RISK = F (Molecular Structure)
The application of new HARD-descriptor available from the CORAL software to building up NOAEL models

Alla P. Toropova¹,*, Andrey A. Toropov¹, Marco Marzo¹, Sylvia E. Escher², Jean Lou Dorne³, Nikolaos Georgiadis³, Emilio Benfenati¹

¹Department of Environmental Health Science, Laboratory of Environmental Chemistry and Toxicology, IRCCS-Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20156 Milano, Italy
²Fraunhofer Institute for Toxicology and Experimental Medicine ITEM Hannover, Germany
³Scientific Committee and Emerging risks Unit, European Food Safety Authority, Via Carlo Magno 1A, 43126 Parma, Italy

Abstract

The HARD-index is a line of eleven symbols, which represents the presence, or absence of eight chemical elements (nitrogen, oxygen, sulfur, phosphorus, fluorine, chlorine, bromine, and iodine) and different kinds of chemical bonds (double bond, triple bond, and stereo chemical bond). Optimal molecular descriptors calculated with the Monte Carlo technique (maximization of correlation coefficient between the descriptor and endpoint) gives satisfactory predictive models for no observed adverse effect levels (NOAEL). The models are built up in accordance with OECD principles.

Keywords: QSAR; NOAEL; OECD principles; Monte Carlo method; CORAL software

*Corresponding author
Alla P. Toropova
Laboratory of Environmental Chemistry and Toxicology, IRCCS - Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20156 Milano, Italy
Tel: +39 02 3901 4595
Fax: +39 02 3901 4735
Email: alla.toropova@marionegri.it
Introduction

According to REACH (REACH, 2007), the NOAEL (no observed adverse effect levels) is a reliable criterion for risk assessment various chemicals. Regulators are continually facing the task of assessing health hazards and environmental effects of various chemicals. In fact, the number of substances, which should be taken into account is enormous. For instance, the database of Environmental Risk assessment (ERA) contains more than 100,000 chemical compounds on the European Inventory of Existing Chemicals (EINECS List) (Zvinavashe et al., 2009). Under such circumstances, computational models are very useful tools which can contribute to the hazard assessment of many compounds while reducing de novo animal testing (Kleandrova et al., 2014a,b; Luan et al., 2014). Computational toxicology, an applied science, utilizes the latest advances in mathematics, biology, chemistry, and computer technologies. Integrating all of these sciences into a biologically based computational model enables the researcher to numerically investigate the impact of exposure to environmental chemicals on living systems (Gadaleta et al., 2016; Dearden, 2016; Roy et al., 2015a; Roy and Ambure, 2016; Benfenati, 2007). In addition, the World Health Organization (WHO, 2000) have established an acceptable daily intake (ADI) for an actual risk assessment.

Traditionally, such safe levels in human risk assessment are derived from results of sub-chronic to chronic in vivo toxicological studies in test species (mostly rat, mouse, dog and rabbit) such as No-observed adverse effect level (NOAEL) or Lowest observed adverse effect level (LOAEL) or more recently a benchmark dose limit (BMDL) on which most often a 100-fold uncertainty factor. In food safety, such safe levels are denominated Acceptable daily intake (ADI) or tolerable daily intake (TDI) for regulated compounds (such as pesticides, food and feed additives) and Tolerable Daily Intake for contaminants of anthropogenic or natural origins (e.g dioxins, brominated flame retardants, marine biotins, mycotoxins, alkaloids) (Dorne, 2010; Pizzo and Benfenati, 2016; Diaza et al., 2015; Marzo et al., 2016; Veselinović et al., 2016; Toropova et al., 2015; Toropov et al., 2015).

The NOAEL is an important indicator of danger in utilization of a substance after repeated use and is requested for regulatory purposes (Pizzo and Benfenati, 2016; Diaza et al., 2015; Marzo et al., 2016; Veselinović et al., 2016; Toropova et al., 2015; Toropov et al., 2015). The NOAEL that are derived directly from toxicological studies are analyzed. If several NOAELs values are available, the regulatory focus selects the highest one selected, leading to the common usage of the term NOAEL as the highest exposure without adverse effects (Toropova et al., 2015).
The NOAEL is defined as dose level at which no adverse effects has been observed in a repeated
dose toxicity study and is used to derive the threshold below which a risk for human health is not
likely. In this process several assessment factors are applied to consider differences in human
susceptibility as well as interspecies differences. In vivo studies are, however, no more permitted by
the cosmetics directive, and in general have to be avoided, when feasible, also within other
regulations (e.g. REACH). So far, alternative methods predicting the NOAELs are currently not
available and several initiatives are ongoing to solve this issue.

This endpoint is measured with animal experiments, which are no more permitted by the cosmetics
directive, and in general it should be avoided, when feasible, also within other regulations.
However, so far, the alternative methods for this endpoint have not proved the possibility to replace
it, and several initiatives are ongoing to solve this issue. Among the alternative methods, there are a
few works modelling this endpoint with quantitative structure – activity relationships (QSARs)
methods (Toropov et al., 2015; Goto 2013; Dobchev et al., 2013; Pizzo and Benfenati, 2016). These
models represent interesting efforts to predict a complex endpoint. The difficulties for the specific
toxicological property are related to the fact that the number of substances with high quality
NOAEL values is quite limited. Further, several mechanisms lead to the NOAEL e.g. different
apical findings in different organs. Beside the biological diversity, NOAELs inherit some
uncertainty e.g. due to differences in the experimental protocol (such as time of treatment, scope of
examination, dose selection and dose spacing) as well as interspecies differences. It is to be noted,
that NOAEL can be interrelated with biochemical processes which influence genetical phenomena
(Mukerji et al., 2016).

The aim of this work is the development and the evaluation of QSAR models for NOAEL (Park
and Cho, 2011; Rupp et al., 2010; Alexeeff et al., 2002; Bitsch et al., 2006) by means of the
CORAL software based on the Monte Carlo technique.

Method

Data

The data set are made using the data on the NOAEL taken from the Fraunhofer RepDose®
database and the EFSA's Chemical Hazards Database (Bitsch et al., 2006; EFSA 2013, 2014).

OpenFoodTox provides open source data for the substance characterization, the links to EFSA’s
output, background regulations, and a summary of hazard identification and hazard characterization
for more than 400 substances from over 1650 Scientific Opinions, Statements and Conclusions
through the work of its scientific Panels, Units and Scientific Committee. The database is available
as an opensource tool under: https://dwh.efsa.europa.eu/bi/asp/Main.aspx?rwtrep=400.
Data were selected from sub-chronic studies on rat and with way of exposure oral. Form EFSA database all studies with experimental duration between 77 and 98 days were selected for a total of 166 compounds of which 96 with multiple value (the lowest value was selected for developing the model). Among these 166 data a cluster of aliphatic chain of aldehydes, carboxylic acids and alcohols was found, this cluster include 39 compounds with the same experimental value, so to prevent a bias in the model development only 13 of these compounds randomly chosen among the three chemical classes was selected. Were then selected 137 compounds from EFSA database whit experimental value for NOAEL. From Fraunhofer database (Bitsch et al., 2006) all studies with reliability “A” and “B” and with experimental duration between 85 and 99 days was selected for a total of 362 compounds of which 44 with multiple value (also here the lowest value was selected for developing the model). The two data set was merged and between the two data set there was 15 compounds in common, the lowest value was selected (5 lowest values were from EFSA and 10 from RepDose data set). The final data set obtained is composed by 475 compounds, 119 from EFSA database and 356 from RepDose database.

These 475 compounds were three times randomly split into the training (≈35%), invisible training (≈35%), calibration (≈15%), and external validation sets (≈15%). In addition, the ranges of endpoint for each set are approximately equivalent.

**Optimal descriptor**

The so-called the hybrid optimal descriptor (Veselinovic et al., 2016; Toropova et al., 2015; Toropov et al., 2015) is utilized in this work. The hybrid descriptor takes into account molecular features extracted from simplified molecular input-line entry system (SMILES) and from hydrogen-suppressed graph (HSG). In fact, SMILES and HSG are different (similar but non-identical) representations of the molecular structure (Toropova et al., 2015).

The paradigm used in this work can be expressed as the following:

\[
\text{NOAEL} = F(\text{SMILES}, \text{HSG})
\]  

In Eq. 1, the SMILES is provider of (i) SMILES-atoms i.e. one or two characters, which cannot be examined separately (e.g. ‘Cl’, ‘Br’, etc.); (ii) pairs of SMILES-atoms (Table 1); (iii) BOND, HALO, and NOSP suggested in the literature (Toropova et al., 2015); and (iv) novel global SMILES attribute integrated BOND, HALO, and NOSP into united HARD index. Also, nearest neighbors codes (NNC) calculated with adjacency matrix of HSG (Toropov and Toropova, 2004) are involved in building up the model (Table 1):

\[
\text{NNC}_k = 100 \times N_{\text{total}} + 10 \times N_{\text{Carbon}} + N_{\text{nonCarbon}}
\]
The $N_{total}$, $N_{carbon}$ and $N_{nonCarbon}$ are the total number of neighbors, the number of neighbors which are carbon atoms, and the number of compounds which are not carbon, respectively.

[Table 1 around here]

Table 2 contains the general scheme of the construction of BOND, HALO, NOSP, and HARD.

[Table 2 around here]

Thus, the models suggested in this work are calculated with descriptor of optimal correlation weights (DCW) using the following two different equations (Toropova et al., 2015; Toropov et al., 2015):

\[
DCW_1(T^*, N^*) = CW(HARD) \\
+ \sum CW(S_k) + \sum CW(SS_k) + \sum CW(NNC_k)
\]  \hspace{1cm} (3)

\[
DCW_2(T^*, N^*) = CW(BOND) + CW(NOSP) + CW(HALO) \\
+ \sum CW(S_k) + \sum CW(SS_k) + \sum CW(NNC_k)
\]  \hspace{1cm} (4)

The correlation weights are calculated with the Monte Carlo optimization aimed to maximize the correlation coefficient between optimal descriptor and NOAEL. The optimization should be stopped when the best statistical quality for calibration set is reached. This situation takes place for special values of parameters of the Monte Carlo optimization: (i) the threshold (T); and (ii) the number of epochs (one round of modification of all correlation weights involved in the modeling process). The threshold is an integer to distribute all features into two classes rare and not rare. The rare features are blocked (their correlation weights are equal to zero). The number of epochs is also a parameter affecting the optimization process (Toropova et al., 2015).

The optimization by balance of correlations is used in this work. The training set plays the role of the builder the model. The invisible training set plays the role of inspector of the model since this set controllers the absence of the overtraining. The calibration set is a preliminary estimator of the predictive potential of the model. The external (with chemicals unknown during building up the model) validation set is the final estimator of the predictive potential of the model.

The model is one-parameter correlation as in Eq.5

\[
NOAEL = C_0 + C_1 \times DCW_x \,(T^*, N^*), \, x = 1, 2
\]  \hspace{1cm} (5)

In order to get more robust and representative statistical parameters, based on the presence of different chemicals in the sub-sets above discussed, we randomly split the overall set of compounds three times, with different compositions of the chemicals in the sub-sets.

**Results and Discussion**
QSAR models for NOAEL calculated with described schemes as in Eqs. 3 and 4, on the three splits, are the following:

Split 1
\[
\text{NOAEL} = 0.6089031 (\pm 0.0038666) + 0.0773483 (\pm 0.0002688) \times \text{DCW}_1(1,13) \quad (6)
\]
\[
\text{NOAEL} = 0.1203621 (\pm 0.0059245) + 0.0412961 (\pm 0.0001898) \times \text{DCW}_2(1,15) \quad (7)
\]

Split 2
\[
\text{NOAEL} = 0.5856837 (\pm 0.0049061) + 0.0682434 (\pm 0.0002828) \times \text{DCW}_1(1,14) \quad (8)
\]
\[
\text{NOAEL} = -0.3066240 (\pm 0.0079996) + 0.0810281 (\pm 0.0003725) \times \text{DCW}_2(1,14) \quad (9)
\]

Split 3
\[
\text{NOAEL} = 0.7329505 (\pm 0.0045662) + 0.0788220 (\pm 0.0002943) \times \text{DCW}_1(1,14) \quad (10)
\]
\[
\text{NOAEL} = 0.1426720 (\pm 0.0074527) + 0.0593202 (\pm 0.0002931) \times \text{DCW}_2(1,12) \quad (11)
\]

Table 3 contains the statistical quality of models related to three different splits into the training, invisible training, calibration and validation sets.

The comparison of the models obtained with the new integrated descriptor HARD with model calculated with separated BOND, NOSP, and HALO shows that for the given total set of 487 compounds, the HARD improves the predictive potential of the models (Table 3).

The correlation coefficient metrics for external validation may be misleading (Roy and Ambure, 2016, Afantitis et al., 2008). In this work the average correlation coefficient value is calculated with three different models based on the three different splits of available data into the training, invisible training, calibration, and external validation sets is used as the measure of the predictive potential of the suggested approach.

Having numerical data on several runs of the Monte Carlo optimization one can extract molecular features of three types: (i) features which have positive correlation weights in all runs (these can be qualified as promoters of increase for the endpoint); (ii) features which have negative correlation weights in all runs (these can be qualified as promoters of decrease for the endpoint); and (iii) features which have both positive and negative correlation weights in several runs of the optimization (these have not a defined role). It is necessary to pay attention to frequency of the molecular features in the training, invisible training, and calibration sets. Table 4 contains examples of the promoters of increase or decrease for NOAEL. Table 5 contains comments on the molecular features extracted from SMILES or HSG.

Various distributions of the data into the training, invisible training, calibration, and validation sets do result in different prevalence of molecular features (extracted from SMILES or HSG) in the training and calibration sets. The measure of influence of a feature $F_k$ for possible predictive
potential of a model can be estimated with the measurement of the defect of the feature \( d(F_k) \) calculated as the following:

\[
d(F_k) = \frac{P_T(F_k) - P_C(F_k)}{N_T(F_k) + N_C(F_k)}
\]

where \( P_T(F_k) \) and \( P_C(F_k) \) are probabilities of attribute \( F_k \) in the training and the calibration set, respectively; \( N_T(F_k) \) and \( N_C(F_k) \) are prevalence (frequency) of attribute \( F_k \) in the training set and the calibration set, respectively. The \( d(F_k) = 1 \), if \( N_C(F_k) = 0 \). This situation represents the case with an unbalanced distributions of \( F_k \) in the two sets and obviously this is the worst situation. Ideally, the value should be close to 0.

The defect of SMILES and HSG \( d(SMILES) \) can be estimated via defects of \( F_k \) which presence in the SMILES and HSG:

\[
d(SMILES & HSG) = \sum_{F_k \in SMILES & HSG} d(F_k) \rightarrow \min
\]

The defect of a split into the training, invisible training, calibration, and validation sets can be estimated with the sum of defects of SMILES from the training and calibration sets:

\[
d(Split) = \sum d(SMILES & HSG)
\]

Computational experiments have shown, that described models have preferable predictive potential if the \( d(Split) \) is minimal, because this is an indication that the single splits are quite representative of common situations, and are not affected by local distribution of chemicals in a set of compounds.

The definition of domain of applicability is important component of a QSAR analysis (Roy et al., 2015b; Roy et al., 2016). The statistically robust domain of applicability can be introduced via inequality:

\[
d(SMILES&HSG) < 2 \times \overline{d(SMILES&HSG)}
\]

where, the \( \overline{d(SMILES&HSG)} \) is average defect of SMILES over the training set.

The number of outliers according to inequality 15 are 4, 2, and 4 for splits 1, 2, and 3 respectively. The split defects calculated with Eq. 14 (which related to the sum of all the contributions of defects) are 76.3, 64.3, and 85.2 for splits 1, 2, and 3 respectively. The second split is the best and third split is the worst according to criterion calculated with Eq.14. In this way it is possible to identify more perspective split among different splits, which are not equally useful for modeling purposes.
Table 6 contains the comparison of the statistical quality of the models studied in this work with models suggested in the literature. The comparison confirms that the suggested approach gives satisfactory models, which are able to have practical applications. Our model is based on a larger set of compounds, and the higher performance obtained in particular on the data in the external validation set are quite encouraging.

Conclusions

The HARD index, which is an integrated attribute of SMILES, improves the predictive potential of QSAR models calculated with the Monte Carlo technique described in this work. The new approach has been tested on a set of compounds with values towards NOAEL, which is quite a complex toxicological property. The statistical quality of the suggested models are comparable or even better than the statistical quality of models reported in the literature. The special probabilistic criteria for the definition of domain of applicability are suggested (Eqs. 12-15). Mechanistic interpretation of the CORAL models in terms of promoters of increase or decrease for NOAEL value is used (Table 4). Thus, the CORAL software is building up predictive models for NOAEL according to OECD principles (OECD, 2007).

Acknowledgment

Authors thank the project EU-ToxRisk (Project reference: 681002) and EFSA contract (NP//EFSA/AFSCO/2016/1) for financial support.
References
Afantitis, A., Melagraki, G., Sarimveis, H., Koutentis, P.A., Markopoulos, J., Igglessi-Markopoulou, O., 2008. Development and evaluation of a QSPR model for the prediction of diamagnetic susceptibility. QSAR Comb. Sci., 27 (4), 432-436.
Alexeeff, G.V., Broadwin, R., Liaw, J., Dawson, S.V., 2002. Characterization of the LOAEL-to-NOAEL Uncertainty Factor for Mild Adverse Effects from Acute Inhalation Exposures. Regul. Toxicol. Pharmacol. 36, 96–105.
Bitsch, A., Jacobi, S., Melber, C., Wahnschaffe, U., Simetska, N., Mangelsdorf, I., 2006. REPDOSE: A database on repeated dose toxicity studies of commercial chemicals—A multifunctional tool. Regul. Toxicol. Pharmacol. 46, 202–210.
Benfenati, E., 2007. Predicting toxicity through computers: a changing world. Chem. Cent. J. 1, 32.
Contrera, J.F., Matthews, E.J., Kruhlak, N.L., Benz, R.D., 2004. Estimating the safe starting dose in phase I clinical trials and no observed effect level based on QSAR modeling of the human maximum recommended daily dose. Regular. Toxicol. Pharmacol. 40, 185-206.
Diaza, R.G., Manganelli, S., Esposito, A., Roncaglioni, A., Manganaro, A., Benfenati, E., 2015. Comparison of in silico tools for evaluating rat oral acute toxicity. SAR QSAR Environ Res. 26, 1-27.
Dearden, J.C., 2016. The History and Development of Quantitative Structure-Activity Relationships (QSARs). Int. J. Quantitative Structure-Property Relationships 1(1), 1-44.
Dobchev, D.A., Tulp, I., Karelson, G., Tamm, T., Tämma, K., Karelson, M., 2013. Subchronic oral and inhalation toxicities: a challenging attempt for modeling and prediction. Mol. Inf. 32, 793–801.
Dorne, J.L.C.M., 2010. Metabolism, variability and risk assessment. Toxicology 268 (3), 156-164.
EFSA, 2013. Report on ‘Data collection and data entry for EFSA’s Chemical Hazards Database’ available online: http://www.efsa.europa.eu/en/supporting/pub/458e.htm
EFSA, 2014. Report on ‘Further development and update of EFSA’s Chemical Hazards Database’ available online: http://www.efsa.europa.eu/it/supporting/pub/654e.htm
Gadaleta, D., Mangiatordi, G.F., Catto, M., Carotti, A., Orazio Nicolotti, O., 2016. Applicability Domain for QSAR Models: Where Theory Meets Reality. Int. J. Quantitative Structure-Property Relationships 1(1), 45-63.
Goto, T., 2013. QSAR modeling using a set of intermediate-duration oral NOELs. PhD thesis, https://etd.library.emory.edu/view/record/pid/emory:d724n
Kleandrova, V.V., Luan, F., Gonzalez-Diaz, H., Ruso, J.M., Speck-Planche, A., Cordeiro, M.N.D.S., 2014a. Computational tool for risk assessment of nanomaterials: Novel QSTR-perturbation model for simultaneous prediction of ecotoxicity and cytotoxicity of uncoated and coated nanoparticles under multiple experimental conditions. Environ. Sci. Technol. 48, 14686-14694.

Kleandrova, V.V., Luan, F., Gonzalez-Diaz, H., Ruso, J.M., Melo, A., Speck-Planche, A., Cordeiro, M.N.D.S., 2014b. Computational ecotoxicology: Simultaneous prediction of ecotoxic effects of nanoparticles under different experimental conditions. Environ. Int. 73C, 288-294.

Luan, F., Kleandrova, V.V., Gonzalez-Diaz, H., Ruso, J.M., Melo, A., Speck-Planche, A., Cordeiro, M.N.D.S., 2014. Computer-aided nanotoxicology: assessing cytotoxicity of nanoparticles under diverse experimental conditions by using a novel QSTR-perturbation approach. Nanoscale 6, 10623-10630.

OECD, 2007. OECD Environment Health and Safety Publications Series on Testing and Assessment No.69. Guidance Document on the Validation of (Quantitative) Structure–Activity Relationship [(Q)SAR] Models. http://search.oecd.org/officialdocuments.

Mukerji, P., Roper, J.M., Stahl, B., Smith, A.B., Burns, F., Rae, J.C., Yeung, N., Lyra, A., Svärd, L., Saarinen, M.T., Alhoniemi, E., Ibarra, A., Ouwehand, A.C., 2016. Safety evaluation of AB-LIFE® (Lactobacillus plantarum CECT 7527, 7528 and 7529): Antibiotic resistance and 90-day repeated-dose study in rats. Food Chem. Toxicol. 92, 117-128.

Park, Y.-C., Cho, M.-H., 2011. A new way in deciding NOAEL based on the findings from GLP-toxicity test. Toxicol. Res. 27, 133–135.

Pery, A., Desmots, S., Mombelli, E., 2010. Substance-tailored testing strategies in toxicology: an in silico methodology based on QSAR modeling of toxicological thresholds and Monte Carlo simulations of toxicological testing. Regul. Toxicol. Pharmacol. 56, 82-92.

Pizzo, F., Benfenati, E., 2016. In Silico Models for Repeated-Dose Toxicity (RDT): Prediction of the No Observed Adverse Effect Level (NOAEL) and Lowest Observed Adverse Effect Level (LOAEL) for Drugs. Methods Mol. Biol. 1425, 163-176.

REACH, 2007. Review of Annex IV of the Regulation No. 1907/2007 (REACH). ec.europa.eu/environment/chemicals/reach/pdf/6b_appendix_2.pdf

Roy, K., Kar, S., Das, R.N., 2015a. Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment, pp. 1-479.

Roy, K., Kar, S., Ambure, P., 2015b. On a simple approach for determining applicability domain of QSAR models. Chemometr. Intell. Lab. Syst. 145, 22-29.
Roy, K., Das, R.N., Ambure, P., Aher, R.B., 2016. Be aware of error measures. Further studies on validation of predictive QSAR models. Chemometr. Intell. Lab. Syst. 152, 18-33.

Roy, K., Ambure, P., 2016. The “double cross-validation” software tool for MLR QSAR model development. Chemometr. Intell. Lab. Syst. 159, 108-126.

Rupp, B., Appel, K.E., Gundert-Remy, U., 2010. Chronic oral LOAEL prediction by using a commercially available computational QSAR tool. Arch. Toxicol. 84, 681–688.

Toropova, A.P., Toropov, A.A., Veselinović, J.B., Veselinović A.M., 2015. QSAR as a random event: a case of NOAEL. Environ. Sci. Poll. Res. 22, 8264-8271.

Toropov, A.A., Toropova, A.P., Pizzo, F., Lombardo, A., Gadaleta, D., Benfenati, E., 2015. CORAL: model for no observed adverse effect level (NOAEL). Mol. Divers. 19, 563–575.

Toropov, A.A., Toropova, A.A., 2004. Nearest neighboring code and hydrogen bond index in labeled hydrogen-filled graph and graph of atomic orbitals: application to model of normal boiling points of haloalkanes. J. Mol. Struct. (Theochem) 711, 173–183.

Veselinović, J.B., Veselinović, A.M., Toropova, A.P., Toropov, A.A., 2016. The Monte Carlo technique as a tool to predict LOAEL. Eur. J. Med. Chem. 116, 71-75.

WHO Regional Publications, European Series, Second Edition. 2000. Air quality guideline for Europe. No 91, 273 pages.

Zvinavashe, E., Murk, A.J., Rietjens, I.M.C.M., 2009. On the number of EINECS compounds that can be covered by (Q)SAR models for acute toxicity. Toxicol. Lett. 184, 67–72.
Table 1
An example of molecular features extracted from SMILES ($S_k$, $SS_k$, HARD) and from hydrogen suppressed graph (NNC$_k$)

| SMILES  | Structure | Adacency matrix |
|---------|-----------|-----------------|
| $\text{C(CSCCCI)}\text{Cl}$ | ![Molecular graph](image) | $\begin{array}{cccccccc}
\text{Cl}_1 & \text{C}_2 & \text{C}_3 & \text{S}_4 & \text{C}_5 & \text{C}_6 & \text{Cl}_7 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 
\end{array}$ |

| $S_k$ | CW($S_k$) | N1 | N2 | N3 | $S_k$ | CW($SS_k$) | N1 | N2 | N3 |
|-------|-----------|----|----|----|-------|-----------|----|----|----|
| $\text{C...........}$ | 2.5048 | 142 | 162 | 74 | $\text{C...........}$ | -2.0003 | 139 | 152 | 69 |
| $\text{...........}$ | -2.0003 | 139 | 152 | 69 | $\text{...........}$ | -0.8774 | 127 | 144 | 61 |
| $\text{S...........}$ | -1.4338 | 32 | 38 | 15 | $\text{S...........}$ | -1.8172 | 15 | 19 | 10 |
| $\text{C...........}$ | 2.5048 | 142 | 162 | 74 | $\text{C...........}$ | -0.8774 | 127 | 144 | 61 |
| $\text{...........}$ | -1.0022 | 32 | 44 | 14 | $\text{...........}$ | -3.8711 | 7 | 3 | 1 |
| $\text{...........}$ | -2.0003 | 139 | 152 | 69 | $\text{Cl...........}$ | 0.6228 | 27 | 42 | 13 |
| $\text{Cl...........}$ | -1.0022 | 32 | 44 | 14 | $\text{Cl...........}$ | 0.6228 | 27 | 42 | 13 |
| NNC | CW(NNC$_k$) | N1 | N2 | N3 |
|------|----------------|----|----|----|
| NNC-$\text{C...211.}$ | -2.5654 | 80 | 105 | 50 |
| NNC-$\text{C...202.}$ | -2.5647 | 21 | 19 | 4 |
| NNC-$\text{S...202.}$ | -2.3795 | 4 | 1 | 0 |
| NNC-$\text{C...202.}$ | -2.5647 | 21 | 19 | 4 |
| NNC-$\text{Cl...101.}$ | -1.8106 | 17 | 12 | 3 |
| NNC-$\text{Cl...101.}$ | -1.8106 | 17 | 12 | 3 |
| HARD | CW(HARD) | N1 | N2 | N3 |
|------|----------------|----|----|----|
| S000000100100 | -16.997 | 1 | 0 | 0 |

*) The N1, N2, and N3 are frequencies of molecular features in the training, invisible training, and calibration sets, respectively.
Table 2
The scheme of definition of twelve symbols codes for BOND, NOSP, HALO, and HARD

| ID | Comment | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|----|---------|---|---|---|---|---|---|---|---|---|----|----|----|
| 1  | Definition of BOND attribute | B | O | N | D | C1 | @ | # | 0 | 0 | 0 | 0 | 0 |
| 2  | Definition of NOSP attribute  | N | O | S | P | C1 | C2 | C3 | C4 | 0 | 0 | 0 | 0 |
| 3  | Definition of HALO attribute  | H | A | L | O | C1 | C2 | C3 | C4 | 0 | 0 | 0 | 0 |
| 4  | Definition of HARD attribute  | $ | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 | C10 | C11 |

*) C1-C11 are (0,1) codes: if a feature presents (e.g. double bond ‘=’; nitrogen atom ‘N’; chlorine ‘Cl’, etc.) the code=1, if the feature absent code=0.
Table 3

The statistical characteristics of QSAR models for three splits of data into the training, invisible training, calibration, and validation sets.

| Split | Descriptor | Set       | n   | \( r^2 \) | Y-randomization*** | \( q^2 \) | s     | F     |
|-------|------------|-----------|-----|----------|---------------------|---------|-------|-------|
| 1     | DCW_1      | Training  | 159 | 0.6684   | 0.0058              | 0.6616  | 0.632 | 316   |
|       |            | Invisible training | 156 | 0.6204   | 0.0091              | 0.6122  | 0.681 | 252   |
|       |            | Calibration | 81  | 0.7584   | 0.0049              | 0.7449  | 0.545 | 249   |
|       |            | Validation  | 79  | 0.7096   |                     | 0.586   |       |       |
|       | DCW_2      | Training  | 159 | 0.5982   | 0.0034              | 0.5892  | 0.696 | 234   |
|       |            | Invisible training | 156 | 0.5826   | 0.0053              | 0.5733  | 0.710 | 215   |
|       |            | Calibration | 81  | 0.7176   | 0.0088              | 0.6937  | 0.603 | 201   |
|       |            | Validation  | 79  | 0.6274   |                     | 0.695   |       |       |
| 2     | DCW_1      | Training  | 147 | 0.6668   | 0.0073              | 0.6589  | 0.666 | 290   |
|       |            | Invisible training | 167 | 0.5957   | 0.0052              | 0.5862  | 0.729 | 243   |
|       |            | Calibration | 80  | 0.7536   | 0.0160              | 0.7380  | 0.496 | 239   |
|       |            | Validation  | 81  | 0.7897   |                     | 0.542   |       |       |
|       | DCW_2      | Training  | 147 | 0.6344   | 0.0118              | 0.6255  | 0.698 | 252   |
|       |            | Invisible training | 167 | 0.5810   | 0.0064              | 0.5718  | 0.726 | 229   |
|       |            | Calibration | 80  | 0.5202   | 0.0142              | 0.4937  | 0.723 | 85    |
|       |            | Validation  | 81  | 0.7065   |                     | 0.615   |       |       |
| 3     | DCW_1      | Training  | 155 | 0.6779   | 0.0050              | 0.6708  | 0.634 | 322   |
|       |            | Invisible training | 163 | 0.6735   | 0.0045              | 0.6657  | 0.660 | 332   |
|       |            | Calibration | 79  | 0.6337   | 0.0146              | 0.6150  | 0.599 | 133   |
|       |            | Validation  | 78  | 0.7711   |                     | 0.536   |       |       |
|       | DCW_2      | Training  | 155 | 0.6156   | 0.0055              | 0.6060  | 0.692 | 245   |
|       |            | Invisible training | 163 | 0.5796   | 0.0056              | 0.5701  | 0.755 | 222   |
|       |            | Calibration | 79  | 0.6041   | 0.0088              | 0.5813  | 0.669 | 117   |
|       |            | Validation  | 78  | 0.6641   |                     | 0.630   |       |       |

* The n is the number of compounds in a set; the \( r^2 \) is correlation coefficient between experimental and calculated NOAEL; the \( q^2 \) is the leave-one-out cross-validated \( r^2 \); the s is root-mean squared error; the F is the Fischer F-ratio

** The best prediction indicated by bold

*** Y-randomization test (Afantitis et al., 2008)
Table 4

Correlation weights of promoters of increase or decrease for NOAEL according to three runs of the optimization procedure for split 1.

| ID | $F_k$       | CW($F_k$) in run 1 | CW($F_k$) in run 2 | CW($F_k$) in run 3 | N1  | N2  | N3  | $d(F_k)$ |
|----|-------------|---------------------|---------------------|---------------------|-----|-----|-----|----------|
| 1  | O........... | 2.12455             | 1.18734             | 1.75450             | 115 | 134 | 67  | 0.0004   |
| 2  | NNC-O......110. | 4.75091             | 4.18682             | 2.50236             | 100 | 107 | 53  | 0.0000   |
| 3  | NNC-C......110. | 3.62112             | 3.24520             | 3.81742             | 90  | 96  | 50  | 0.0002   |
| 4  | c...........  | 0.43777             | 0.25460             | 0.44041             | 74  | 86  | 41  | 0.0001   |
| 5  | O......C.... | 0.06182             | 0.62412             | 1.37367             | 58  | 91  | 39  | 0.0010   |
| 6  | 1........... | 6.06188             | 3.62115             | 4.75479             | 54  | 61  | 29  | 0.0000   |
| 7  | =........... | 2.87929             | 0.43360             | 0.24694             | 51  | 53  | 24  | 0.0005   |
| 8  | NNC-N.....220. | 0.62933             | 1.55809             | 1.75412             | 42  | 55  | 28  | 0.0010   |
| 9  | $10001000000 | 14.12210            | 6.80752             | 11.50477            | 24  | 26  | 21  | 0.0022   |
| 10 | $10011000000 | 8.87922             | 2.12073             | 6.62465             | 24  | 22  | 13  | 0.0000   |
| 1  | (........... | -2.18671            | -1.24938            | -1.99693            | 139 | 152 | 69  | 0.0003   |
| 2  | C..........( | -1.62650            | -0.49663            | -0.87781            | 127 | 144 | 61  | 0.0005   |
| 3  | NNC-C.....211. | -1.62416            | -0.68737            | -1.62907            | 80  | 105 | 50  | 0.0007   |
| 4  | 2........... | -3.12351            | -0.87663            | -4.06681            | 64  | 70  | 25  | 0.0013   |
| 5  | S........... | -2.37869            | -1.24620            | -0.68599            | 32  | 38  | 15  | 0.0006   |
| 6  | NNC-C.....303. | -4.25456            | -1.06102            | -1.81192            | 30  | 37  | 12  | 0.0012   |
| 7  | NNC-N.....211. | -5.37450            | -0.12447            | -4.62101            | 29  | 30  | 6   | 0.0033   |
| 8  | S..........( | -6.30856            | -3.87902            | -4.62974            | 28  | 29  | 10  | 0.0016   |
| 9  | n........... | -3.69037            | -0.68336            | -3.50397            | 27  | 34  | 18  | 0.0010   |
| 10 | N......C.... | -1.80855            | -0.12102            | -0.68759            | 26  | 33  | 13  | 0.0003   |

*) The N1, N2, and N3 are frequencies of molecular features in the training, invisible training, and calibration sets, respectively.
Table 5
Comments on molecular features which are promoters of increase or decrease for NOAEL

| ID | $F_k$ | Comments |
|----|------|----------|
| 1  | O........... | Presence of oxygen atoms |
| 2  | NNC–O...110. | Presence of oxygen connected to carbon |
| 3  | NNC–C...110. | Presence of carbon connected to carbon |
| 4  | c...(........ | Branching in aromatic fragment |
| 5  | o...c........ | Oxygen connected to carbon |
| 6  | 1....(........ | Presence of ring and branching of molecular skeleton |
| 7  | =...(........ | Presence of double covalent bonds and branching of molecular skeleton |
| 8  | NNC–N...220. | Nitrogen connected with two atoms of carbon |
| 9  | $10001000000$ | Presence of double bonds and oxygen |
| 10 | $10011000000$ | Presence of double bonds nitrogen and oxygen |
| 1  | (............. | Branching of the molecular skeleton |
| 2  | c...(........ | Branching of molecular skeleton started from carbon |
| 3  | NNC–C...211. | Carbon which has two neighbors one of which is carbon and other is not carbon |
| 4  | 2............. | Presence of two rings |
| 5  | s............. | Presence of sulfur |
| 6  | NNC–C...303. | Carbon which has three neighbors and all are not carbon |
| 7  | NNC–N...211. | Nitrogen which has two neighbors: the first is carbon, the second is not carbon |
| 8  | s...(........ | Branching of molecular skeleton which started from sulfur |
| 9  | n............. | Presence of nitrogen in sp² state |
| 10 | N...C........ | Presence nitrogen connected to carbon |
Table 6
Comparison of NOAEL models suggested in the literature

| Training set | Validation set | Reference |
|--------------|----------------|-----------|
| n | $R^2$ or other statistics | n | $R^2$ |
| - | 0.31 | - | 0.21 | Goto, 2013 |
| 97 | 0.53 | 26 | 0.57 | Toropova et al., 2015 |
| 90 | 0.78 | 15 | 0.68 | Toropov et al., 2015 |
| 160 | Sensitivity 77% Specificity 78% | - | - | Contrera et al., 2004 |
| - | Oral toxicity 88% Inhalation toxicity 87% | - | - | Dobchev et al., 2013 |
| 286 | 0.65 | 55 | 0.76 | Veselinović et al., 2016 |
| 396* | 0.68 | 79 | 0.71 | In this work (split 1) |

*This is a “total” training set, which is structured into a group of (i) training set (n=159), (ii) invisible training set (n=156), and (iii) calibration set (n=81).
Highlights

- Predictive model for NOAEL (sub-chronic toxicity) is suggested;
- The Monte Carlo method is basis to build up the model;
- New descriptor (HARD) calculated from SMILES is suggested;
- The model is built up in accordance with OECD principles.