Mol Divers (2015) 19:563–575
DOI 10.1007/s11030-015-9587-1

SHORT COMMUNICATION

CORAL: model for no observed adverse effect level (NOAEL)

Andrey A. Toropov[1] · Alla P. Toropova[1] · Fabiola Pizzo[1] · Anna Lombardo[1] · Domenico Gadaleta[1,2] · Emilio Benfenati[1]

Received: 24 November 2014 / Accepted: 21 March 2015 / Published online: 8 April 2015

Abstract The in vivo repeated dose toxicity (RDT) test is intended to provide information on the possible risk caused by repeated exposure to a substance over a limited period of time. The measure of the RDT is the no observed adverse effect level (NOAEL) that is the dose at which no effects are observed, i.e., this endpoint indicates the safety level for a substance. The need to replace in vivo tests, as required by some European Regulations (registration, evaluation authorization and restriction of chemicals) is leading to the searching for reliable alternative methods such as quantitative structure–activity relationships (QSAR). Considering the complexity of the RDT endpoint, for which data quality is limited and depends anyway on the study design, the development of QSAR for this endpoint is an attractive task. Starting from a dataset of 140 organic compounds with NOAEL values related to oral short term toxicity in rats, we developed a QSAR model based on optimal descriptors calculated with simplified molecular input-line entry systems and the graph of atomic orbitals by the Monte Carlo method, using CORAL software. Three different splits into the training, calibration, and validation sets are studied. The mechanistic interpretation of these models in terms of molecular fragment with positive or negative contributions to the endpoint is discussed. The probabilistic definition for the domain of applicability is suggested.

Keywords QSAR · Monte Carlo method · Ecology · Repeated dose toxicity · NOAEL · CORAL software

Abbreviations

NOAEL	No observed adverse effect level
LOAEL	Lowest observed adverse level
QSAR	Quantitative structure–activity relationship
REACH	Registration, evaluation authorization, and restriction of chemicals
RDT	Repeated dose toxicity
SMILES	Simplified molecular input-line entry systems
MoS	Margin of safety
CSA	Chemical safety assessment
DNEL	Derived no effect levels
TOPKAT	Toxicity prediction by komputer assisted technology
BMD	Benchmark dose
CORAL	CORrelation and logic
QSPR	Quantitative structure–property relationship
MLR	Multiple linear regression
MRTD	Maximum recommended therapeutic dose

Introduction

In risk assessment, repeated dose toxicity (RDT) provides information on the adverse toxicological effects which can be induced by repeated exposure to a substance over a limited period of time [1,2]. No observed adverse effect level (NOAEL) indicates the dose at which no effects are observed,

Electronic supplementary material The online version of this article (doi:10.1007/s11030-015-9587-1) contains supplementary material, which is available to authorized users.

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

[1] Laboratory of Environmental Chemistry and Toxicology, IRCCS - Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20159 Milan, Italy

[2] Dipartimento di Farmacia, Scienze del Farmaco, Università degli Studi di Bari "Aldo Moro", Via Orabona 4, 70125 Bari, Italy

and the lowest observed adverse effect level (LOAEL) is the lowest dose at which adverse effect can serve as measure of the RDT [2]. The NOAEL is a toxicological requirement imposed by registration, evaluation authorization, and restriction of chemicals (REACH). When the NOAEL is not available, LOAEL can be used for the same purpose [1,3–7].

REACH is recommended development of methods such as quantitative structure–activity relationships (QSARs) to assess the potential harmful effects of chemicals [3]. The aim of QSAR models is to establish correlation between the chemical structure of a substance and biological activity [8]. QSAR for estimating NOAEL/LOAEL has been described in the scientific literature [2,9,10].

The aim of this work was to build up QSAR model for the NOAEL by means of the CORAL software (http://www.insilico.eu/coral/).

Materials and methods

Data

Experimental data were obtained from the OECD toolbox version 3.2. (http://www.qsartoolbox.org/download), downloading HESS and Munro databases and from the US EPA's integrated risk information system (IRIS) database (http://www.epa.gov/IRIS/). For all the compounds the canonical simplified molecular input-line entry systems (SMILES) were obtained by searching through CAS numbers on the PubChem Compound website (https://pubchem.ncbi.nlm.nih.gov/). The correspondence between CAS numbers and chemical structures was further checked using ChemID Plus Advanced website (http://chem.sis.nlm.nih.gov/chemidplus/). The numerical data from various sources were compared. If several values were available for the same compound, we used the lowest.

All doubtful or inorganic compounds, salts, and mixtures were eliminated, because the relationships between molecular structure and the NOAEL are very complex. We considered only data referring to 90 days of oral administration in rats and rejected reproductive toxicity studies. It is to be noted that the exchange of the 90-day study by shorter testing is an attractive alternative [11]. Taking into account this circumstance, values for 28 days of treatment were considered but, in order to have consistent data, they were divided by a factor of 3, as specified by the scientific committee on consumer safety (SCCS) in order to approximate the 90-day NOAEL [1]. After the above selection, about four hundreds of various substances with small molecules (e.g., 2–3 atoms) and vice versa with extremely large molecules (e.g., 100 or more atoms), molecules with specific groups, such as [N+], [NH4+], [nH], etc., and substances with molecules containing many various cycles / heterocycles were remained. Under such circumstances, the following limitations were used in the selection of compounds for the work set: (i) too large and, vice versa, too small molecules were removed (practically, molecules which can be represented by SMILES with length less than 70 and larger than 10 symbols, were selected); (ii) molecules which have only one cycles or have no cycles at all were selected; and (iii) molecules with special groups (indicated by square brackets) were removed from the work set. Thus, the dataset of 140 compounds has been selected. All values were converted to decimal logarithms (lgNOAEL). These compounds were randomly split into training, calibration, and validation sets three times.

Optimal descriptors

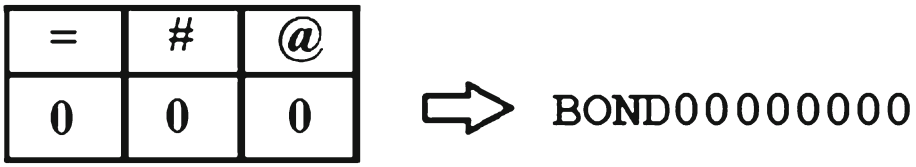
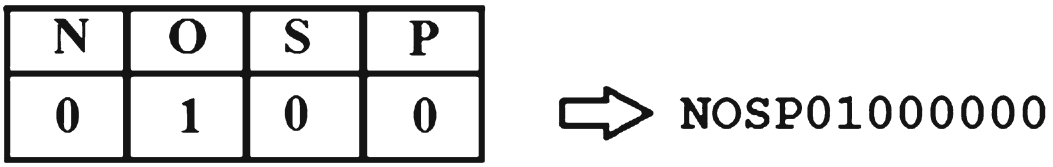
The hybrid optimal descriptors calculated with two representation of the molecular structure by SMILES [12] and by graph of atomic orbitals (GAO) [13,14] were used for QSAR analysis:

$$\begin{aligned} \mathrm{DCW}(T, N) = {} & \mathrm{CW(NOSP)} + \mathrm{CW(HALO)} \\ & + \mathrm{CW(BOND)} + \sum \mathrm{CW}(\mathrm{EC0_k}) \\ = {} & \sum \mathrm{CW}(\mathrm{SA_k}), \end{aligned} \quad (1)$$

where $\mathrm{SA_k}$ is a structural attribute extracted from SMILES (NOSP, HALO, and BOND represented in Table 1) or from the graph of atomic orbitals ($\mathrm{EC0_k}$ represented in Table 2); the CW(x) is so-called correlation weight for a structural attribute extracted from SMILES (i.e., NOSP, HALO, and BOND) or from GAO (i.e., $\mathrm{EC0_k}$). The correlation weights are coefficients which are used for calculation with Eq. 1: the numerical data on the correlation weights are calculated with the Monte Carlo optimization which gives maximum of the determination coefficient (r^2) between experimental and predicted lgNOAEL for the training set. The T is threshold, i.e., coefficient for discrimination of $\mathrm{SA_k}$ into two categories (i) rare (if frequency of $\mathrm{SA_k}$ in the training set is less than T) and (ii) not rare (if frequency of $\mathrm{SA_k}$ is larger than T in the training set). The N is the total number of the Monte Carlo method epochs ($N = 1, 2, \ldots, 30$). The NOSP is a descriptor indicating the presence (absence) of nitrogen, oxygen, sulfur, and phosphorus; HALO is a descriptor indicating the presence (absence) of halogens (i.e., "F," "Cl," "Br," and "I"); the BOND is a descriptor indicating the presence (absence) of double, triple, and stereo chemical bonds; Table 1 contains clarifications for SMILES attributes involved in building up a model. Table 2 contains an example of representation of the molecular structure by the molecular graphs. The $\mathrm{EC0_k}$ are extended connectivity of zero order in the GAO [13,14].

The modeling approach examined in this study includes three steps:

Table 1 The examples of BOND, NOSP, and HALO descriptors

SMILES attribute	Example of the representation for the CORAL software
BOND	The presence / absence of double ("="), triple ("#"), and stereo chemical ("@") bonds, e.g., if SMILES = "CCC(O)CC"
	
NOSP	Presence (absence) of nitrogen, oxygen, sulfur, and phosphorus, e.g., if SMILES = "CCC(O)CC"
	
HALO	Presence (absence) of fluorine, chlorine, bromine, and iodine atoms, e.g., if SMILES = 'ClCC(=O)CCl
	

Step 1 Preparation of the list of attributes extracted from SMILES and from GAO (Tables 1 and 2).

Step 2 Calculation by the Monte Carlo method series of models with various values of the threshold ($T = 1, 2$, and 3) and different values of the number epochs (N). The preferable T^* and N^* are selected according to scheme represented by Fig. 1.

Step 3 Building up model with $T = T^*$ and $N = N^*$ and estimation of the predictive potential of the model with the validation set, i.e., with substances which are invisible during building up the model.

Thus, functions of the training set, the test set, and the validation set are considerably different ones. The substances from the training set are the basis to build up model. The substances from the test set are a tool to examine the "objectivity" of the model which is built up with the training set. In other words, it is evaluating "if the overtraining is absent". Finally, the external validation set is a tool to estimate the true predictive potential of model with data on substances which are invisible during building up the model using the above-mentioned parameters $T = T^*$ and $N = N^*$.

The user can calculate the DCW (T^*, N^*) and build up the model.

$$\text{Endpoint} = C_0 + C_1 * \text{DCW}\left(T^*, N^*\right). \quad (2)$$

The predictive potential of the model calculated with Eq. 2 should be validated with external validation set invisible during building up the model [12,15]. It is to be noted that similar nonlinear models as rule are considerable better for visible training set (i.e., for the system of training and test sets) but poorer for the external invisible validation set [16, 17].

The measure of the statistical prevalence of various molecular features (SA_k) which are extracted from SMILES and graph of atomic orbitals can be calculated as the following equation:

$$SA_k\text{Defect} = \begin{cases} \dfrac{|P_{TRN}(SA_k) - P_{TST}(SA_k)|}{N_{TRN}(SA_k) + N_{TST}(SA_k)}, & \text{if\_}N_{TST}(SA_k) > 0 \\ 1, \quad \text{otherwise} \end{cases} \quad (3)$$

where the $P_{TRN}(SA_k)$ is the probability of the presence of the SA_k in SMILES of the training set, i.e.,

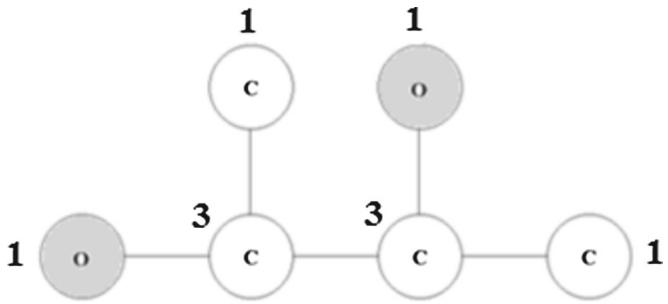
$$P_{TRN}(SA_k) = N_{TRN}(SA_k)/N_{TRN}.$$

The $P_{TRN}(SA_k)$ is the probability of the presence of the SA_k in SMILES of the test set, i.e.,

$$P_{TST}(SA_k) = N_{TST}(SA_k)/N_{TST}.$$

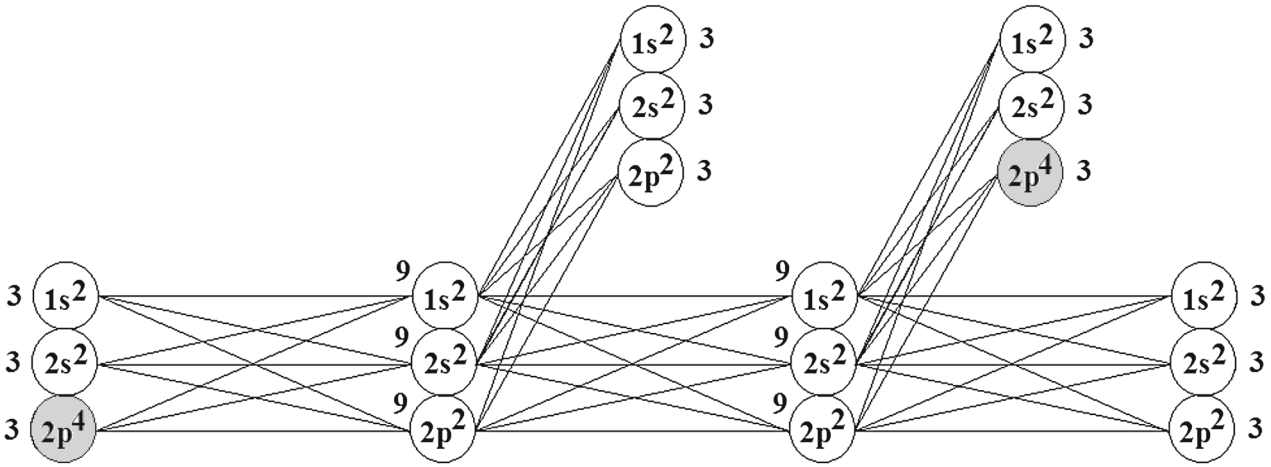
Table 2 Example of the representation of Acetoin (CAS 513-86-0; and SMILES="O=C(C)C(O)C") by means of (i) hydrogen-suppressed graph; and (ii) Graph of atomic orbitals

Hydrogen suppressed graph



	O_1	C_2	C_3	C_4	O_5	C_6	$EC0_k$
O_1	0	1	0	0	0	0	1
C_2	1	0	1	1	0	0	3
C_3	0	1	0	0	0	0	1
C_4	0	1	0	0	1	1	3
O_5	0	0	0	1	0	0	1
C_6	0	0	0	1	0	0	1

Graph of atomic orbitals (GAO)



	O_1			C_2			C_3			C_4			O_5			C_6			$EC0_k$
	$1s^2$	$2s^2$	$2p^4$	$1s^2$	$2s^2$	$2p^2$	$1s^2$	$2s^2$	$2p^2$	$1s^2$	$2s^2$	$2p^2$	$1s^2$	$2s^2$	$2p^4$	$1s^2$	$2s^2$	$2p^2$	
$1s^2$	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	3
$2s^2$	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	3
$2p^4$	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	3
$1s^2$	1	1	1	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	9
$2s^2$	1	1	1	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	9
$2p^2$	1	1	1	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	9
$1s^2$	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	3
$2s^2$	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	3
$2p^2$	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	3
$1s^2$	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	1	1	1	9
$2s^2$	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	1	1	1	9
$2p^2$	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	1	1	1	9

Table 2 continued

	O_1			C_2			C_3			C_4			O_5			C_6			$EC0_k$
	$1s^2$	$2s^2$	$2p^4$	$1s^2$	$2s^2$	$2p^2$	$1s^2$	$2s^2$	$2p^2$	$1s^2$	$2s^2$	$2p^2$	$1s^2$	$2s^2$	$2p^4$	$1s^2$	$2s^2$	$2p^2$	
$1s^2$	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	3
$2s^2$	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	3
$2p^4$	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	3
$1s^2$	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	3
$2s^2$	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	3
$2p^2$	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	3

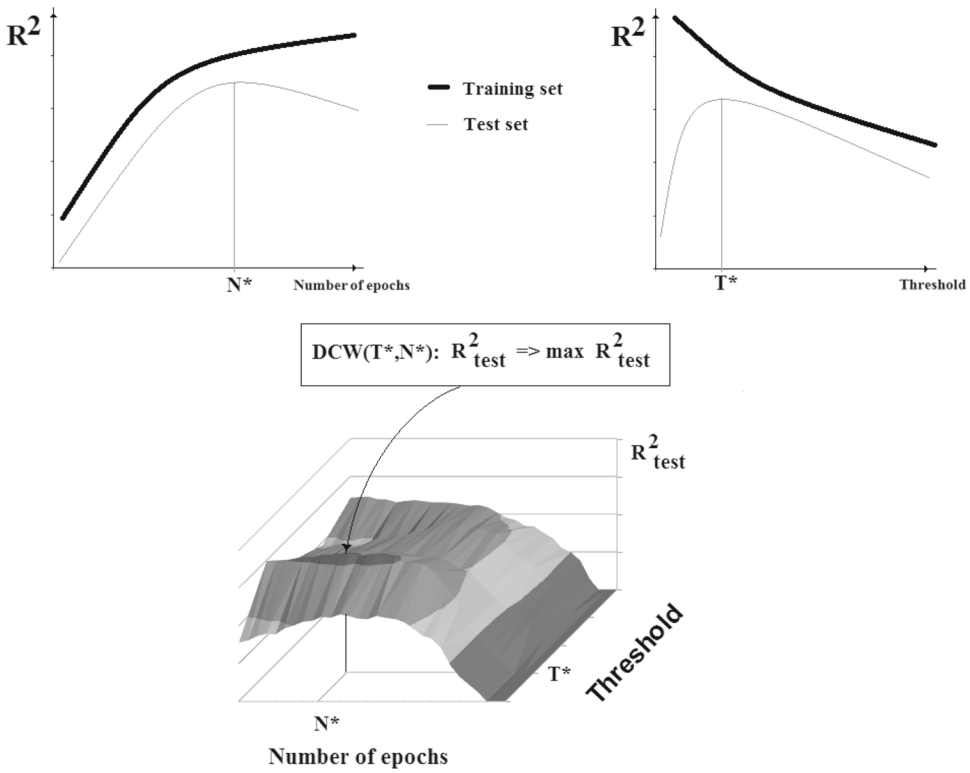


Fig. 1 Scheme of the definition of the preferable CORAL model. T is threshold; N is the number of epochs of the Monte Carlo optimization; T^* and N^* are values which give maximum for the correlation coefficient between experimental and calculated endpoint values for the test set

The N_{TRN} (SA_k) is the number (frequency) of SMILES which contain SA_k in the training set.

The N_{TRN} is the total number of SMILES in the training set.

The N_{TST} (SA_k) is the number (frequency) of SMILES which contain SA_k in the test set.

The N_{TST} is the total number of SMILES in the test set.

If the probability of SA_k in the training set is equal to the probability of SA_k in the test set, it is the ideal situation and the defect of the SA_k in this case should be estimated as minimal (zero). However, this situation is not typical, i.e., the difference between the probability of SA_k in the training set and the probability of SA_k in the test set, as rule, is not zero. Under such circumstances, the numbers of SA_k in the training set and in the test set also should be taken into account.

The small frequency or the absence of SA_k in the training set most probably will lead to decrease of statistical significance of the SA_k for the model. The absence (or even the small frequency) of SA_k in the test set together with significant prevalence of the SA_k in the training set will lead to overfitting: the improvement of the model for the training set due to the correlation weight of the SA_k will be accompanied by unpredictable influence of this correlation weight for the model within the test set. The Eq. 3 gives two criteria of expedient distribution into the training set and test set: (i) the difference between probabilities of attributes be in the training and be in the test sets should be as small as possible; and (ii) the numbers of attributes in the training set and test sets should be as large as possible. If the above-mentioned two conditions take place, the split into the training and test sets should be estimated as "satisfactory" one. Finally, the mole-

cular features which are absent in the test set cannot improve model, and their defect should be defined as maximum (i.e., unit).

Thus, the measure calculated with Eq. 3 can be used for the classification of the active (not blocked) attributes in accordance with their prevalence in the training and test set.

Having the numerical data on the defects of SA_k, one can compare reliability of the prediction for a substance, using the following criterion (DefectSMILES):

$$\text{DefectSMILES} = \sum_{\text{Active}SA_k} SA_k\text{defect} \tag{4}$$

The domain of applicability can be defined as follows: a substance is fall into the domain of applicability if its DefectSMILES obeys the condition:

$$\text{DefectSMILES} < 2 \times \overline{\text{DefectSMILES}}, \tag{5}$$

where $\overline{\text{DefectSMILES}}$ is average for visible set (training and test sets). Thus, the DefectSMILES gives possibility to define the domain of applicability for the models.

Using the summation of the Defect SMILES calculated with Eq. 4 one can define an integral characteristic of a split into the training and test sets:

$$\text{SplitDefect} = \sum_{\text{Training \& TestSets}} \text{DefectSMILES}. \tag{6}$$

The Split Defect can be a useful characteristic of the distribution into the training set and test set from heuristic point of view.

Results

The CORAL software gives for three above-mentioned splits the following models (the n is the number of compounds in a set, r^2 is determination coefficient, q^2 is cross-validated r^2, RMSE is root-mean-square error; F is Fischer F ratio):

Split 1

$$\begin{aligned}
&\text{lgNOAEL} = -2.1959849(\pm 0.0086409) \\
&\quad + 0.0675751(\pm 0.0005022)*\text{DCW}(1, 30) \\
&n = 97, \quad r^2 = 0.5312, \quad q^2 = 0.5158, \\
&\text{RMSE} = 0.617, \; F = 108 \text{ (training set)} \\
&n = 16, \quad r^2 = 0.6610, \quad \text{RMSE} = 0.444 \text{ (test set)} \\
&n = 27, \quad r^2 = 0.5832, \quad \text{RMSE} = 0.447 \text{ (validation set)} \\
&n = 26, \quad r^2 = 0.5749, \quad \text{RMSE} = 0.449 \\
&\quad \text{(validation set, domain of applicability)}
\end{aligned} \tag{7}$$

Split 2

$$\begin{aligned}
&\text{lgNOAEL} = -1.8713499 \; (\pm 0.0073147) \\
&\quad + 0.0633027 \; (\pm 0.0004966)*\text{DCW}(1, 30) \\
&n = 97, \quad r^2 = 0.5015, \quad q^2 = 0.4852, \\
&\text{RMSE} = 0.613, \quad F = 96 \text{ (training set)} \\
&n = 16, \quad r^2 = 0.6799, \quad \text{RMSE} = 0.524 \text{ (test set)} \\
&n = 27, \quad r^2 = 0.5843, \quad \text{RMSE} = 0.457 \text{ (validation set)} \\
&n = 25, \quad r^2 = 0.5890, \text{RMSE} = 0.453 \\
&\quad \text{(validation set, domain of applicability)}
\end{aligned} \tag{8}$$

Split 3

$$\begin{aligned}
&\text{lgNOAEL} = -2.1680835(\pm 0.0082917) \\
&\quad + 0.0737528(\pm 0.0005127)*\text{DCW}(1, 30) \\
&= 97, \quad r^2 = 0.5301, \quad q^2 = 0.5153, \\
&\text{RMSE} = 0.611, \quad \text{F} = 107 \text{ (training set)} \\
&n = 16, \quad r^2 = 0.7306, \quad \text{RMSE} = 0.494 \text{ (test set)} \\
&n = 27, \quad r^2 = 0.6049, \quad \text{RMSE} = 0.427 \text{ (validation set)} \\
&n = 26, \quad r^2 = 0.6143, \text{RMSE} = 0.425 \\
&\quad \text{(validation set, domain of applicability)}
\end{aligned} \tag{9}$$

Table 3 contains experimental and calculated lgNOAEL with Eqs. 7–9 for the training set and test set. The distributions of compounds into the training set and test set also are represented in Table 3. Table 4 contains experimental and calculated lgNOAEL for the external validation set.

Figure 2 contains the graphical representation of these models.

The additional analysis of twelve splits (including three represented by models which are calculated with Eqs. 7–9; these are the splits 1, 2, and 3; Table 5 contains the data) gives possibility to study the criterion calculated with Eq. 6. Figure 3 represents the correlation between Split Defect calculated with Eq. 6 and the determination coefficient between experimental and predicted lgNOAEL for the validation set (12 random splits). Figure 4 represents the correlation between Split Defect calculated with Eq. 6 and the root-mean-square error for the validation set (twelve random splits). Unexpectedly, the increase of the Split Defect is accompanied by increase of the determination coefficient and by decrease for root-mean-square error for the external validation set. Thus, very likely, these correlations can be useful criteria to compare different splits into the training set and test set.

Table 3 The distribution of available data into the training (+) and test (#) sets; experimental and calculated lgNOAEL values; and domain of applicability for suggested models

Random splits			Substances		lgNOAEL				Domain of applicability		
1	2	3	CAS	SMILES	Experiment	Eq. 6	Eq. 7	Eq. 8	1	2	3
+	+	+	513-86-0	O=C(C)C(O)C	−2.5190	−1.9727	−1.9115	−2.0152	Y	Y	Y
+	+	+	2432-99-7	O=C(O)CCCCCCCCCCN	−3.1760	−2.2346	−2.2466	−2.3286	Y	Y	Y
+	+	+	123-31-9	Oc1ccc(O)cc1	−1.3980	−1.7691	−1.7416	−1.6771	Y	Y	Y
+	+	+	94-26-8	O=C(OCCCC)c1ccc(O)cc1	−2.9540	−1.6171	−1.6004	−1.6094	Y	Y	Y
+	+	+	87-20-7	O=C(OCCC(C)C)c1ccccc1(O)	−0.6720	−1.5413	−1.5328	−1.5912	Y	Y	Y
+	#	#	534-73-6	OCC(O)C(O)C(O)C(O)CO C1OC(CO)C(O)C(O)C1(O)	−3.4660	−3.7482	−3.9690	−3.8429	Y	Y	Y
+	+	+	503-74-2	O=C(O)CC(C)C	−3.3330	−2.0023	−1.9394	−2.0377	Y	Y	Y
+	+	+	38502-29-3	OC(Cc1ccccc1)CC(C)C	−1.0000	−1.4098	−1.3657	−1.3763	Y	Y	Y
+	+	+	108-39-4	Oc1cccc(c1)C	−1.6990	−1.5155	−1.4522	−1.4417	Y	Y	Y
+	+	+	93-92-5	O=C(OC(c1ccccc1)C)C	−1.6990	−1.1564	−1.1245	−1.1917	Y	Y	Y
+	+	+	122-99-6	OCCOc1ccccc1	−1.9030	−1.4852	−1.4373	−1.3486	Y	Y	Y
+	+	+	698-87-3	OC(C)Cc1ccccc1	−1.0000	−1.4266	−1.3774	−1.3495	Y	Y	Y
#	+	+	142-19-8	O=C(OCC=C)CCCCCC	−1.6950	−1.9551	−1.8495	−1.9142	Y	Y	Y
+	+	+	431-03-8	O=C(C(=O)C)C	−1.9540	−1.9727	−1.9115	−2.0152	Y	Y	Y
+	+	+	78-59-1	O=C1C=C(C)CC(C)(C)C1	−2.6990	−1.7408	−1.6072	−1.7236	Y	Y	Y
#	#	+	110-43-0	O=C(C)CCCCC	−1.3010	−1.8836	−1.7734	−1.8657	Y	Y	Y
+	+	+	93-65-2	O=C(O)C(Oc1ccc(cc1C)Cl)C	−0.3980	−1.7192	−1.6708	−1.8684	Y	Y	Y
+	+	#	94-81-5	O=C(O)CCCOc1ccc(cc1C)Cl	−1.0790	−1.8542	−1.7942	−1.9318	Y	Y	Y
#	#	+	99911-45-2	O=C(NS(=O)(=O)O)CC(=O)C	−3.2700	−2.8728	−2.6868	−3.2731	Y	Y	N
+	+	+	1646-88-4	O=C(ON=CC(C)(C)S(=O)(=O)C)NC	0.2220	0.2244	0.2254	0.2230	N	N	N
+	+	+	834-12-8	n1c(nc(nc1NC(C)C)SC)NCC	−0.9340	−0.9468	−0.9410	−0.9367	N	N	N
+	+	+	108-60-1	O(C(C)CCl)C(C)CCl	−2.0000	−1.8711	−1.6220	−1.5866	Y	N	N
+	+	+	56-23-5	C(Cl)(Cl)(Cl)Cl	0.1490	0.1483	0.1495	0.1442	N	N	Y
+	+	+	609-20-1	Nc1cc(c(N)c(c1)Cl)Cl	−2.0000	−1.0139	−1.0008	−1.2012	Y	Y	Y
+	+	+	95-50-1	c1ccc(c(c1)Cl)Cl	−1.7780	−0.8650	−0.8626	−0.8850	Y	Y	Y
+	+	+	94-75-7	O=C(O)COc1ccc(cc1Cl)Cl	−0.0000	−1.5459	−1.5050	−1.6059	Y	Y	Y
+	#	+	2164-17-2	O=C(Nc1cccc(c1)C(F)(F)F)N(C)C	−0.9030	−0.9052	−1.4701	−0.8976	N	Y	N
+	#	+	87-68-3	C(C(=C(Cl)Cl)Cl)(=C(Cl)Cl)Cl	−0.0000	−0.2582	−0.3141	−0.3019	N	Y	N
+	+	+	108-78-1	n1c(nc(nc1N)N)N	−2.2300	−2.1459	−2.1641	−2.1103	N	Y	Y
+	+	#	85-91-6	O=C(OC)c1ccccc1(NC)	−1.1760	−1.0047	−1.0141	−1.1216	Y	Y	Y
+	+	+	150-68-5	O=C(Nc1ccc(cc1)Cl)N(C)C	−1.8750	−0.7268	−0.7058	−0.8508	Y	Y	Y
+	#	+	76-01-7	C(C(Cl)(Cl)Cl)(Cl)Cl	−2.0970	−1.5918	−1.4090	−1.5649	N	Y	Y
+	#	#	108-45-2	Nc1cccc(N)c1	−0.7780	−0.9538	−0.9772	−1.0405	Y	Y	Y
+	+	+	23950-58-5	O=C(NC(C#C)(C)C)c1cc(cc(c1)Cl)Cl	−0.3980	−0.3970	−0.3929	−0.3984	Y	N	Y
+	+	+	95-94-3	c1c(c(cc(c1Cl)Cl)Cl)Cl	0.4690	−0.5110	−0.5222	−0.5616	Y	Y	Y
#	+	+	58-90-2	Oc1c(cc(c(c1Cl)Cl)Cl)Cl	−1.3980	−1.1494	−1.0723	−1.1600	Y	Y	Y
+	#	+	95-95-4	Oc1cc(c(cc1Cl)Cl)Cl	−2.0000	−1.3264	−1.2425	−1.3217	Y	Y	Y
+	#	+	5989-27-5	C=C(C)C1CC=C(C)CC1	−2.1760	−1.3180	−1.2706	−1.3956	Y	Y	Y
++	+	+	513-37-1	C(=C(C)C)Cl	−2.0970	−1.5291	−1.4164	−1.5086	Y	N	Y
+	+	+	98-85-1	OC(c1ccccc1)C	−1.9680	−1.3970	−1.3495	−1.3270	Y	Y	Y
+	+	+	6731-36-8	O(OC1(OOC(C)(C)C)(CC(C) CC(C)(C)C1))C(C)(C)C	−1.5230	−1.1612	−1.0179	−1.1545	Y	Y	Y
#	+	+	112-26-5	O(CCOCCCl)CCCl	−1.2220	−2.0647	−1.8052	−1.6490	Y	N	N
+	+	+	102-47-6	c1cc(c(cc1CCl)Cl)Cl	−0.5220	−0.8529	−0.8034	−0.7324	Y	N	Y

Table 3 continued

Random splits			Substances		lgNOAEL				Domain of applicability		
1	2	3	CAS	SMILES	Experiment	Eq. 6	Eq. 7	Eq. 8	1	2	3
+	+	#	526-73-8	c1cc(c(c(c1)C)C)C	−1.0000	−1.0215	−1.0283	−1.0815	Y	Y	Y
+	+	+	3319-31-1	O=C(OCC(CC)CCCC)c1ccc(C(=O)OCC(CC)CCCC)c(c1)C(=O)OCC(CC)CCCC	−2.5230	−2.0315	−2.0743	−2.2175	Y	Y	Y
+	+	+	95-63-6	c1cc(c(cc1C)C)C	−1.5230	−1.0215	−1.0283	−1.0815	Y	Y	Y
+	+	+	88-44-8	O=S(=O)(O)c1cc(ccc1(N))C	−2.0000	−2.1248	−2.1279	−2.1226	Y	Y	N
+	+	+	149-57-5	O=C(O)C(CC)CCCC	−1.7850	−2.0909	−2.0232	−2.1053	Y	Y	Y
+	+	#	79-39-0	O=C(N)C(=C)C	−1.0000	−1.6393	−1.6383	−1.8721	Y	Y	Y
+	+	+	88-18-6	Oc1ccccc1C(C)(C)C	−0.8240	−1.3553	−1.2901	−1.3173	Y	Y	Y
+	+	+	118-75-2	O=C1C(=C(C(=O)C(=C1Cl)Cl)Cl)Cl	−1.0000	−1.2981	−1.2472	−1.3940	Y	Y	Y
+	+	#	96-76-4	Oc1ccc(cc1C(C)(C)C)C(C)(C)C	−0.8240	−1.2672	−1.1911	−1.3119	Y	Y	Y
+	#	#	123-63-7	O1C(OC(OC1C)C)C	−1.5230	−1.6382	−1.5109	−1.6124	Y	Y	Y
+	+	+	4130-42-1	Oc1c(cc(cc1C(C)(C)C)CC)C(C)(C)C	−0.6990	−1.3690	−1.2822	−1.4535	Y	Y	Y
+	+	+	828-00-2	O=C(OC1OC(OC(C)C1)C)C	−1.4910	−1.6978	−1.6110	−1.7541	Y	Y	Y
+	+	+	103-44-6	O(C=C)CC(CC)CCCC	−0.4270	−1.6720	−1.5322	−1.6563	Y	Y	Y
+	+	+	121-47-1	O=S(=O)(O)c1cccc(N)c1	−2.0000	−2.0526	−2.0648	−2.0036	Y	Y	N
+	+	#	620-17-7	Oc1cccc(c1)CC	−2.0000	−1.5451	−1.4802	−1.4643	Y	Y	Y
+	+	+	4435-53-4	O=C(OCCC(OC)C)C	−2.5230	−1.7439	−1.6625	−1.7866	Y	Y	Y
+	+	+	111-17-1	O=C(O)CCSCCC(=O)O	−1.8240	−1.8177	−1.8179	−1.8154	N	N	N
+	+	#	108-69-0	Nc1cc(cc(c1)C)C	−0.5220	−0.6922	−0.6644	−0.8002	Y	Y	Y
#	+	+	140-66-9	Oc1ccc(cc1)C(C)(C)CC(C)(C)C	−0.6990	−1.2246	−1.1560	−1.2154	Y	Y	Y
+	+	+	121-60-8	O=C(Nc1ccc(cc1)S(=O)(=O)Cl)C	−1.8240	−1.8293	−1.8238	−1.8274	N	N	N
+	+	+	1570-64-5	Oc1ccc(cc1C)Cl	−1.3010	−1.7526	−1.6460	−1.7641	Y	Y	Y
+	+	+	20265-96-7	Nc1ccc(cc1)Cl	−0.6990	−0.7848	−0.7320	−0.8844	Y	Y	Y
#	+	+	123-07-9	Oc1ccc(cc1)CC	−1.5230	−1.5451	−1.4802	−1.4643	Y	Y	Y
+	+	+	137-09-7	Oc1ccc(N)cc1(N)	−1.0790	−1.8417	−1.9075	−2.0125	N	Y	Y
#	+	#	4286–23-1	Oc1ccc(cc1)C(=C)C	−1.0000	−1.2921	−1.2350	−1.3256	Y	Y	Y
+	+	+	99-71-8	Oc1ccc(cc1)C(C)CC	−1.5230	−1.4988	−1.4405	−1.4686	Y	Y	Y
+	+	+	98-54-4	Oc1ccc(cc1)C(C)(C)C	−1.7780	−1.3553	−1.2901	−1.3173	Y	Y	Y
+	+	+	100-40-3	C=CC1CC=CCC1	−0.3010	−1.4697	−1.4057	−1.4321	Y	Y	Y
+	+	+	88-53-9	O=S(=O)(O)c1cc(c(cc1(N))C)Cl	−2.5230	−2.3619	−2.3217	−2.4450	Y	Y	N
+	+	+	13718-94-0	O=C(CO)C(O)C(O)C(O)COC1OC(CO)C(O)C(O)C1(O)	−3.8450	−3.5711	−3.7914	−3.7225	Y	Y	Y
+	+	+	141-17-3	O=C(OCCOCCOCCCC)CCCCC(=O)OCCOCCOCCCC	−2.0000	−2.4014	−2.3677	−2.3061	Y	Y	Y
+	+	+	591-87-7	O=C(OCC=C)C	−0.7780	−1.8074	−1.7098	−1.8015	Y	Y	Y
+	+	+	156-43-4	O(c1ccc(N)cc1)CC	−0.5220	−1.1944	−1.1993	−1.3021	Y	Y	Y
+	+	+	103-69-5	c1ccc(cc1)NCC	0.4770	−0.3870	−0.3367	−0.3075	Y	Y	Y
+	+	+	1825-21-4	O(c1c(c(c(c(c1Cl)Cl)Cl)Cl)Cl)C	−1.6020	−0.7017	−0.6050	−0.7439	Y	Y	Y
#	#	+	99-94-5	O=C(O)c1ccc(cc1)C	−1.5230	−1.5457	−1.5244	−1.5609	Y	Y	Y
+	+	+	100-47-0	N#Cc1ccccc1	−0.8240	−0.3914	−0.3964	−0.4232	Y	N	Y
+	+	+	140-11-4	O=C(OCc1ccccc1)C	−2.4950	−1.2322	−1.1921	−1.2099	Y	Y	Y
#	+	+	141-02-6	O=C(OCC(CC)CCCC)C=CC(=O)OCC(CC)CCCC	−2.0000	−2.2003	−2.1518	−2.2560	Y	Y	Y
#	+	#	127-90-2	O(CC(C(Cl)(Cl)Cl)Cl)CC(C(Cl)(Cl)Cl)Cl	−1.1250	−1.9210	−1.4803	−1.8715	Y	Y	Y
+	+	+	134-72-5	OC(c1ccccc1)C(NC)C	−0.7960	−1.0805	−1.0967	−1.1465	Y	Y	Y

Table 3 continued

Random splits			Substances		lgNOAEL				Domain of applicability		
1	2	3	CAS	SMILES	Experiment	Eq. 6	Eq. 7	Eq. 8	1	2	3
#	+	+	56539-66-3	OCCC(OC)(C)C	−1.3010	−1.8706	−1.7650	−1.7932	Y	Y	Y
+	+	#	78-44-4	O=C(OCC(C)(COC(=O)NC(C)C)CCC)N	−2.0000	−1.9136	−1.9687	−2.1215	Y	Y	Y
#	#	+	131-17-9	O=C(OCC=C)c1ccccc1(C(=O)OCC=C)	−1.3980	−1.5832	−1.5809	−1.6171	Y	Y	Y
+	+	+	3648-21-3	O=C(OCCCCCCC)c1ccccc1(C(=O)OCCCCCCC)	−1.3190	−1.8195	−1.8043	−1.7975	Y	Y	Y
+	+	+	205687-03-2	O=C(OCc1ccc(O)c(OC)c1)CCCCCCC(C)C	−2.4770	−1.7440	−1.7278	−1.8038	Y	Y	Y
+	+	+	71-55-6	CC(Cl)(Cl)Cl	−2.4620	−1.9971	−1.7681	−2.0254	Y	Y	Y
+	+	+	79-34-5	C(C(Cl)Cl)(Cl)Cl	−1.0040	−1.0263	−1.0376	−0.9753	N	Y	N
+	+	+	544-76-3	CCCCCCCCCCCCCCCC	−1.1250	−1.8362	−1.7829	−1.7280	Y	Y	Y
+	+	#	461-72-3	O=C1NC(=O)CN1	−2.0000	−1.9688	−1.9362	−1.9359	Y	Y	Y
+	+	+	100-74-3	O1CCN(CC)CC1	−1.2220	−1.0855	−0.9785	−0.9901	Y	Y	Y
#	+	+	100-61-8	c1ccc(cc1)NC	−0.2230	−0.3574	−0.3088	−0.2850	Y	Y	Y
+	+	+	100-54-9	N#Cc1cnccc1	−0.2230	−0.6535	−0.6457	−0.6203	Y	N	Y
+	+	+	4390-04-9	CC(CC(C)(C)C)CC(C)(C)CC(C)(C)C	−1.5230	−0.9842	−0.9498	−1.1109	Y	Y	Y
+	#	+	95-51-2	Nc1ccccc1Cl	−1.0000	−0.7848	−0.7320	−0.8844	Y	Y	Y
+	+	+	110-30-5	O=C(NCCNC(=O)CCCCCCCCCCCCCCC)CCCCCCCCCCCCCC	−2.5230	−2.8504	−2.7942	−2.6794	Y	Y	Y
+	+	+	629-62-9	CCCCCCCCCCCCCCCC	−2.5230	−1.8067	−1.7550	−1.7055	Y	Y	Y
+	#	+	67-72-1	C(C(Cl)(Cl)Cl)(Cl)(Cl)Cl	−1.6720	−2.1574	−1.7804	−2.1544	Y	Y	Y
+	+	+	106-48-9	Oc1ccc(cc1)Cl	−1.5230	−1.6804	−1.5829	−1.6450	Y	Y	Y
+	+	+	61-76-7	Oc1cccc(c1)C(O)CNC	−2.0970	−1.4821	−1.5167	−1.5191	Y	Y	Y
#	+	#	108-73-6	Oc1cc(O)cc(O)c1	−2.0000	−2.0949	−2.0941	−2.0315	Y	Y	Y
+	+	+	9016-45-9	OCCOCCOc1ccc(cc1)CCCCCCCCC	−2.5230	−1.8356	−1.7719	−1.6740	Y	Y	Y
+	+	+	51-52-5	O=C1C=C(NC(N1)=S)CCC	−0.0000	−0.0003	−0.0016	0.0016	Y	N	Y
+	#	+	657-84-1	O=S(=O)(O)c1ccc(cc1)C	−2.0000	−1.7612	−1.2099	−1.8467	Y	Y	N
+	+	#	585-07-9	O=C(OC(C)(C)C)C(=C)C	−0.8240	−1.5418	−1.4523	−1.6362	Y	Y	Y
#	#	+	614-45-9	O=C(OOC(C)(C)C)c1ccccc1	−1.4770	−1.0253	−0.9943	−1.0437	Y	Y	Y
+	+	+	126-33-0	O=S1(=O)(CCCC1)	−1.3010	−1.5379	−1.3014	−1.4755	Y	N	N
+	+	+	98-51-1	c1cc(ccc1C)C(C)(C)C	0.3010	−0.7890	−0.8031	−0.8379	Y	Y	Y
+	+	+	1025-15-6	O=C1N(C(=O)N(C(=O)N1CC=C)CC=C)CC=C	−0.2230	−0.6570	−0.6924	−0.6435	N	Y	Y
+	+	+	598-77-6	CC(C(Cl)Cl)Cl	−1.1760	−1.1360	−1.1238	−1.1594	N	Y	Y

Discussion

Sakuratani et al. [2] developed a read-across approach to predict LOAEL within repeated dose toxicity (RDT) using toxicological grouping categories. They defined 33 chemical categories to be used for the gap filling based on RDT data. Mazzatorta et al. [10] developed a QSAR model for predicting LOAEL using chronic data (exposure longer than 180 days) in the rat. The model gave $R^2 = 0.54$ and RMSE = 0.7. However, the limit of this model is the absence of validation with an external dataset and, thus, the lack of a vital point for assessing the real predictive power of the model. The same group also calculated an experimental variability of 0.64 (logarithmic scale) for LOAEL used in their dataset. A model for the NOAEL suggested in the literature (the model is built up with involving various physicochemical descrip-

Table 4 The experimental and calculated lgNOAEL values and domain of applicability for validation set

Substances		lgNOAEL				Domain of applicability for splits 1, 2, and 3		
CAS	SMILES	Expr	Eq. 7	Eq. 8	Eq. 9	1	2	3
5471-51-2	O=C(C)CCc1ccc(O)cc1	−2.0000	−1.6047	−1.5802	−1.6060	Y	Y	Y
108-46-3	Oc1cccc(O)c1	−1.5050	−1.7691	−1.7416	−1.6771	Y	Y	Y
5977-14-0	O=C(N)CC(=O)C	−2.2480	−1.9224	−1.9556	−2.1300	Y	Y	Y
2835-39-4	O=C(OCC=C)CC(C)C	−1.4910	−1.7906	−1.6981	−1.8283	Y	Y	Y
105-60-2	O=C1NCCCCC1	−1.6990	−1.7362	−1.6516	−1.7041	Y	Y	Y
108-90-7	c1ccc(cc1)Cl	−1.6990	−1.0420	−1.0328	−1.0466	Y	Y	Y
77-47-4	C=1(C(=C(C(C=1Cl)(Cl)Cl)Cl)Cl)Cl	−1.0000	−0.5285	−0.3132	−0.5233	Y	Y	Y
108-31-6	O=C1OC(=O)C=C1	−1.6020	−2.1079	−2.0230	−2.0417	Y	Y	Y
924-42-5	O=C(C=C)NCO	−1.0970	−1.8079	−1.7938	−1.8712	Y	Y	Y
1918-16-7	O=C(N(c1ccccc1)C(C)C)CCl	−1.1240	−0.6298	−0.5675	−0.6628	Y	N	N
122-42-9	O=C(OC(C)C)Nc1ccccc1	−1.6990	−0.8862	−0.9113	−1.0069	Y	Y	Y
630-20-6	C(C(Cl)(Cl)Cl)Cl	−2.0000	−1.9128	−1.6458	−1.7538	Y	N	Y
1948-33-0	Oc1ccc(O)c(c1)C(C)(C)C	−2.0970	−1.6810	−1.6426	−1.6717	Y	Y	Y
118-91-2	O=C(O)c1ccccc1Cl	−2.5230	−1.7105	−1.6550	−1.7642	Y	Y	Y
95-57-8	Oc1ccccc1Cl	−1.1250	−1.6804	−1.5829	−1.6450	Y	Y	Y
102-81-8	OCCN(CCCC)CCCC	−0.9210	−1.3516	−1.2960	−1.3116	Y	Y	Y
100-69-6	n1ccccc1C=C	−0.6200	−0.9915	−1.0017	−0.9426	Y	Y	Y
87-59-2	Nc1cccc(c1C)C	−0.6020	−0.6922	−0.6644	−0.8002	Y	Y	Y
2416-94-6	Oc1c(ccc(c1C)C)C	−1.5230	−1.6599	−1.5784	−1.6799	Y	Y	Y
108-42-9	Nc1cccc(c1)Cl	−1.0000	−0.7848	−0.7320	−0.8844	Y	Y	Y
95-64-7	Nc1ccc(c(c1)C)C	−0.5220	−0.6922	−0.6644	−0.8002	Y	Y	Y
87-62-7	Nc1c(cccc1C)C	−1.0000	−0.6922	−0.6644	−0.8002	Y	Y	Y
626-17-5	N#Cc1cccc(C#N)c1	−0.4270	−0.8270	−0.8634	−0.9242	N	N	Y
50-81-7	O=C1OC(C(O)=C1(O))C(O)CO	−3.0970	−2.7001	−2.7443	−2.7380	Y	Y	Y
1477-55-0	NCc1cccc(c1)CN	−1.6990	−1.0129	−1.0331	−1.0856	Y	Y	Y
608-93-5	c1c(c(c(c(c1Cl)Cl)Cl)Cl)Cl	−0.3420	−0.3340	−0.3520	−0.3999	Y	Y	Y
87-86-5	Oc1c(c(c(c(c1Cl)Cl)Cl)Cl)Cl	−1.0000	−0.9724	−0.9021	−0.9984	Y	Y	Y

tors) is characterized by $n = 218$, $r^2 = 0.35$, and $q^2 = 0.21$ [18].

Although the correlation between molecular structure and the NOAEL takes place, there is a considerable percentage of other factors that can influence this endpoint [8,10,18–20].

Thus, the suggested approach gives quantitative models of the NOAEL for three random splits into the training set, the test set, and the validation sets, and the predictive potential of these models are comparable with the predictive potential of models for the NOAEL [21,22] described in the literature [18].

Instead of the NOAEL/LOAEL approach, the Benchmark Dose Methodology (BDM) [23] can be used. However, the BDM is more expensive approach. Besides, in some cases, the BDM cannot be used to estimate toxicological behavior of substances. Taking this into account, the NOAEL/LOAEL approach should be estimated as a useful alternative for the BDM.

Finally, we deem that the principle "a QSAR is a random event" can be useful from regulatory point of view. In other words, the reliability of a QSAR approach for any endpoint in general, and for NOAEL in particular, should be validated with a group of different splits into the visible training set and invisible validation set [15]. The use of the approach (analysis of a group of distribution into the training set, test set, and external validation set, i.e., not only one split) in the case of the lgNOAEL numerical data for other set of organic compounds [18] gave the models which are statistically characterized by [24]: $\overline{n} \approx 174$, $\overline{r^2} \approx 0.70$, $\overline{s} \approx 0.41$(training set), and $\overline{n} \approx 21$, $\overline{r^2} \approx 0.64$, $\overline{s} \approx 0.39$ (test set). These models [24] are based on the representation of the molecular structure by

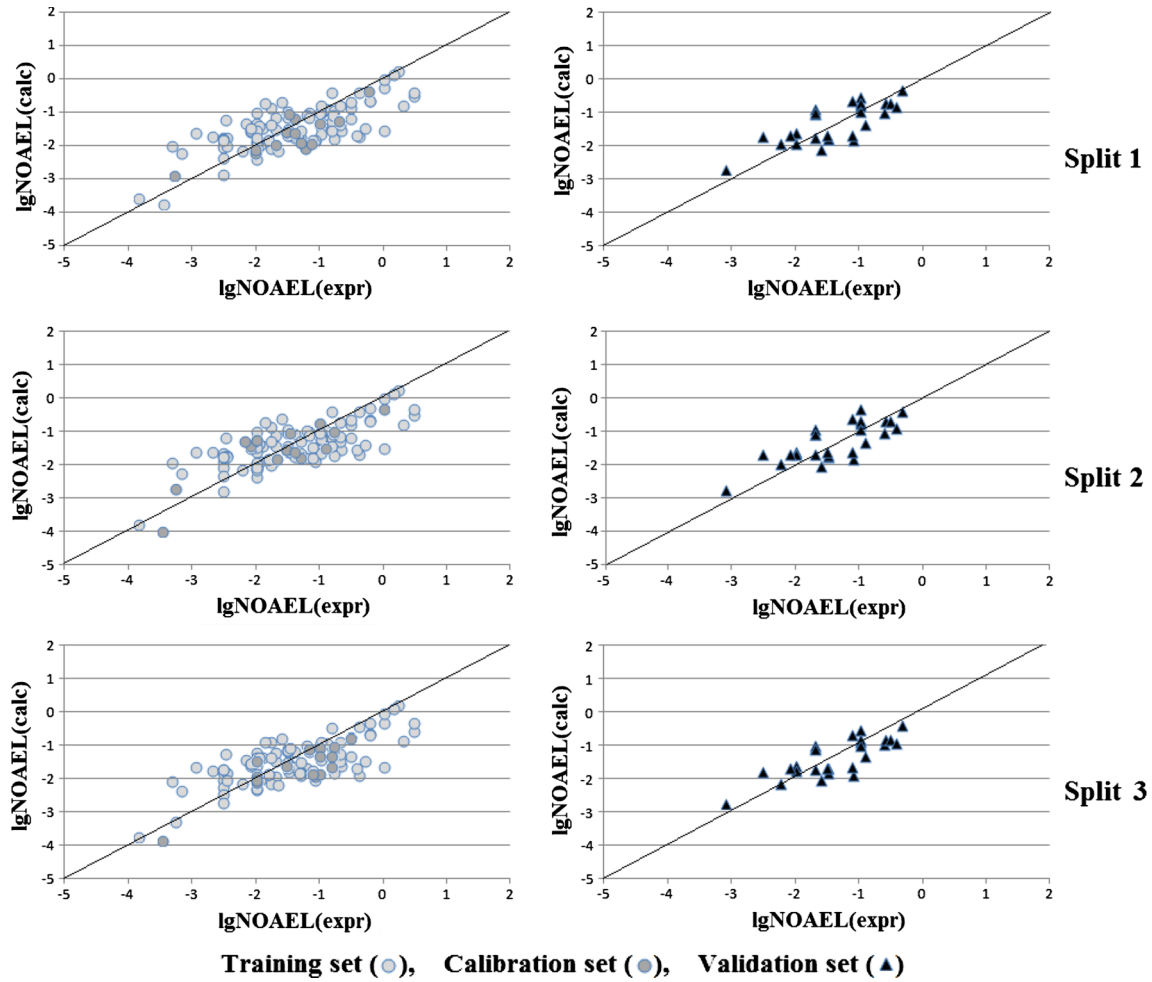


Fig. 2 Experimental (expr) and calculated (calc) lgNOAEL values for three random splits

Table 5 Correlation between split defect calculated with Eq. 6 and the predictive potential of the models

Split	Split defect	$r^2_{validation}$	$RMSE_{validation}$
1	30.1	0.5832	0.4473
2	21.3	0.5843	0.4571
3	38.1	0.6049	0.4272
4	18.2	0.5188	0.4896
5	40.1	0.6300	0.4424
6	25.2	0.5672	0.4507
7	26.1	0.6210	0.4424
8	24.2	0.5224	0.4859
9	22.3	0.6006	0.4500
10	25.2	0.5800	0.4400
11	16.4	0.5676	0.4800
12	16.3	0.5156	0.5334

solely SMILES (without data on the molecular graph). In fact, compounds examined here are characterized by a bigger variety of the molecular structure; therefore, QSAR models for these substances which are based on solely SMILES are characterized by a poor statistical quality. Fortunately, the hybrid approach (SMILES together with the molecular graph) gives the satisfactory statistical quality of the models for these very varied compounds calculated with Eqs. 7–9.

The suggested criteria for the estimation of the defect for the individual SMILES and GAO (Defect SMILES, Eq. 4) and for the estimation of the distribution into the training set and test set (Split Defect, Eq. 6) can be a convenient tool for the QSAR analysis. The Defect SMILES gives possibility to detect "suspected" compounds, thus this criterion is a tool to define the domain of applicability. The Split Defect gives possibility to compare different distribution into the training set and test set and to select preferable distribution from point of view of robustness of a QSAR. The disadvantage of these criteria is their dependence upon the distribution of available data into the training set and test set.

The Supplementary materials section (Table S1) contains the numerical data on the correlation weights for SA_k calculated with three different splits into the training and test sets.

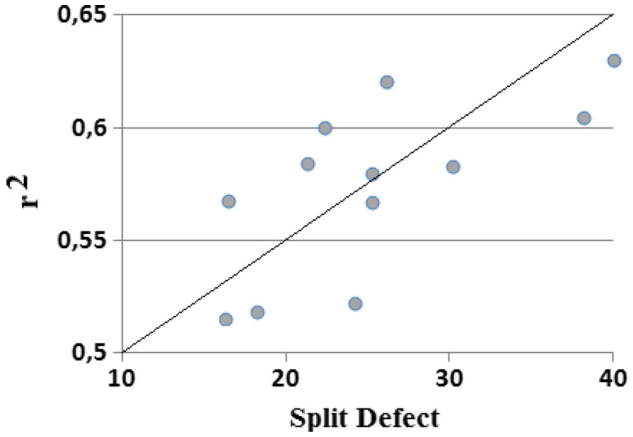


Fig. 3 Correlations between the Split Defect and the determination coefficient (r^2) for the external validation set (for 12 random splits)

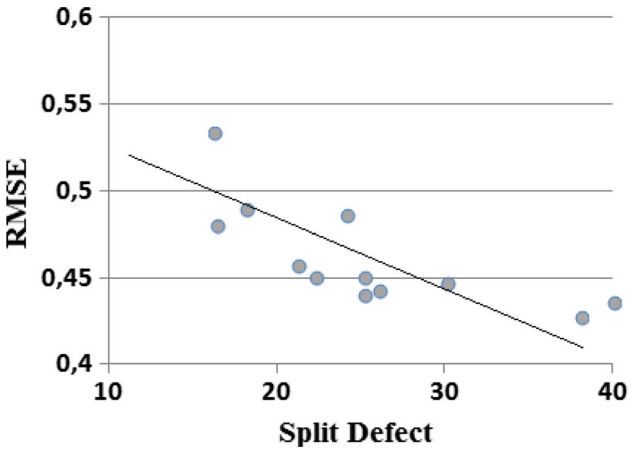


Fig. 4 Correlations between the Split Defect and the root-mean-square error (RMSE) for the external validation set (for 12 random splits)

There are **a** group of SA_k which are stable promoters of the lgNOAEL increase (i.e., SA_k characterized by (i) significant frequency in the training set and (ii) stable positive values of correlation weights) and group of stable promoters of the lgNOAEL decrease (i.e., SA_k characterized by (i) significant frequency in the training set and (ii) stable negative values of correlation weights). The above described SA_k defects are also represented in Table S1 for split 1, 2, and 3. Table S2 contains the numerical data on the correlation weights of SMILES attributes and GAO invariants for twelve random splits into the training set and the test set. Table S3 contains an example of the DCW (1, 30) calculation for a substance represented by the SMILES and GAO (acetoin, CAS 513-86-0). Table S4 contains the list of compounds that were selected as the external validation set. Table S5 contains twelve random splits into the training and test set which are examined in this work.

Thus, the suggested models have (i) definition of the domain of applicability (Tables 3, 4); (ii) the mechanistic interpretation in terms of the promoters of increase / decrease for lgNOAEL; and (iii) unambiguous algorithm to build up model. Consequently, described models are built up in accordance with OECD principles [21,22].

Conclusions

The NOAEL can be modeled by the Monte Carlo technique using SMILES and graph of atomic orbitals for the representation of the molecular structure. The statistical quality of models for the NOAEL calculated with the 2D descriptors is comparable with the statistical quality of models based on the 3D representation of the molecular structure with additional input of the physicochemical data. There are correlations between predictive potential of the models and the Split Defect calculated with Eq. 6 (Table 5). It should be noted that the described approach based on 2D descriptors can be used to build up predictive models for the cases of other complex endpoints, i.e., endpoints related to nanomaterials [25,26] and endpoints related to peptides [27].

Acknowledgments The authors are grateful for the contribution of the project HEALTH-F5-2010-267042 ToxBank (Supporting Integrated Data Analysis and servicing of Alternative Testing Methods in Toxicology) funded by European Commission and Cosmetics Europe under the Seventh Framework programme and the EU project PROSIL funded under the LIFE program (Project LIFE12 ENV/IT/000154).

Conflict of interest All the authors declare that there are no conflicts of interest.

References

1. SCCS -Scientific Committee on Consumer Safety (2012) The SCCS's notes of guidance for the testing of cosmetics substances and their safety evaluation 8th revision. http://ec.europa.eu/health/scientific_committees/consumer_safety/docs/sccs_s_006.pdf. Accessed April 2014
2. Sakuratani Y, Zhang HQ, Nishikawa S, Yamazaki K, Yamada T, Yamada J, Gerova K, Chankov G, Mekenyan O, Hayashi M (2013) Research hazard evaluation support system (HESS) for predicting repeated dose toxicity using toxicological categories. SAR QSAR Environ Res 24:351–363. doi:10.1080/1062936X.2013.773375
3. ECHA (2011) Comments and contributions received from MSC participants following the UK proposal and analysis concerning 'Possibilities for waiving repeat dose studies for low-toxicity substances (ECHA/MSC-16/2011/002)'. ECHA/MSC-18/2011/001 (6 May 2011)
4. REACH (2011) REACH: registration, evaluation, authorisation and restriction of chemicals (REACH) Regulation (EU) No 253/2011 of the European Parliament and of the Council of 15 March
5. Pauwels M, Rogiers V (2010) Human health safety evaluation of cosmetics in the EU: a legally imposed challenge to science. Toxicol Appl Pharm 243:260–274. doi:10.1016/j.taap.2009.12.007
6. Russell WMS, Burch RL (1959) The principles of humane experimental technique. Methuen, London

7. Lilienblum W, Dekant W, Foth H, Gebel T, Hengstler JG, Kahl R, Kramer P-J, Schweinfurth H, Wollin K-M (2008) Alternative methods to safety studies in experimental animals: role in the risk assessment of chemicals under the new European Chemicals Legislation (REACH). Arch Toxicol 82:211–236. doi:10.1007/s00204-008-0279-9
8. Pery A, Henegar A, Mombelli E (2009) Maximum-likelihood estimation of predictive uncertainty in probabilistic QSAR modeling. QSAR Comb Sci 3:338–344. doi:10.1002/qsar.200860116
9. Matthews EJ, Kruhlak NL, Benz RD, Contera JF (2004) Assessment of the health effects of chemical in humans: I. QSAR estimation of the maximum recommended therapeutic dose (MRTD) and no effect level (NOEL) of organic chemicals based on clinical trial data. Curr Drug Discov Technol 1:61–76. doi:10.2174/1570163043484789
10. Mazzatorta P, Dominguez Estevez M, Coulet M, Schilter B (2008) Modeling oral rat chronic toxicity. J Chem Inf Model 48:1949–1954. doi:10.1021/ci8001974
11. Taylor K, Andrew DJ, Rego L (2014) The added value of the 90-day repeated dose oral toxicity test for industrial chemicals with a low (sub)acute toxicity profile in a high quality dataset. Regul Toxicol Pharm 69:320–332. doi:10.1016/j.yrtph.2014.04.008
12. Toropova AP, Toropov AA, Benfenati E, Gini G, Leszczynska D, Leszczynski J (2011) CORAL: quantitative structure-activity relationship models for estimating toxicity of organic compounds in rats. J Comput Chem 32:2727–2733. doi:10.1002/jcc.21848
13. Toropov AA, Toropova AP (2003) QSPR modeling of alkanes properties based on graph of atomic orbitals. J Mol Struct THEOCHEM 637:1–10. doi:10.1016/S0166-1280(02)00492-X
14. Toropov AA, Toropova A, Martyanov SE, Benfenati E, Gini G, Leszczynska D, Leszczynski J (2011) Comparison of SMILES and molecular graphs as the representation of the molecular structure for QSAR analysis for mutagenic potential of polyaromatic amines. Chemometr Intell Lab Syst 109:94–100. doi:10.1016/j.chemolab.2011.07.008
15. Toropov AA, Toropova AP, Puzyn T, Benfenati E, Gini G, Leszczynska D, Leszczynski J (2013) QSAR as a random event: modeling of nanoparticles uptake in PaCa2 cancer cells. Chemosphere 92:31–37. doi:10.1016/j.chemosphere.2013.03.012
16. Peruzzo PJ, Marino DJG, Castro EA, Toropov AA (2001) Calculation of pK values of flavylium salts from the optimization of correlation weights of local graph invariants. J Mol Struct THEOCHEM 572:53–60. doi:10.1016/S0166-1280(01)00559-0
17. Krenkel G, Castro EA, Toropov AA (2001) Improved molecular descriptors to calculate boiling points based on the optimization of correlation weights of local graph invariants. J Mol Struct THEOCHEM 542:107–113. doi:10.1016/S0166-1280(00)00822-8
18. Goto T (2013) QSAR modeling using a set of intermediate-duration oral NOELs, Master's Thesis. https://etd.library.emory.edu/view/record/pid/emory:d724n
19. Filipson FA, Sand S, Nilsson J, Victorin K (2003) The benchmark dose method—review of available models, and recommendations for application in health assessment. Crit Rev Toxicol 33:505–542. doi:10.1080/10408440390242360
20. Kalberlah F, Schneider K, Schuhmacher-Wolz U (2003) Uncertainty in toxicological risk assessment for non-carcinogenic health effects. Regul Toxicol Pharm 37:92–104. doi:10.1016/S0273-2300(02)00032-6
21. OECD (2005) Guidance document on the validation and international acceptance of new or updated test methods for hazard assessment. OECD series on testing and assessment number 34, OECD, Paris (25 August 2005)
22. OECD (2008) Repeated dose 28-day oral toxicity study in rodents, Test Guideline No. 407. OECD Guidelines for the testing of chemicals. OECD, Paris (Adopted: 3 October 2008)
23. EPA (2014) US Environmental Protection Agency http://www.epa.gov/ncea/bmds/bmds_training/methodology/intro.htm
24. Toropova AP, Toropova AA, Veselinović JB, Veselinović AM (2015) QSAR as a random event: a case of NOAEL. Environ Sci Pollut 11:1–8. doi:10.1007/s11356-014-3977-2 (Published online December 19, 2014)
25. Toropova AP, Toropov AA (2014) Optimal descriptor as a translator of eclectic information into the prediction of membrane damage by means of various TiO2 nanoparticles. Chemosphere 93:2650–2655. doi:10.1016/j.chemosphere.2013.09.089
26. Toropov AA, Toropova AP (2014) Optimal descriptor as a translator of eclectic data into endpoint prediction: mutagenicity of fullerene as a mathematical function of conditions. Chemosphere 104:262–264. doi:10.1016/j.chemosphere.2013.10.079
27. Toropov AA, Toropova AP, Raska I Jr, Benfenati E, Gini G (2012) QSAR modeling of endpoints for peptides which is based on representation of the molecular structure by a sequence of amino acids. Struct Chem 23:1891–1904. doi:10.1007/s11224-012-9995-0