Weininger D (1988) J Chem Inf Model 28:31–36. https://doi.org/10.1021/ci00057a005
24. Weininger D, Weininger A, Weininger JL (1989) J Chem Inf Comput Sci 29:97–101. https://doi.org/10.1021/ci00062a008
25. Toropova AP, Toropov AA, Fjodorova N (2022) NanoImpact 28:100427. https://doi.org/10.1016/j.impact.2022.100427
26. Kumar P, Kumar A, Lal S, Singh D, Lotfi S, Ahmadi S (2022) J Mol Struct 1265:133437. https://doi.org/10.1016/j.molstruc.2022.133437
27. Azimi A, Ahmadi S, Kumar A, Qomi M, Almasirad A (2022) Polycycl Aromat Compd 1–21. https://doi.org/10.1080/10406638.2022.2067194
28. Ahmadi S, Lotfi S, Afshari S, Kumar P, Ghasemi E (2021) Environ Res 32:1013–1031. https://doi.org/10.1080/1062936X.2021.2003429
29. Ahmadi S, Mehrabi M, Rezaei S, Mardafkan N (2019) J Mol Struct 1191:165–174. https://doi.org/10.1016/j.molstruc.2019.04.103
30. Nimbhal M, Bagri K, Kumar P, Kumar A (2020) Struct Chem 31:831–839. https://doi.org/10.1007/s11224-019-01468-w
31. Toropova AP, Duchowicz PR, Saavedra LM, Castro EA, Toropov AA (2020) Mol Inform 39:1900070. https://doi.org/10.1002/minf.201900070
32. Toropova AP, Toropov AA, Carnesecchi E, Benfenati E, Dorne JL (2020) Environ Sci Pollut Res 27:13339–13347. https://doi.org/10.1007/s11356-020-07820-6
33. Kumar P, Kumar A (2021) J Mol Struct 1246:131205. https://doi.org/10.1016/j.molstruc.2021.131205
34. Shayanfar A, Shayanfar S (2014) Eur J Pharm Sci 59:31–35. https://doi.org/10.1016/j.ejps.2014.03.007
35. Consonni V, Ballabio D, Todeschini R (2009) J Chem Inf Model 49:1669–1678. https://doi.org/10.1021/ci900115y
36. Roy K, Kar S (2014) Eur J Pharm Sci 62:111–114. https://doi.org/10.1016/j.ejps.2014.05.019
37. Lin LI-K (1992) Biometrics 48:599. https://doi.org/10.2307/2532314
38. Rücker C, Rücker G, Meringer M (2007) J Chem Inf Model 47:2345–2357. https://doi.org/10.1021/ci700157b
39. Manisha, Chauhan S, Kumar P, Kumar A (2019) Environ Res 30:145–159. https://doi.org/10.1080/1062936X.2019.1568299
40. Kumar P, Kumar A, Sindhu J, Lal S (2019) Drug Res (Stuttg) 69:159–167. https://doi.org/10.1055/a-0652-5290
41. Kumar P, Kumar A, Sindhu J (2019) Environ Res 30:63–80. https://doi.org/10.1080/1062936X.2018.1564067
42. Kumar P, Kumar A, Sindhu J (2019) SAR QSAR Environ Res 30:525–541. https://doi.org/10.1080/1062936X.2019.1629998
43. Toropov AA, Toropova AP, Benfenati E, Gini G, Leszczynska D, Leszczynski J (2012) Anticancer Agents Med Chem 12:807–817. https://doi.org/10.2174/187152012802650255
44. Nesmerak K, Toropov AA, Toropova AP, Kohoutova P, Waisser K (2013) Eur J Med Chem 67:111–114. https://doi.org/10.1016/j.ejmech.2013.05.031
45. Veselinović AM, Milosavljević JB, Toropov AA, Nikolić GM (2013) Eur J Pharm Sci 48:532–541. https://doi.org/10.1016/j.ejps.2012.12.021
46. Toropov AA, Kjeldsen F, Toropova AP (2022) Chemosphere 303:135086. https://doi.org/10.1016/j.chemosphere.2022.135086
47. Trinh TX, Choi J-S, Jeon H, Byun H-G, Yoon T-H, Kim J (2018) Chem Res Toxicol 31:183–190. https://doi.org/10.1021/acs.chemrestox.7b00303
48. Leone C, Bertuzzi EE, Toropova AP, Toropov AA, Benfenati E (2018) Chemosphere 210:52–56. https://doi.org/10.1016/j.chemosphere.2018.06.161
49. Toropova AP, Toropov AA, Benfenati E, Korenstein R, Leszczynska D, Leszczynski J (2015) Environ Sci Pollut Res 22:745–757. https://doi.org/10.1007/s11356-014-3566-4