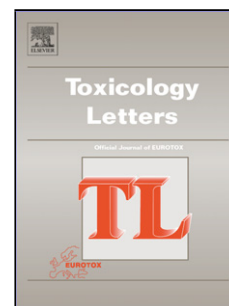


Accepted Manuscript

Title: Hybrid optimal<!--<query id="Q1">Please check whether the author name(s) and affiliation(s) are correct.</query>--> descriptors as a tool<!--<query id="Q2">Please check the footnote given to authors and correct if necessary.</query>--> to predict skin sensitization in accordance to OECD principles



Authors: Alla P. Toropova, Andrey A. Toropov

PII: S0378-4274(17)30125-X
DOI: <http://dx.doi.org/doi:10.1016/j.toxlet.2017.03.023>
Reference: TOXLET 9734

To appear in: *Toxicology Letters*

Received date: 15-2-2017
Revised date: 24-3-2017
Accepted date: 24-3-2017

Please cite this article as: {<http://dx.doi.org/>

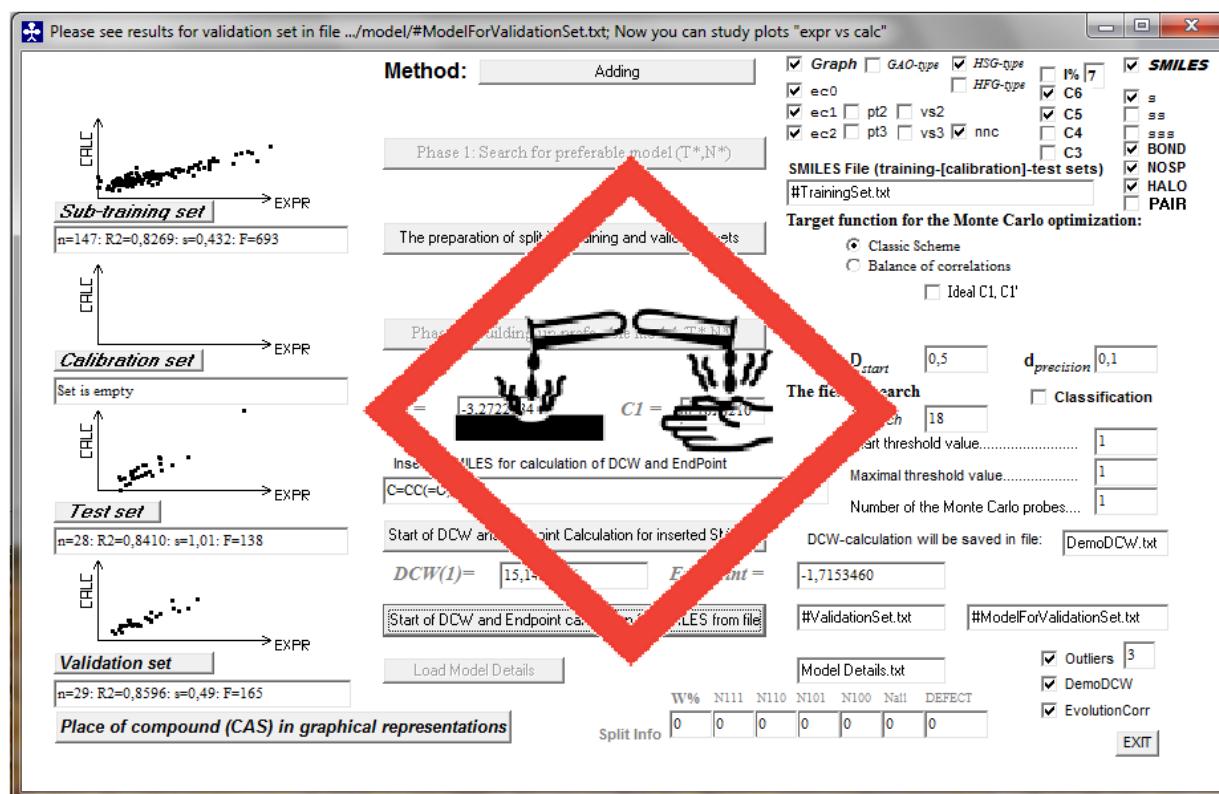
This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

**Hybrid Optimal Descriptors as a Tool to Predict Skin Sensitization
in accordance to OECD principles**

Alla P. Toropova,* Andrey A. Toropov

IRCCS, Istituto di Ricerche Farmacologiche Mario Negri, 20156, Via La Masa 19, Milano, Italy

Graphical Abstract



Highlights

- Quantitative structure – activity relationships for skin Sensitization are built up;
- The statistical quality of the predictive models are quite good;
- The CORAL software (available on the Internet) is utilized to build up the model.

ABSTRACT

Skin sensitization (allergic contact dermatitis) is a widespread problem arising from the contact of chemicals with the skin. The detection of molecular features with undesired effect for skin is a complex task owing to unclear biochemical mechanisms and uncertainty of conditions of action of chemicals to skin. The development of computational methods for estimation of this endpoint in order to reduce animal testing is recommended (Cosmetics Directive EC regulation 1907/2006; EU Regulation, 1223/2009). The CORAL software (<http://www.insilico.eu/coral>) gives good predictive models for the skin sensitization. Simplified molecular input-line entry system (SMILES) together with molecular graph are used to represent the molecular structure for these models. So-called hybrid optimal descriptors are used to establish quantitative structure-activity relationships (QSARs). The aim of this study is the estimation of the predictive potential of the hybrid descriptors. Three different distributions into the training ($\approx 70\%$), calibration ($\approx 15\%$), and validation ($\approx 15\%$) sets are studied. QSAR for these three distributions are built up with using the Monte Carlo technique. The statistical characteristics of these models for external validation set are used as a measure of predictive potential of these models. The best model, according to the above criterion, is characterized by $n_{\text{validation}}=29$, $r^2_{\text{validation}}=0.8596$, $\text{RMSE}_{\text{validation}}=0.489$. Mechanistic interpretation and domain of applicability for these models are defined.

Keywords: SMILES; Skin Sensitization; OECD principles; QSAR; 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 0239014595; fax: +39 0239014735

E-mail: alla.toropova@marionegri.it

* Conflicts of interest: none

* Authors have equivalent contributions into the development of suggested models, preparation of the text, and design of Tables and Figures.

1. Introduction

Skin sensitization is considered a human health risk resulting from skin contact with a wide range of chemicals, including industrial pollutants, household chemicals (detergents), cosmetics, etc. Skin sensitization occurs when small molecules penetrating the skin, and interact with skin proteins blocking immune response. Thus, skin sensitization is an important ecological indicator (Barratt et al., 1994; Cronin et al., 1994; Dimitrov et al., 2005; Hoffmann et al., 2008; Moore et al., 2014a,b; Verheyen et al., 2017). The evaluation of various substances as potential agents of the skin sensitization involves animal testing (OECD, 1992, 2010, 2012, 2015). However, considerable effort has been put into the development of computational methods for estimation of this endpoint in order to reduce animal testing (Cosmetics Directive EC regulation 1907/2006; EU Regulation, 1223/2009). In addition, The European Chemical Agency has defined recommendations for standardization of QSAR models (OECD, 2007).

Before a substance (pharmaceutical, cosmetic, chemical, etc.) can be brought into the European market, it has to be studied in aspect of its impact on human health and environment. The European chemicals legislation on the Registration, Evaluation and Authorization of Chemicals (REACH; Regulation (EC) No 1907/2006) foresees evaluating an estimated 30,000 chemicals by the year 2018, many of which will have to be tested for hazardous effects (EC 1907/2006). Skin sensitization is an endpoint described in Annex VII of the REACH regulation, which implies that its evaluation is mandatory for all REACH-relevant substances manufactured or imported in quantities of one ton or more per year (Mehling et al., 2012). There is a contradiction between the requirement of EU REACH to have the completed data on all compounds involved in the everyday life (Grindon et al., 2008) and the animal welfare (van der Jagt et al., 2004). In vitro tests for use as alternatives to animal testing must imitate the complex interactions between substances and the different parts of the immune system (Nukada et al., 2012; Mehling et al., 2012). In recent years, advances have been made in the field of so-called "in chemico" methods. However possibilities of expansion of applications of such methods is limited (Natsch and Gfeller, 2008; Mehling et al., 2012).

Quantitative structure – activity relationships (QSARs) are the possible way to predict the skin sensitization without animal testing. QSARs are a most popular tool to develop predictive models for variety of endpoints and variety of substances (Afantitis et al., 2010; Furtula and Gutman, 2011; Afantitis et al., 2011; Duchowicz et al., 2012; Torrens and Castellano, 2012; Scotti et al., 2012; Scotti et al., 2014; Torrens and Castellano, 2014; Melagraki et al., 2015; Speck-Planche et al., 2015). In spite of the widely using of QSAR for various tasks, QSAR-researches related to skin

sensitivity remain rare (Mehling et al., 2012). However, there are a few works dedicated to the development of QSARs to predict this endpoint (Roberts et al., 2006; Ren et al., 2006; Roberts et al., 2010; Nandy et al., 2013).

The CORAL software is a tool for the QSAR analyses (Toropov et al., 2013). The molecular structures represented by simplified molecular input-line entry system (SMILES) is the input for the software. User of the CORAL software can build up a model based on solely SMILES. In addition, the user can build up model based on molecular features extracted from SMILES and molecular graph (Gobbi et al., 2016; Toropova et al., 2015; Toropova et al., 2016). Molecular features related to presence of various rings are available in the recent version of the CORAL (Toropova et al., 2016). The aim of present work is estimation of ability of the CORAL software to be a tool to predict the skin sensitization, using so-called hybrid optimal descriptors calculated via representations of the molecular structure by SMILES together with molecular graph. Molecular features related to presence of rings are involved in building up of the models for skin sensitization.

2. Method

2.1. Data

The numerical data on the negative decimal logarithm of concentration (weight percent) giving a stimulation index value of 3 in the local lymph node assay (*pEC3*) is taken in the literature (Dearden et al., 2015). These compounds (n=204) were distributed into the training ($\approx 70\%$), calibration ($\approx 15\%$), and the external validation ($\approx 15\%$) sets. The distributions are random.

2.2. Building up QSAR

The CORAL software builds up QSAR for *pEC3* in form of one-variable correlation

$$pEC3 = C_0 + C_1 \times DCW_x(T^*, N^*), \quad x = 1, 2, 3, 4 \quad (1)$$

The $DCW_x(T^*, N^*)$ is descriptor of correlation weights (DCW). Four version of the $DCW_x(T^*, N^*)$ are analyzed in this work. Table 1 contains formulae to calculate $DCW_1(T^*, N^*)$, $DCW_2(T^*, N^*)$, $DCW_3(T^*, N^*)$, and $DCW_4(T^*, N^*)$.

In order to build up the model, one should extract from set of SMILES the full list of molecular features corresponding to selected CORAL-method and define the optimal values of correlation weights for these features (Figure 1). The optimization of the correlation weights is performed by the Monte Carlo method. The target function is correlation coefficient between *pEC3* and $DCW_x(T^*, N^*)$ for training set. The maximal value for the correlation coefficient is the aim of the

calculation. However, in order to avoid the overtraining, the calculation is stopping when maximal value of the correlation coefficient for the calibration set is reached.

In addition, the optimization has two parameters (i) the threshold, T ; and (ii) the number of epochs of the calculations, N . The threshold is used to separate molecular features into two classes (i) rare; and (ii) not rare. If the frequency of a feature in the training set is less than T , the feature should be recognized as rare. The correlation weights of rare attributes are not involved to building up a model (their correlation weights are zeroth).

Having the numerical data on correlation weights of features involved in building up model, one can using the training set calculate the $DCW_X(T^*, N^*)$ and by the Least squares method define C_0 and C_1 for Eq. 1. Predictive potential of the model should be checked with the external set (Ghaedi et al., 2015; Li et al., 2015; Worachartcheewan et al., 2015; Veselinović, et al., 2015; Veselinović et al., 2015a,b; Worachartcheewan et al., 2014; Toropova et al., 2015).

Different molecular features involved to build up a model are represented by codes of twelve symbols. Table 2 contains description and example of these codes for global (i.e. related to molecule in whole) SMILES attributes. Table 3 contains description of local SMILES attributes. Table 4 contains list of invariants of hydrogen-suppressed graph (HSG) which are using in the CORAL models. Table 5 contains examples of graph invariants C_5 and C_6 in molecular structures. These invariants are calculating with the adjacency matrix (Veselinovic et al., 2014, Das et al., 2017). Figure 2 contains example of translation of molecular structure in HSG and calculation of HSG invariants.

Figure 3 contains of the graphical representation of the selection for the T^* and N^* .

2.3. Mechanistic interpretation

Having the numerical data on the correlation weights of features which takes place in several runs of the Monte Carlo optimization, one can extract three categories of these features: (i) Features which have positive value of the correlation weight in all runs. These are promoters of endpoint increase; (ii) Features which have negative value of the correlation weight in all runs. These are promoters of endpoint decrease; (iii) Features which have both negative and positive value of the correlation weight in different runs of the optimization. These are features with unclear role (one cannot classify these features as promoter of increase or decrease for endpoint).

2.4. Domain of applicability

Different distributions into the training, calibration, and validation sets are characterized by different prevalence of features 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 via defect of F_k , $d(F_k)$ calculated as the following (Toropova et al., 2015; Gobbi et al., 2016; Toropova et al., 2016):

$$d(F_k) = \frac{P_T(F_k) - P_C(F_k)}{N_T(F_k) + N_C(F_k)} \quad (2),$$

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 of attribute F_k in the training and the calibration set, respectively. The $d(F_k) = 1$, if $N_C(F_k) = 0$.

The defect of the system SG i.e. “SMILES and graph (HSG)” $d(SG)$ can be estimated via defects of F_k which presence in the SMILES or HSG:

$$d(SG) = \sum_{F_k \in SG} d(F_k) \rightarrow \min \quad (3)$$

The defect of a distribution into the training, calibration, and validation sets can be estimated via sum of defects of SG:

$$d(Distribution) = \sum d(SG) \quad (4)$$

Computational experiments have shown, that described models have preferable predictive potential if, each $d(SG)$ is minimal. The statistically robust domain of applicability can be introduced via inequality

$$d(SG) < 2 \times \overline{d(SG)} \quad (5),$$

where, the $\overline{d(SG)}$ is average defect of systems “SMILES and graph” over the training set.

2.5. Validation of a model

Internal validation. The Y-randomization test suggested in the literature (Zhang et al., 2006; Afantitis et al., 2008; Ojha and Roy, 2011.) has been used for the internal validation of examined models. *External validation.* The statistical characteristics of the CORAL model for external validation set have been used as the basic criterion for estimation of the predictive potential of models.

3. Results

3.1. QSAR

Table 6 contains the statistical characteristics of models based on the hybrid optimal descriptor $DCW_1(T^*, N^*)$. Computational experiments show that this is the best version of the optimal descriptors examined in this work.

The comparison of statistical characteristics of model based on $DCW_1(T^*,N^*)$ and $DCW_2(T^*,N^*)$ (which calculated without correlation weights for C5 and C6) confirms the significance of the correlation weights of C5 and C6 for predictive potential of the model. In other words, presence (absence) of rings has strong influence to studied endpoint.

In addition, Table 6 contains:

- (a) the statistical characteristics of models, which are calculated with correlation weights of all available SMILES attributes without any correlation weights of HSG ($DCW_3(T^*,N^*)$);
- (b) the statistical characteristics of models calculated with correlation weights of all available invariants of HSG without any correlation weights of SMILES attributes ($DCW_4(T^*,N^*)$)

One can see from Table 6, that for all distribution take place inequalities:

$r_1^2 > r_2^2$; $r_2^2 \gg r_3^2$; $r_2^2 \gg r_4^2$, where the r_1^2 is related to model calculated with $DCW_1(T^*,N^*)$; the r_2^2 is related to model that calculated with $DCW_2(T^*,N^*)$; the r_3^2 is related to model that calculated with $DCW_3(T^*,N^*)$; and the r_4^2 is related to model that calculated with $DCW_4(T^*,N^*)$.

Thus, Table 6 indicates: (i) predictive potential of the hybrid optimal descriptors is better than predictive potential of optimal descriptors based on solely SMILES or based on solely molecular graph; (ii) the involving of correlation weights of molecular features expressed via C5 and C6 improves the statistical quality of models.

The hybrid optimal descriptors, which involve correlation weights of C5 and C6 give the following models:

$$pEC3 = -3.2722(\pm 0.0083) + 0.1028(\pm 0.0003) \times DCW(1,18) \quad (6)$$

$$pEC3 = -1.5607(\pm 0.0053) + 0.06165(\pm 0.0003) \times DCW(1,10) \quad (7)$$

$$pEC3 = -2.5741(\pm 0.0076) + 0.08630(\pm 0.0003) \times DCW(1,13) \quad (8)$$

Figure 4 contains the graphical representation of these models.

3.2. Mechanistic interpretation

Table 7 contains examples of the promoters for increase or decrease of $pEC3$. It is to be noted the selection should be done with taking into account the prevalence of features in the training and calibration sets.

3.3. Domain of applicability

The numbers of outliers for the hybrid optimal descriptors according to inequality 5 for distributions 1, 2, and 3 are 2, 1, and 8, respectively. This confirms the principle: “QSAR is a random event” (Toropov et al., 2013b; Toropova et al., 2014) which takes place for each randomized distribution into the training set and the validation sets.

4. Discussion

The overall design of the CORAL model is based on the following principles: (i) a QSAR model is a random event; (ii) a distribution of available data into the training, calibration, and validation sets is a basis of establishing the random event; and (iii) the statistical characteristics observed for a given the distribution is the representation of the random event.

The described approach has been applied for the QSAR analysis of various endpoints in recent works (Toropov et al., 2013; Masand et al., 2014; Veselinovic et al., 2015; Veselinovic et al., 2015a,b; Worachartcheewan et al., 2015; Toropova et al., 2015; Gobbi et al., 2016; Toropova et al., 2016). The main specificity of the approach is an interpretation of QSPR/QSAR models as random events (Toropov et al., 2013b; Toropova et al., 2014). According to this interpretation, the predictive potential of any QSPR/QSAR model should be confirmed for a group of different distributions into the training, calibration, and validation set.

The comparison of models calculated with the CORAL software with models suggested in the literature (Grace et al., 2003; Promkatkaew et al., 2014; Dearden et al., 2015) has shown that the CORAL software can be a tool to predict the skin sensitization (Table 8). It is to be noted, that the majority of models for skin sensitization are classification models i.e. Yes / No or Active / Inactive (Golla et al., 2009; Jaworska et al., 2011, 2013; Alves et al., 2015) whereas CORAL models give numerical prediction for skin sensitization. In addition, QSAR models for skin sensitization suggested in the literature (Grace et al., 2003; Promkatkaew et al., 2014; Dearden et al., 2015; Golla et al., 2009; Jaworska et al., 2011, 2013; Alves et al., 2015), as the rule, are based on different physicochemical properties, 3D molecular geometry, and descriptors of quantum mechanics (Dearden et al., 2015), whereas the CORAL software is building up models for the endpoint from data represented by solely SMILES.

The suggested models indicate the following structural alerts (molecular features that lead to increase skin sensitization): (i) presence of five-member rings; (ii) presence aromatic six-member rings; (iii) presence of double covalent bonds (Table 7).

Table 9 contains outliers, which take place for distribution-2 and distribution-3. Figure 4 contains these outliers (together with their CAS numbers) indicated by squares.

Supplementary materials section contains (i) three distributions into the training, calibration, and validation sets; experimental and calculated with Eqs 6-8 values of $pEC3$; (ii) the detailed description of the used the CORAL methods; and (iii) the numerical data on correlation weights obtained in the case of distribution 1 for the hybrid optimal descriptor $DCW_1(T^*, N^*)$.

The CORAL software is free available on the Internet. Thus, a user can check up the reproducibility of the described models. The suggested models are also available on request in form a folder with described models available to calculate the $pEC3$ by Eqs. 6-8 for arbitrary substances, represented by SMILES.

5. Conclusions

The distribution into the training, calibration, and validation sets has apparent influence upon the predictive potential models of the $pEC3$. However, for all examined distributions, the statistical quality of the models based on hybrid optimal descriptors is more or less satisfactory and comparable with models suggested in the literature (Table 8). The suggested models built up according to OECD principles, i.e. (i) there is a defined endpoint; (ii) unambiguous algorithm; (iii) a defined domain of applicability; (iv) appropriate measure of goodness-of-fit, robustness and predictivity; and (v) a mechanistic interpretation (OECD, 2007). There is the mechanistic interpretation of suggested model: the presence of aromatic six-members rings and presence of double bonds are promoter of skin sensitization increase (Table 7). The hybrid optimal descriptors, which are calculated with Eqs. 6-8 are characterized by the best predictive potential. The hybrid optimal descriptors, which are calculated without correlation weights for C5 and C6 are characterized by slightly decreased predictive potential (Table 6). The predictive potential of CORAL models based on solely SMILES ($DCW_3(T^*, N^*)$) as well as models based on solely hydrogen-suppressed graph ($DCW_4(T^*, N^*)$) is considerably poorer in comparison with hybrid optimal descriptor ($DCW_1(T^*, N^*)$).

Acknowledgements

Authors thank the project EU-ToxRisk (Project reference: 681002) and LIFE-COMBASE contract (LIFE15 ENV/ES/000416) for financial support. In addition, authors express gratitude to Prof. D.W. Roberts for valuable consultations.

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.
- Afantitis, A., Melagraki, G., Koutentis, P.A., Sarimveis, H., Kollias, G., 2011. Ligand - Based virtual screening procedure for the prediction and the identification of novel β -amyloid aggregation inhibitors using Kohonen maps and Counterpropagation Artificial Neural Networks. *Eur. J. Med. Chem.* 46 (2), 497-508.
- Afantitis, A., Melagraki, G., Sarimveis, H., Koutentis, P.A., Igglessi-Markopoulou, O., Kollias, G., 2010. A combined LS-SVM & MLR QSAR workflow for predicting the inhibition of CXCR3 receptor by quinazolinone analogs. *Mol. Divers.* 14 (2), 225-235.
- Alves V. M., Muratov E., Fourches D., Strickland J., Kleinstreuer N., Andrade C. H., Tropsha A., 2015. Predicting chemically-induced skin reactions. Part I: QSAR models of skin sensitization and their application to identify potentially hazardous compounds. *Toxicol. Appl. Pharm.* 284, 262-272.
- Barratt, M.D., Basketter, D.A., Roberts, D.W., 1994. Skin sensitization structure-activity relationships for phenyl benzoates. *Toxicol. In Vitro* 8 (4), 823-826.
- Cronin, M.T.D., Basketter, D.A., 1994. Multivariate QSAR analysis of a skin sensitization database. *SAR QSAR Environ. Res.* 2 (3), 159-179.
- Das, K.C., Gutman, I., Furtula, B. 2017. On spectral radius and energy of extended adjacency matrix of graphs. *Appl. Math. Comput.*, 296, 116-123.
- Dearden, J.C., Hewitt, M., Roberts, D.W., Enoch, S.J., Rowe, P.H., Przybylak, K.R., Vaughan-Williams, G.D., Smith, M.L., Pillai, G.G., Katritzky, A.R., 2015. Mechanism-Based QSAR Modeling of Skin Sensitization. *Chem. Res. Toxicol.* 28 (10), 1975-1986.
- Dimitrov, S.D., Low, L.K., Patlewicz, G.Y., Kern, P.S., Dimitrova, G.D., Comber, M.H.I., Phillips, R.D., Niemela, J., Bailey, P.T., Mekenyan, O.G., 2005. Skin sensitization: Modeling based on skin metabolism simulation and formation of protein conjugates. *Int. J. Toxicol.* 24 (4), 189-204.
- Duchowicz, P.R., Comelli, N.C., Ortiz, E.V., Castro, E.A., 2012. QSAR study for carcinogenicity in a large set of organic compounds. *Curr. Drug Saf.* 7 (4), 282-288.
- EC regulation 1907/2006. Regulation (EC) No 1907/2006 of the European Parliament and the council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and

- Restriction of chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC.
- EU Regulation, 1223/2009. Of The European Parliament And Of The Council of 30 November 2009 on cosmetic products. OJEU L 342/59-209 .
- Furtula, B., Gutman, I., 2011. Relation between second and third geometric-arithmetic indices of trees. *J. Chemom.* 25 (2), 87-91.
- Hoffmann, S., Saliner, A.G., Patlewicz, G., Eskes, C., Zuang, V., Worth, A.P., 2008. A feasibility study developing an integrated testing strategy assessing skin irritation potential of chemicals. *Toxicol. Lett.* 180 (1), 9-20.
- Ghaedi, A., 2015. Predicting the cytotoxicity of ionic liquids using QSAR model based on SMILES optimal descriptors. *J. Mol. Liq.* 208, 269-279.
- Gobbi, M., Beeg, M., Toropova, M.A., Toropov, A.A., Salmona, M., 2016. Monte Carlo method for predicting of cardiac toxicity: hERG blocker compounds. *Toxicol. Lett.* 250–251, 42–46.
- Golla, S., Madhally, S., Robinson Jr, R.L., Gasem, K.A.M., 2009. Quantitative structure–property relationship modeling of skin sensitization: A quantitative prediction. *Toxicol. In Vitro* 23, 454–465.
- Grace, P., Roberts, D.W., Walker, J.D., 2003. QSARs for the skin sensitization potential of aldehydes and related compounds. *QSAR Comb. Sci.* 22 (2), 196-203.
- Grindon, C., Combes, R., Cronin, M.T.D., Roberts, D.W., Garrod, J.F., 2008. An integrated decision-tree testing strategy for skin sensitisation with respect to the requirements of the EU REACH legislation. *Altern. Lab. Anim.* 36 (SUPPL. 1), 75-89.
- Jaworska, J., Harol, A., Kern, P.S., Gerberick, G.F., 2011. Integrating non-animal test information into an adaptive testing strategy - skin sensitization proof of concept case. *ALTEX* 28(3), 211-225.
- Jaworska, J., Dancik, Y., Kern, P., Gerberick, F., Natsch, A., 2013. Bayesian integrated testing strategy to assess skin sensitization potency: from theory to practice. *J. Appl. Toxicol.* 33(11), 1353–1364.
- Li, Q., Ding, X., Si, H., Gao, H., 2014. QSAR model based on SMILES of inhibitory rate of 2, 3-diarylpropenoic acids on AKR1C3. *Chemometr. Intell. Lab. Syst.* 139, 132-138.
- Masand, V.H., Toropov, A.A., Toropova, A.P., Mahajan, D.T., 2014. QSAR models for anti-malarial activity of 4-aminoquinolines. *Curr. Comput. Aided. Drug Des.* 10 (1), 75-82.

- Mehling, A., Eriksson, T., Eltze, T., Kolle, S., Ramirez, T., Teubner, W., van Ravenzwaay, B., Landsiedel, R., 2012. Non-animal test methods for predicting skin sensitization potentials. *Arch. Toxicol.* 86, 1273–1295.
- Melagraki, G., Afantitis, A., 2015. A risk assessment tool for the virtual screening of metal oxide nanoparticles through enalos insiliconano platform. *Curr. Top. Med. Chem.* 15 (18), 1827-1836.
- Moore, C.A., Wilkinson, S.C., Blain, P.G., Dunn, M., Aust, G.A., Williams, F.M., 2014a. Use of a human skin in vitro model to investigate the influence of 'every-day' clothing and skin surface decontamination on the percutaneous penetration of organophosphates. *Toxicol. Lett.* 229(1), 257-264.
- Moore, C.A., Wilkinson, S.C., Blain, P.G., Dunn, M., Aust, G.A., Williams, F.M., 2014b. Percutaneous absorption and distribution of organophosphates (chlorpyrifos and dichlorvos) following dermal exposure and decontamination scenarios using in vitro human skin model. *Toxicol. Lett.* 229 (1), 66-72.
- Nandy, A., Kar, S., Roy, K., 2013. Development and validation of regression-based QSAR models for quantification of contributions of molecular fragments to skin sensitization potency of diverse organic chemicals. *SAR QSAR Environ. Res.* 24 (12), 1009-1023.
- Natsch, A., Gfeller, H., 2008. LC-MS–Based characterization of the peptide reactivity of chemicals to improve the in vitro prediction of the skin sensitization potential. *Toxicol. Sci.* 106(2), 464–478.
- Nukada, Y., Ashikaga, T., Miyazawa, M., Hirota, M., Sakaguchi, H., Sasa, H., Nishiyama, N., 2012. Prediction of skin sensitization potency of chemicals by human Cell Line Activation Test (h-CLAT) and an attempt at classifying skin sensitization potency. *Toxicol. in Vitro* 26, 1150–1160.
- OECD, 1992. OECD Guideline for the testing of chemicals 406: Skin sensitization, adopted: 17 July 1992. http://www.oecd-ilibrary.org/environment/oecd-guidelines-for-the-testing-of-chemicals-section-4-health-effects_20745788
- OECD, 2010. OECD Guideline for the testing of chemicals 429: Skin sensitization: local lymph node assay revised. 2010. http://www.oecd-ilibrary.org/environment/oecd-guidelines-for-the-testing-of-chemicals-section-4-health-effects_20745788
- OECD, 2012. The adverse outcome pathway for skin sensitisation initiated by covalent binding to proteins. OECD environment, health and safety publications: No.168; ENV/JM/MONO(2012)10.

- OECD, 2015. Performance standards for assessment of proposed similar or modified in vitro skin sensitization ARE–NRF2 Luciferase test method series on testing and assessment No. 213; ENV/JM/MONO (2015) 6.
- OECD, 2007. (Organization for Economic Co-operation and Development). Guidance Document on The Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models No. 69.
- Ojha, P.K., Roy, K., 2011. Comparative QSARs for antimalarial endochins: Importance of descriptor-thinning and noise reduction prior to feature selection. *Chemometr. Intell. Lab. Syst.* 109(2), 146-161.
- Promkatkaew, M., Gleeson, D., Hannongbua, S., Gleeson, M. P., 2014. Skin sensitization prediction using quantum chemical calculations: A theoretical model for the SNAr Domain. *Chem. Res. Toxicol.* 27 (1), 51–60.
- Ren, Y., Liu, H., Xue, C., Yao, X., Liu, M., Fan, B., 2006. Classification study of skin sensitizers based on support vector machine and linear discriminant analysis. *Anal. Chim. Acta* 572 (2), 272-282.
- Roberts, D.W., Aptula, A.O., Patlewicz, G., 2006. Mechanistic applicability domains for non-animal based prediction of toxicological endpoints. QSAR analysis of the Schiff base applicability domain for skin sensitization. *Chem. Res. Toxicol.* 19 (9), 1228-1233.
- Roberts, D.W., Patlewicz, G.Y., 2010. Updating the skin sensitization in vitro data assessment paradigm in 2009 - A chemistry and QSAR perspective. *J. Appl. Toxicol.* 30 (3), 286-288.
- Scotti, L., Lima, E.O., da Silva, M.S., Ishiki, H., Lima, I.O., Pereira, F.O., Mendonça Jr, F.J.B., Scotti, M.T., 2014. Docking and PLS studies on a set of thiophenes RNA polymerase inhibitors against *Staphylococcus aureus*. *Curr. Top. Med. Chem.* 14 (1), 64-80.
- Scotti, L., Ishiki, H., Ferreira, M.J.P., Francisco Jr., J.B.M., De P. Emerenciano, V., Silva, M.S., Scotti, M.T., 2012. In silico methods applied in food chemistry: A short review with bitter and Mutagenic compounds. *Lett. Drug Des. Discov.* 9 (5), 527-534.
- Speck-Planche, A., Kleandrova, V.V., Luan, F., Ds Cordeiro, M.N., 2015. Computational modeling in nanomedicine: Prediction of multiple antibacterial profiles of nanoparticles using a quantitative structure-activity relationship perturbation model. *Nanomedicine*, 10 (2), 193-204.
- Toropov, A.A., Toropova, A.P., 2002a. QSAR modeling of toxicity on optimization of correlation weights of Morgan extended connectivity. *J. Mol. Struc. THEOCHEM* 578, 129-134.

- Toropov, A.A., Toropova, A.P., 2002b. Modeling of acyclic carbonyl compounds normal boiling points by correlation weighting of nearest neighboring codes. *J. Mol. Struc. THEOCHEM* 581, 11-15.
- Toropov, A.A., Toropova, A.P., 2003. QSPR modeling of alkanes properties based on graph of atomic orbitals. *J. Mol. Struc. THEOCHEM* 637, 1-10.
- Toropov, A.A., Toropova, A.P., Benfenati, E., Gini, G., Leszczynska, D., Leszczynski, J., 2013a. CORAL: QSPR model of water solubility based on local and global SMILES attributes. *Chemosphere* 90 (2), 877-880.
- Toropov, A.A., Toropova, A.P., Puzyn, T., Benfenati, E., Gini, G., Leszczynska, D., Leszczynski, J., 2013b. QSAR as a random event: Modeling of nanoparticles uptake in PaCa2 cancer cells. *Chemosphere* 92 (1), 31-37.
- Toropova, A.P., Toropov, A.A., Veselinović, J.B., Veselinović, A.M., 2014. QSAR as a random event: a case of NOAEL. *Environ. Sci. Pollut. Res.* 22 (11), 8264-8271.
- Toropova, A.P., Schultz, T.W., Toropov, A.A., 2016. Building up a QSAR model for toxicity toward *Tetrahymena pyriformis* by the Monte Carlo method: A case of benzene derivatives. *Environ. Toxicol. Pharm.* 42, 135-145.
- Toropova, M.A., Toropov, A.A., Raška, I., Rašková, M., 2015. Searching therapeutic agents for treatment of Alzheimer disease using the Monte Carlo method. *Comput. Biol. Med.* 64, 148-154.
- Torrens, F., Castellano, G., 2014. QSPR prediction of chromatographic retention times of pesticides: Partition and fractal indices. *J. Environ. Sci. Health. B.* 49 (6), 400-407.
- Torrens, F., Castellano, G., 2012. QSPR prediction of retention times of phenylurea herbicides by biological plastic evolution. *Curr. Drug Saf.* 7 (4), 262-268.
- Vandebriel, R.J., Loveren, H.V., 2010. Non-animal sensitization testing: State-of-the-art. *Crit. Rev. Toxicol.* 40 (5), 389-404.
- van der Jagt, K., Munn, S., Torslov, J., de Bruijn, J., 2004. Alternative approaches can reduce the use of test animals under REACH. Addendum to: Assessment of additional testing needs under REACH effects of (Q)SARS, risk based testing and voluntary industry initiatives. IHCP report EUR 21405 EN
- Verheyen, G.R., Braeken, E., Van Deun, K., Van Miert, S., 2017. Evaluation of existing (Q)SAR models for skin and eye irritation and corrosion to use for REACH registration. *Toxicol. Lett.* 265, 47-52.

- Veselinović, A.M., Veselinović, J.B., Toropov, A.A., Toropova, A.P., Nikolić, G.M. 2014. Qsar models for the reactivation of sarin inhibited acetylcholinesterase by quaternary pyridinium oximes based on Monte Carlo method. *Cur. Comput.-Aid. Drug Des.* 10 (3), 266-273.
- Veselinović, A.M., Veselinović, J.B., Toropov, A.A., Toropova, A.P., Nikolić, G.M., 2015a. In silico prediction of the β -cyclodextrin complexation based on Monte Carlo method. *Int. J. Pharm.* 495 (1), 404-409.
- Veselinović, A.M., Veselinović, J.B., Živković, J.V., Nikolić, G.M., 2015b. Application of smiles notation based optimal descriptors in drug discovery and design. *Curr. Top. Med. Chem.* 15(18), 1768-1779.
- Veselinović, J.B., Nikolić, G.M., Trutić, N.V., Živković, J.V., Veselinović, A.M., 2015. Monte Carlo QSAR models for predicting organophosphate inhibition of acetylcholinesterase. *SAR QSAR Environ. Res.* 26(6), 449-460.
- Worachartcheewan, A., Prachayasittikul, V., Toropova, A.P., Toropov, A.A., Nantasenamat, C., 2015. Large-scale structure-activity relationship study of hepatitis C virus NS5B polymerase inhibition using SMILES-based descriptors. *Molec. Divers.* 19 (4), 955-964.
- Worachartcheewan, A., Mandi, P., Prachayasittikul, V., Toropova, A.P., Toropov, A.A., Nantasenamat, C., 2014. Large-scale QSAR study of aromatase inhibitors using SMILES-based descriptors. *Chemometr. Intell. Lab. Syst.* 138, 120-126.
- Zhang, S., Golbraikh, A., Oloff, S., Kohn, H., Tropsha, A., 2006. A novel Automated Lazy Learning QSAR (ALL-QSAR) approach: Method development, applications, and virtual screening of chemical databases using validated ALL-QSAR models. *J. Chem. Inf. Model.* 46 (5), 1984-1995.

SMILES

```

***
+6668-24-2 O=C(C(C)C(C)=O)c1cccc1 -1.462
+170928-69-5 Cc1c(cc(C)c(C)c1C)C(=O)CC(=O)OCC -1.519
+107-75-5 O=CCC(C)CCCC(C)(C)O -1.519
+97-96-1 CCC(CC)C=O -1.881
+109-55-7 NCCCN(C)C -0.342
+107-15-3 NCCN -0.342
+111-40-0 NCCNCCN -0.763
#89-25-8 O=C2CC(C)=NN2c1cccc1 -0.929
+106-24-1 C\C(C)=C\CCC(\C)=C\CO -1.415
#1086-00-6 ClCc4cccc1cccc2cccc3cccc4c1c23 2.301
+26172-55-4 ClC1=CC(=O)N(C)S1 2.046
+70-25-7 N=C(N[N+](=O)[O-])N(C)N=O 1.523
+684-93-5 NC(=O)N(C)N=O 1.301
+57-57-8 O=C1CCO1 0.824
#100-39-0 BrCc1cccc1 0.699
+2374-65-4 O=S(=O)(CCCCCCCCC)OC 0.409
+55406-53-6 O=C(NCCCC)OCC#CI 0.046
+759-73-9 NC(=O)N(C)N=O -0.041
+2634-33-5 O=C2N8c1cccc12 -0.362
#112-82-3 BrCCCCCCCCCCCCCCCCC -0.362
+118-58-1 Oc2cccc2C(=O)OCc1cccc1 -0.462
+64-67-5 O=S(=O)(OCC)OCC -0.519
+3508-00-7 BrCCCCCCCCCCCCCCCCC -0.681
*629-72-1 BrCCCCCCCCCCCCCCCCC -0.708
+137-26-8 CN(C)C(=S)SSC(=S)N(C)C -0.716
#4276-49-7 BrCCCCCCCCCCCCCCCCC -0.785
*103-63-9 BrCc1cccc1 -0.792
#66-27-3 CS(=O)(=O)OC -0.908
#6938-66-5 BrCCCCCCCCCCCCCCCCC -0.919
*51323-71-8 CS(=O)(=O)OCCCCCCCCCCCCC -0.944
+4860-03-1 CCCCCCCCCCCCCCCCCC1 -0.959
*112-71-0 BrCCCCCCCCCCCCCCCCC -0.964
*111-25-1 BrCCCCC -1.000
*765-09-3 BrCCCCCCCCCCCCCCCCC -1.000
+4292-19-7 CCCCCCCCCCCCCI -1.114
*3386-33-2 CCCCCCCCCCCCCCCCCC1 -1.204
+120-51-4 O=C(OCc1cccc1)c2cccc2 -1.230
-73367-80-3 O=C(O)CCCCCCCCCCCCBr -1.255
*544-77-4 CCCCCCCCCCCCCCCCCI -1.279
+693-67-4 BrCCCCCCCCCCCCC -1.301
***

```

Features, F

```

***
1.....
2.....
3.....
4.....
5.....
=.....
@.....
@@.....
C.....
C5.....0...
C5.....2...
C5.....H.1...
C5.....H.2...
C5...AH.1...
C6.....0...
C6.....1...
C6.....3...
C6...A..1...
C6...A..2...
C6...A..3...
C6...A..4...
C6...A..5...
C6...AH.1...
C6...AH.2...
C6...AH.4...
BOND00000000
BOND01000000
BOND10000000
BOND10100000
BOND11000000
F.....
EC0-C...1...
EC0-C...2...
EC0-C...3...
EC0-C...4...
EC0-F...1...
EC0-Br..1...
EC0-I...1...
EC0-Cl..1...
EC0-N...1...
EC0-N...2...

```

CW(F)

```

***
-0.50312
-1.93946
5.62601
1.05781
1.37538
-0.74818
0.0
0.12043
10.56604
1.94000
14.30947
3.75314
-7.50259
4.93271
4.12627
0.0
4.06649
1.87710
-6.50360
8.37627
5.75267
14.25272
4.37070
-0.25145
5.75211
5.37661
-1.62488
-2.25086
11.87461
-0.93446
-1.93468
-0.56262
-1.25261
1.68826
-0.62671
-3.18295
2.00293
-0.56471
-0.24808
-1.87050

```



Figure 1

Scheme of building up the CORAL models

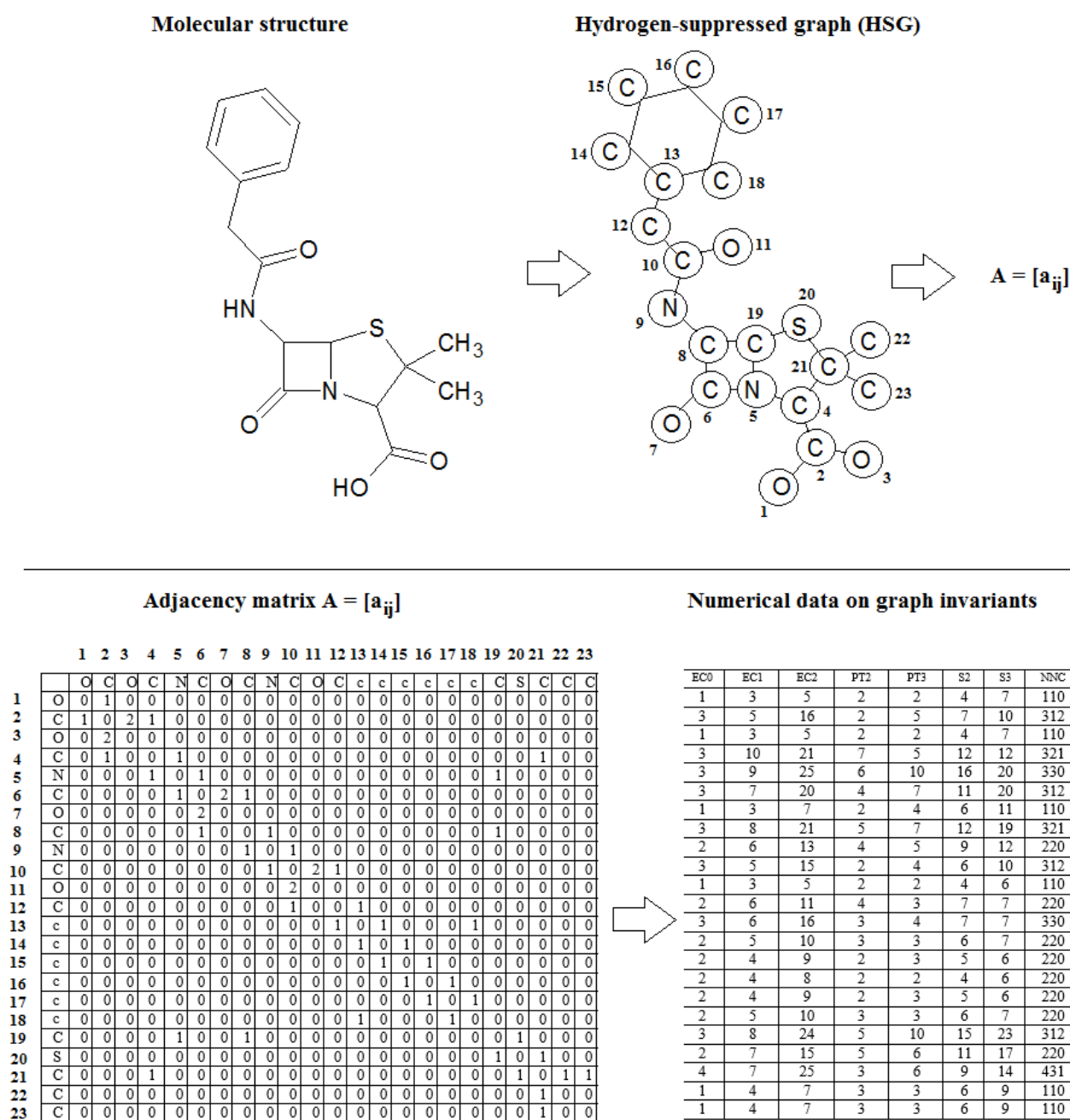


Figure 2

Example of adjacency matrix and numerical data on graph invariants (for HSG)

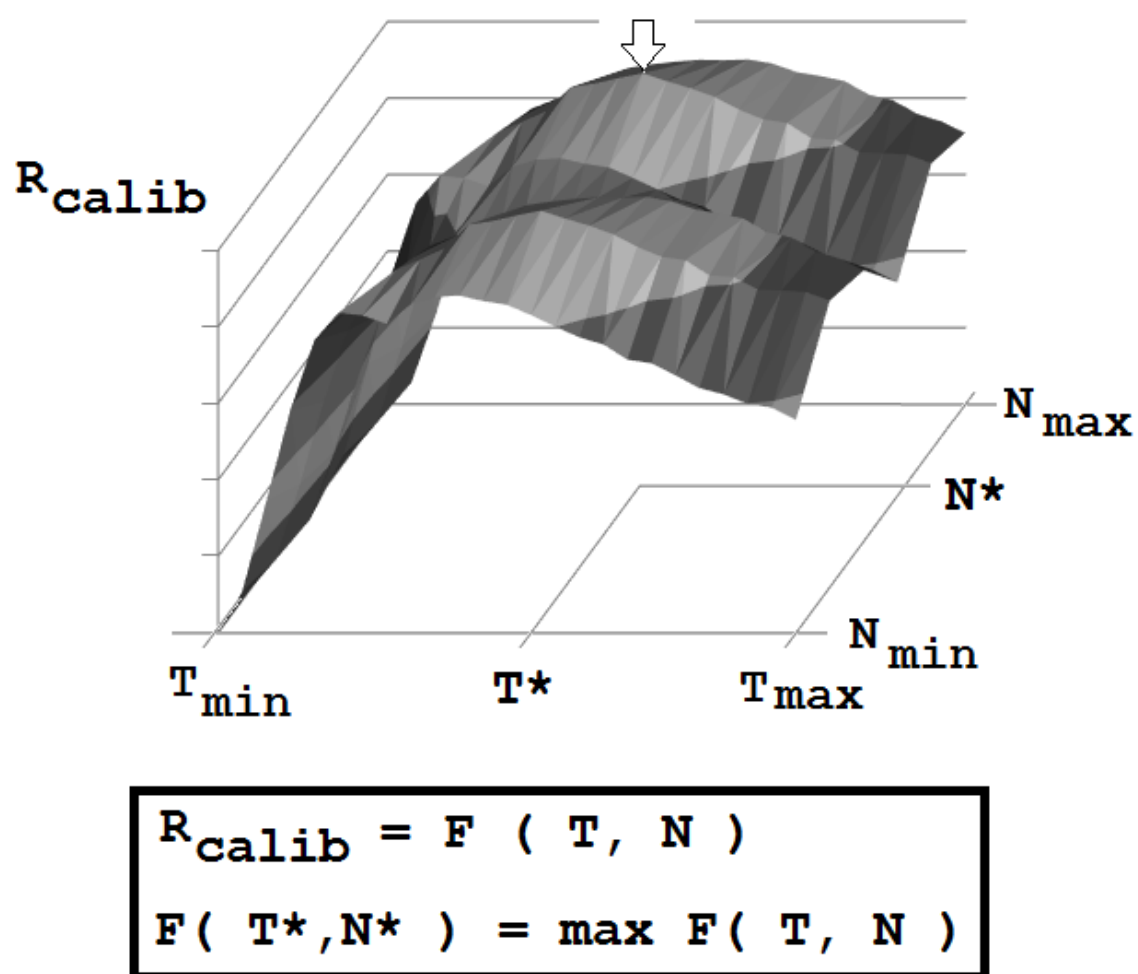


Figure 3

Graphical representation of the scheme of definition of the T^* and N^* for the CORAL models, based on the correlation coefficient (R_{calib}) between experimental and predicted values of endpoint for calibration set.

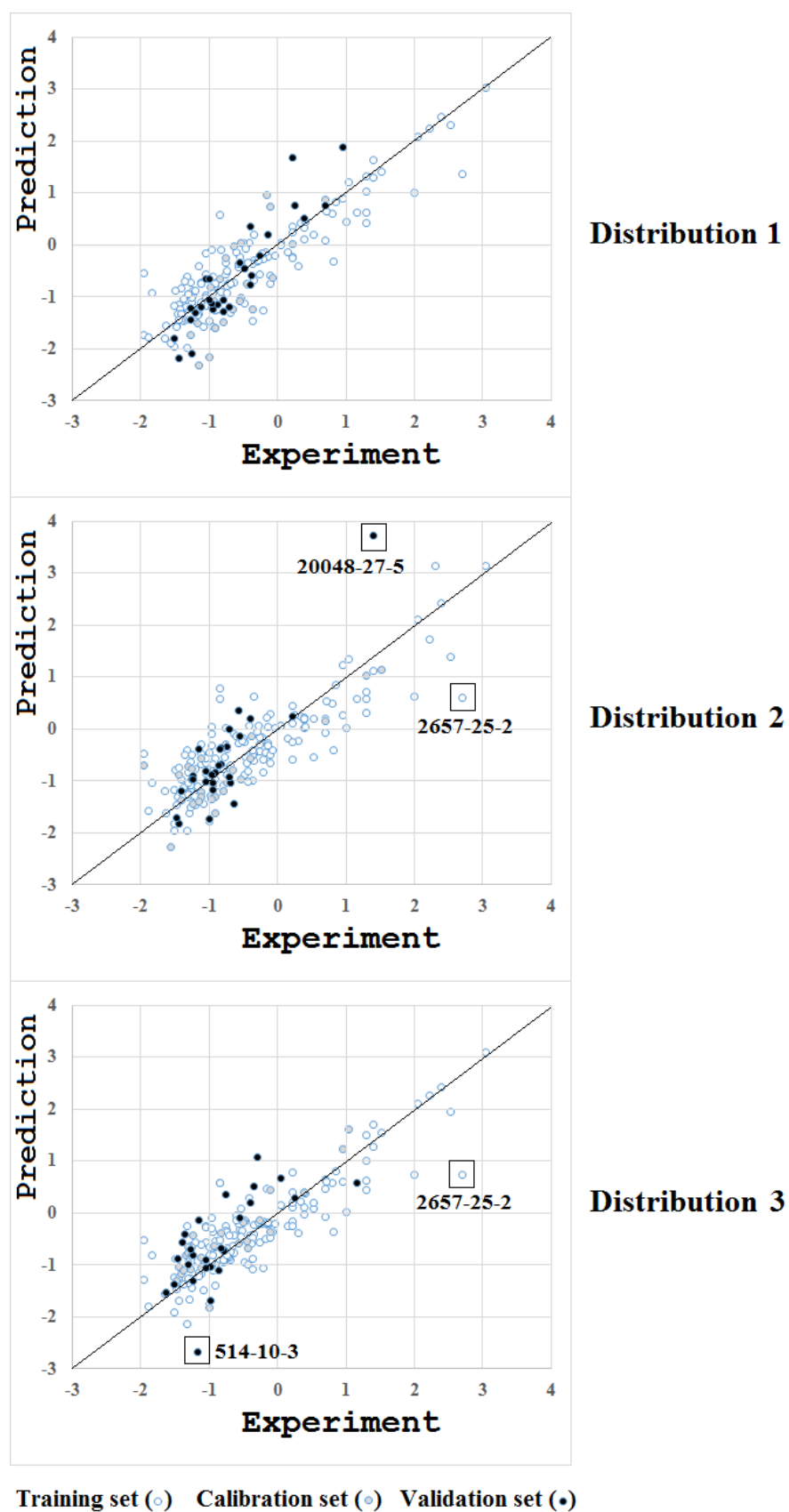


Figure 4

Graphical representation of the model based on hybrid descriptors. Outliers indicated by squares.

Table 1

Optimal descriptors used in this study

ID	Formula	Comment
1	$DCW(T^*, N^*) = CW(C5) + CW(C6) +$ $\sum CW(EC0_k) + \sum CW(EC1_k) + \sum CW(EC2_k) +$ $\sum CW(NNC_k) +$ $CW(NOSP) + CW(HALO) + CW(BOND) +$ $\sum CW(S_k)$	Hybrid descriptor calculated with correlation weights of SMILES attributes and correlation weights of graph invariants
2	$DCW(T^*, N^*) = \sum CW(EC0_k) + \sum CW(EC1_k) + \sum CW(EC2_k) +$ $\sum CW(NNC_k) +$ $CW(NOSP) + CW(HALO) + CW(BOND) +$ $\sum CW(S_k)$	Hybrid descriptor calculated with correlation weights of SMILES attributes and correlation weights of graph invariants, but without correlation weights of C5 and C6
3	$DCW(T^*, N^*) = CW(NOSP) + CW(HALO) + CW(BOND) +$ $\sum CW(S_k) + \sum CW(SS_k) + \sum CW(SSS_k)$	Optimal descriptor calculated with solely correlation weights of SMILES attributes
4	$DCW(T^*, N^*) = CW(C5) + CW(C6) +$ $\sum CW(EC0_k) + \sum CW(EC1_k) + \sum CW(EC2_k) +$ $\sum CW(PT2_k) + \sum CW(PT3_k) +$ $\sum CW(S2_k) + \sum CW(S3_k) +$ $\sum CW(NNC_k)$	Optimal descriptor calculated with solely correlation weights of graph invariants

Table 2

Definition of twelve symbols codes for BOND, NOSP, and HALO.

ID	Comment	1	2	3	4	5	6	7	8	9	10	11	12
1	Definition of BOND attribute					=	#	@					
		B	O	N	D	C1*	C2	C3	0	0	0	0	0
2	Definition of NOSP attribute					N	O	S	P				
		N	O	S	P	C1	C2	C3	C4	0	0	0	0
3	Definition of HALO attribute					F	Cl	Br	I				
		H	A	L	O	C1	C2	C3	C4	0	0	0	0
For instance, if SMILES is "O=C(/C=C/c1cccc1)c2ccc(O)cc2", then the definitions are													
1	Definition of BOND attribute					=	#	@					
		B	O	N	D	1	0	0	0	0	0	0	0
2	Definition of NOSP attribute					N	O	S	P				
		N	O	S	P	0	1	0	0	0	0	0	0
3	Definition of HALO attribute					F	Cl	Br	I				
		H	A	L	O	0	0	0	0	0	0	0	0

*) C1-C4 are (0,1) codes: if a feature is present (e.g. double bond '='; nitrogen atom 'N'; chlorine 'Cl', etc.) then code=1, if the feature is absent then code=0.

Table 3

Translation of SMILES attributes S_k , SS_k , and SSS_k , into code of twelve symbols
(SMILES = “O=C(/C=C/c1ccccc1)c2ccc(O)cc2”)

S_k	SS_k	SSS_k
O.....		
=.....	O...=.....	
C.....	C...=.....	O...=...C...
(.....	C...(.....	=...C...(...
/.....	/...(.....	C...(.../...
C.....	C.../.....	C.../...(...
=.....	C...=.....	=...C.../...
C.....	C...=.....	C...=...C...
/.....	C.../.....	=...C.../...
c.....	c.../.....	c.../...C...
1.....	c...1.....	1...c.../...
c.....	c...1.....	c...1...c...
c.....	c...c.....	c...c...1...
c.....	c...c.....	c...c...c...
c.....	c...c.....	c...c...c...
c.....	c...c.....	c...c...c...
1.....	c...1.....	c...c...1...
(.....	1...(.....	c...1...(...
c.....	c...(.....	c...(...1...
2.....	c...2.....	2...c...(...
c.....	c...2.....	c...2...c...
c.....	c...c.....	c...c...2...
c.....	c...c.....	c...c...c...
(.....	c...(.....	c...c...(...
O.....	O...(.....	c...(...O...
(.....	O...(.....	(...O...(...
c.....	c...(.....	c...(...O...
c.....	c...c.....	c...c...(...
2.....	c...2.....	c...c...2...

Table 4

Graph invariants, which can be involved to build up the CORAL model

No.	Invariants of hydrogen-suppressed graph (HSG)	Comment
1	$EC0_k = \sum_{(kj)edge} a_{kj}$	Vertex degree
2	$EC1_k = \sum_{(kj)edge} EC0_j$	Extended connectivity of the first order
3	$EC2_k = \sum_{(kj)edge} EC1_j$	Extended connectivity of the second order
4	$PT2_k = \sum_{(kj)edge} \sum_{(jl)edge, l \neq k} a_{kl}$	The number of paths of length 2 which starts from k-vertex
5	$PT3_k = \sum_{(kj)edge} \sum_{(jl)edge, l \neq k} \sum_{(lm)edge, m \neq j, m \neq k} a_{km}$	The number of paths of length 3 which starts from k-vertex
6	$S2_k = \sum_{(kj)edge} \sum_{(jl)edge, l \neq k} EC0_l$	Valence shell of the second order
7	$S3_k = \sum_{(kj)edge} \sum_{(jl)edge, l \neq k} \sum_{(lm)edge, m \neq j, m \neq k} EC0_m$	Valence shell of the third order
8	$NNC_k = N_{total} \times 100 + N_{carbon} \times 10 + N_{non-carbon}$	The nearest neighboring code, N_{total} is total number of neighbors of k-the vertex; N_{carbon} is number of neighbors vertices represent carbon atoms; and $N_{non-carbon}$ is number of neighbors vertices represent non-carbon atoms

Table 5
Examples of C5 and C6 for different molecular structures

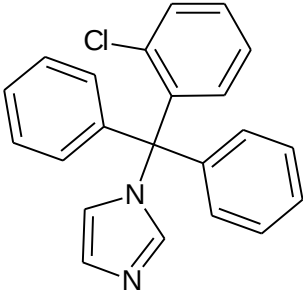
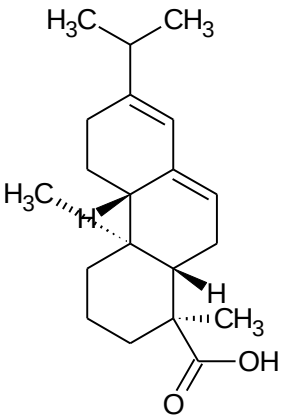
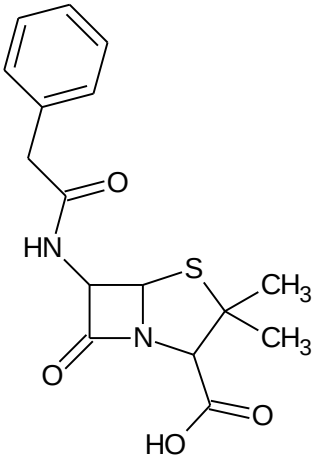
SMILES	Molecular structure	Representation of C5 and C6 by code of twelve symbols
<chem>Clc1ccccc1C(c2ccccc2)(c3ccccc3)n4ccnc4</chem>		C6...A..3... C5...AH.1...
<chem>CC(C)C=3CC[C@H]1C(=CC[C@H]2[C@@](C)(CCC[C@]12C)C(=O)O)C=3</chem>		C6.....3... C5.....0...
<chem>OC(=O)C2N3C(=O)C(NC(=O)Cc1ccccc1)C3SC2(C)C</chem>		C6...A..1... C5....H.1... C4....H.1...

Table 6

The statistical characteristics of the CORAL models for three different random distributions into the training, calibration, and validation sets. The statistical quality of a model for the validation set is examined as the criterion of predictive potential of a model. The best results are indicated by bold.

Distribution	Descriptor	Set	n*	r ²	q ²	RMSE	Y-test
1	1	Training	147	0.8269	0.8224	0.432	0.0077
		Calibration	28	0.8410	0.8223	1.01	0.0634
		Validation	29	0.8596	0.8088	0.489	0.0250
	2	Training	147	0.8276	0.8232	0.431	0.0073
		Calibration	28	0.8190	0.7813	1.40	0.0377
		Validation	29	0.8498	0.7491	0.543	0.0235
	3	Training	147	0.9167	0.9149	0.300	0.0093
		Calibration	28	0.3790	0.0985	0.691	0.0281
		Validation	29	0.2867	0.1189	0.747	0.0310
	4	Training	147	0.7255	0.7180	0.544	0.0110
		Calibration	28	0.3571	0.0119	0.835	0.0314
		Validation	29	0.5118	0.4529	0.493	0.0244
2	1	Training	142	0.7358	0.7277	0.530	0.0069
		Calibration	31	0.6203	0.5540	0.491	0.0260
		Validation	31	0.7959	0.7933	0.595	0.0423
	2	Training	142	0.7391	0.7308	0.527	0.0130
		Calibration	31	0.6021	0.5308	0.494	0.0184
		Validation	31	0.7828	0.6441	0.683	0.0413
	3	Training	142	0.9212	0.9191	0.290	0.0068
		Calibration	31	0.3690	0.2336	0.705	0.0191
		Validation	31	0.2904	0.1363	0.828	0.0381
	4	Training	142	0.8535	0.8498	0.395	0.0054
		Calibration	31	0.5810	0.5020	0.511	0.0679
		Validation	31	0.1260	0.0243	0.647	0.0417
3	1	Training	142	0.7796	0.7738	0.472	0.0078
		Calibration	31	0.7508	0.7061	0.431	0.0124
		Validation	31	0.7249	0.6745	0.761	0.0390
	2	Training	142	0.7588	0.7525	0.494	0.0078
		Calibration	31	0.6970	0.6423	0.433	0.0397
		Validation	31	0.6652	0.5916	0.842	0.0364
	3	Training	142	0.9347	0.9332	0.257	0.0059
		Calibration	31	0.1860	0.0516	0.937	0.0288
		Validation	31	0.4258	0.3084	1.07	0.0355
	4	Training	142	0.7117	0.7042	0.540	0.0062
		Calibration	31	0.4298	0.3192	0.600	0.0207
		Validation	31	0.5428	0.4583	0.772	0.0285

*) The n is the number of compounds in set; r² is determination coefficient; q² is leave-one-out cross-validated r²; RMSE is root mean squared error; Y-test is Y-randomization test (Zhang et al., 2006; Afantitis et al., 2008)

Table 7

Examples of the promoters for increase or decrease of skin sensitization (*pEC3*).

Code of a feature	Comment
<i>Promoters of <i>pEC3</i> increase</i>	
C5.....0...	Absence of five members rings
C6...A...1...	Presence of six members aromatic ring
=.....	Presence at least one double covalent bond
NNC-C...321.	Presence of nearest neighboring code for carbon atom equal to 321
EC2-C...8...	Presence of Morgan connectivity second order for carbon equal to 8
<i>Promoters of <i>pEC3</i> decrease</i>	
NNC-C...220.	Presence of nearest neighboring code for carbon atom equal to 220
NNC-C...211.	Presence of nearest neighboring code for carbon atom equal to 211
(.....	Branching of the molecular skeleton
EC0-C...1...	Presence of the vertex degree of carbon atom equal to 1
EC1-C...6...	Presence of Morgan connectivity first order for carbon equal to 6

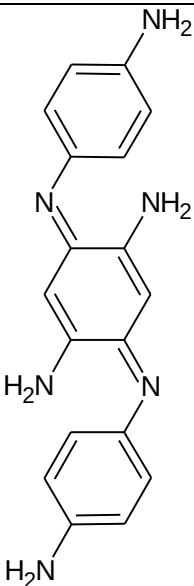
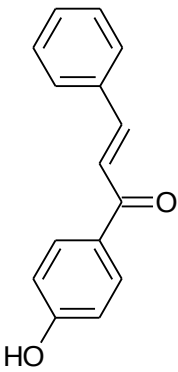
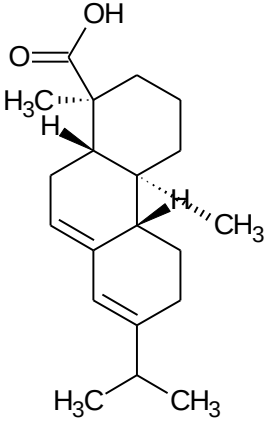
Table 8

Comparison of the statistical quality of QSAR model of skin sensitization

n	Correlation coefficient, r^2	Root-mean squared error (RMSE)	Reference
12	0.825	0.17	Grace et al., 2003
12	0.64	-	Promkatkaew et al., 2014
204	0.496	0.689	Dearden et al., 2015
204	0.734	0.555	In this work: Distribution 1
204	0.715	0.547	In this work: Distribution 2
204	0.714	0.526	In this work: Distribution 3

Table 9

Analysis of outliers (Figure 4)

Distribution	set	CAS	Structure	Comment
2	Validation	20048-27-5		High symmetry
2,3	Training	2657-25-2		Presence of group of rare molecular features
3	Validation	514-10-3		Presence of group of rare molecular features

Hybrid Optimal Descriptors as a Tool to Predict Skin Sensitization in accordance to OECD principles

Alla P. Toropova,* Andrey A. Toropov

IRCCS, Istituto di Ricerche Farmacologiche Mario Negri, 20156, Via La Masa 19, Milano, Italy

Supplementary materials

SM-1

The distribution (1, 2, and 3) into the training (+), calibration (#), and validation (*) sets; experimental $pEC3$ (exp.) and predicted $pEC3$ (pr.) (Dearden et al., 2015)

SM-2

The CORAL methods utilized to build up QSAR models for skin sensitization ($pEC3$)

SM-3

Correlation weights for hybrid descriptors utilized in the case of distribution 1

SM-4

Distribution 1 into the training (+), calibration (#), and validation (*) sets: format of each line is (i) Type of set (i.e. '+', '#', '*'); (ii) CAS number; (iii) SMILES; (iv) $pEC3$ value.

SM-1

The distribution (1, 2, and 3) into the training (+), calibration (#), and validation (*) sets; experimental $pEC3(\text{exp.})$ and predicted $pEC3(\text{pr.})$ (Dearden et al., 2015)

1	2	3	CAS	SMILES	$pEC3(\text{exp.})$	Distributio1 $pEC3(\text{pr.})$	Distributio2 $pEC3(\text{pr.})$	Distributio3 $pEC3(\text{pr.})$
-	+	+	2657-25-2	<chem>O=C(/C=C/c1ccccc1)c2ccc(O)cc2</chem>	2.6990	1.3617	0.6025	0.7381
+	+	#	1482-74-2	<chem>Oc2ccc(C(=O)/C=C/c1ccccc1)c(O)c2O</chem>	0.9590	1.8775	1.2376	1.2264
+	+	+	111-12-6	<chem>O=C(C#CCCCC)OC</chem>	0.3470	0.4140	0.2420	0.2143
-	+	*	1776-30-3	<chem>O=C(/C=C/c1ccccc1)c2ccc(O)cc2O</chem>	0.2520	0.7566	0.2617	0.2878
+	+	+	2953-00-7	<chem>CC(C)Oc1cc(ccc1O)\C=C\C</chem>	0.2220	-0.2246	-0.5908	-0.2505
#	+	+	1210-39-5	<chem>O=C/C=C(/c1ccccc1)c2ccccc2</chem>	0.2220	0.0184	0.1098	0.4072
+	-	+	3775-21-1	<chem>CC2=NS(=O)(=O)N=C2c1ccccc1</chem>	-0.1460	-0.2286	-0.2852	-0.2261
*	+	+	13041-12-8	<chem>Oc1c(C)cc(cc1OC)\C=C\C</chem>	-0.2040	-1.2544	-0.8376	-1.0699
+	+	+	100-43-6	<chem>C=Cc1ccncc1</chem>	-0.2040	-0.2771	-0.5897	-0.1525
*	+	+	29043-97-8	<chem>O=C1OC(C)(C)CC1=C</chem>	-0.2550	-0.5627	-0.4330	-0.5749
*	#	+	3913-71-1	<chem>CCCCCCC\C=C\C=O</chem>	-0.3980	-0.7622	-0.5567	-0.4214
*	*	*	111-80-8	<chem>O=C(C#CCCCC)OC</chem>	-0.3980	0.3560	0.1879	0.1929
+	+	+	577-71-9	<chem>O=[N+](O-)]c1cc(O)ccc1[N+](O-)]=O</chem>	-0.4150	0.0318	0.0641	0.2897
+	+	#	104-55-2	<chem>O=C/C=C/c1ccccc1</chem>	-0.4770	-0.4636	-0.2541	-0.1362
*	*	*	142-83-6	<chem>C\C=C\C=C\C=O</chem>	-0.5440	-0.3500	-0.1475	-0.0930
+	+	#	122-57-6	<chem>CC(=O)/C=C/c1ccccc1</chem>	-0.5680	-0.4388	-0.2862	-0.5788
+	+	+	533-75-5	<chem>OC1=CC=CC=CC1=O</chem>	-0.6330	-0.0200	-0.0901	-0.4327
#	*	+	21834-92-4	<chem>O=C/C(=C\C(C)(C)C)c1ccccc1</chem>	-0.6430	-0.7221	-1.4347	-0.7671
+	#	+	101-39-3	<chem>C\C(=C/c1ccccc1)C=O</chem>	-0.6530	-1.2430	-0.8029	-0.6318
#	*	+	6728-26-3	<chem>CCC\C=C\C=O</chem>	-0.7400	-0.5302	-0.3401	-0.3357
*	+	+	141-05-9	<chem>O=C(\C=C/C(=O)OCC)OCC</chem>	-0.7630	-0.3405	0.0266	-0.2579
+	+	+	116-26-7	<chem>CC=1C=CCC(C)(C)C=1C=O</chem>	-0.8750	-1.1490	-1.2145	-0.6549
#	*	#	2111-75-3	<chem>CC(=C)C1CC=C(C=O)CC1</chem>	-0.9080	-0.8019	-0.8459	-0.6499
+	#	*	NA-1	<chem>CC(O)(C=C)CC\C=C(/C)C=O</chem>	-0.9780	-1.0815	-1.2816	-1.0312
#	*	#	103-11-7	<chem>O=C(C=C)OCC(CCCC)CC</chem>	-1.0000	-1.4746	-1.7361	-1.8323
-	*	+	122-40-7	<chem>O=CC(\CCCC)=C/c1ccccc1</chem>	-1.0410	-0.9695	-0.8111	-0.8737
*	+	+	7492-44-6	<chem>O=CC(\CCCC)=C/c1ccccc1</chem>	-1.0410	-0.6527	-0.5413	-0.5714
+	+	*	101-86-0	<chem>O=CC(\CCCC)=C/c1ccccc1</chem>	-1.0410	-1.0275	-0.8652	-0.8951
*	*	*	141-32-2	<chem>C=CC(=O)OCCCC</chem>	-1.0410	-1.1720	-1.0260	-1.0587
+	+	+	6485-40-1	<chem>CC1=CC[C@H](CC1=O)C(C)=C</chem>	-1.1110	-1.1173	-0.3037	-0.8929
-	+	+	103-41-3	<chem>O=C(OCc1ccccc1)/C=C/c2ccccc2</chem>	-1.2650	-0.7210	-0.6269	-0.6314
+	+	+	104-54-1	<chem>OC\C=C/c1ccccc1</chem>	-1.3220	-0.6050	-0.4080	-0.2604
*	+	#	127-51-5	<chem>CC1(C)CCC=C(C)C1\C=C(/C)C(C)=O</chem>	-1.3380	-1.1947	-1.1949	-0.8218
-	#	#	140-88-5	<chem>C=CC(=O)OCC</chem>	-1.4470	-1.1511	-0.8719	-1.0480
-	*	+	97-90-5	<chem>O=C(OCCOC(=O)C(=C)C)C(=C)C</chem>	-1.4470	-2.1903	-1.8313	-1.6938
+	+	+	1565-94-2	<chem>C=C(C)C(=O)OCC(O)COc1ccc(cc1)C(C)(C)c2ccc(OCC(O)COC(=O)C(=C)C)cc2</chem>	-1.6530	-1.7978	-1.1926	-1.5458
+	#	+	80-62-6	<chem>C=C(C)C(=O)OC</chem>	-1.9540	-1.7408	-0.6985	-1.2797
+	*	+	20048-27-5	<chem>Nc1ccc(cc1)/N=C3/C=C(N)C(=N)c2ccc(N)cc2)/C=C3N</chem>	1.3980	1.6263	3.7168	1.7083
#	+	+	99-56-9	<chem>Nc1ccc(cc1N)[N+](O-)]=O</chem>	1.3010	0.6134	0.5811	0.4349
+	+	*	65235-31-6	<chem>O=[N+](O-)]c1cc(O)ccc1NCCO</chem>	1.1550	0.6219	0.5886	0.5775
+	+	+	123-31-9	<chem>Oc1ccc(O)cc1</chem>	0.9590	0.8949	0.2711	0.6030
*	+	+	106-50-3	<chem>Nc1ccc(N)cc1</chem>	0.7960	0.5933	0.4878	0.5707
+	+	+	95-70-5	<chem>Nc1ccc(N)c(C)c1</chem>	0.6990	0.0842	-0.0686	-0.0752
#	#	+	610-81-1	<chem>O=[N+](O-)]c1cc(O)ccc1N</chem>	0.6990	0.7487	0.1559	0.4681

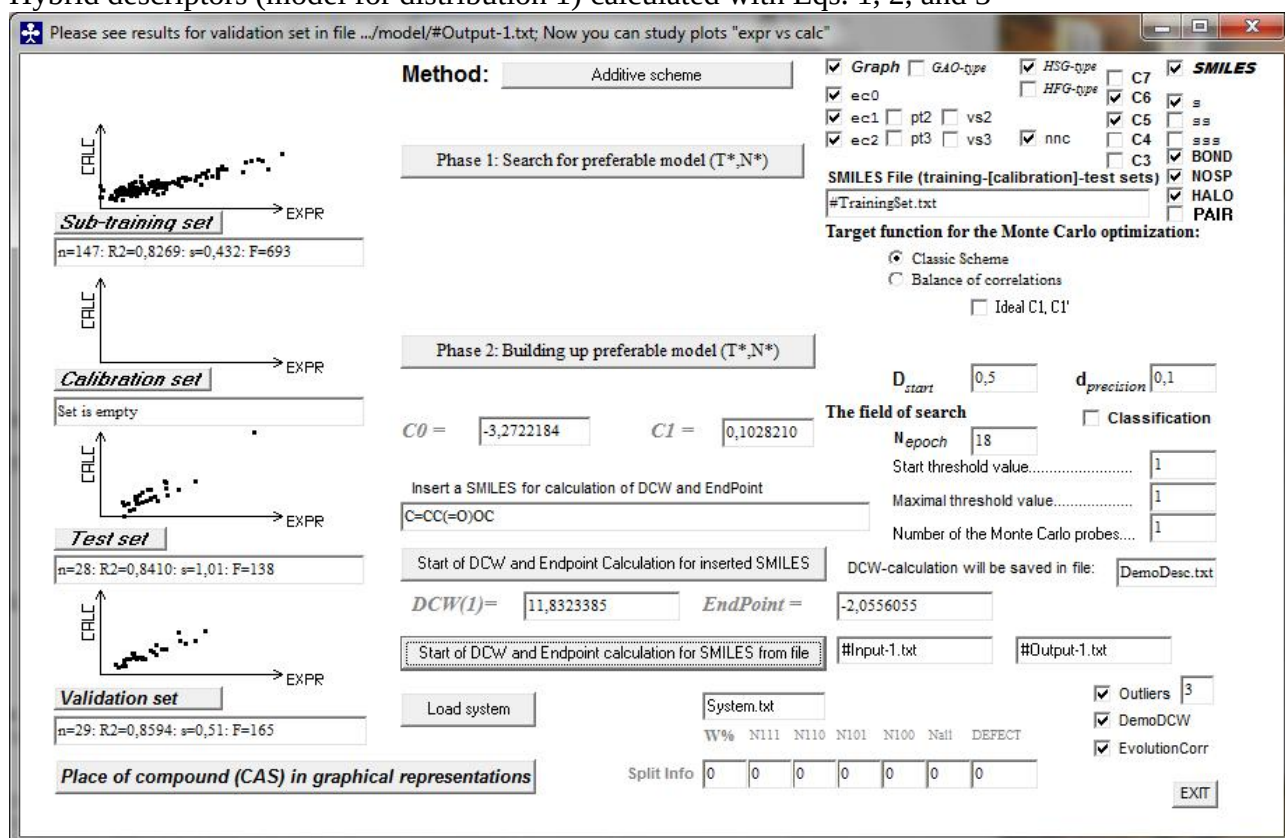
-	+	+	95-55-6	Nc1cccc1O	0.3980	0.5005	0.1786	0.1074
-	+	#	55302-96-0	Cc1ccc(NCCO)cc1O	0.3980	0.4828	0.2191	0.3896
*	+	+	5307-14-2	O=[N+][O-]c1cc(N)ccc1N	0.3980	0.3248	0.1907	0.3981
#	+	+	108-45-2	Nc1cccc(N)c1	0.3100	-0.4111	-0.2333	-0.3977
*	+	+	55658-55-4	CC1=CC[C@H](C/C1=N\O)C(C)=C	0.2220	0.3572	0.2357	0.1093
-	*	+	10897-60-1	Oc1ccc(CCO)cc1O	0.2220	1.6825	0.2350	0.7885
+	+	*	35691-65-7	N#CCCC(Br)(CBr)C#N	0.0460	0.0468	0.0607	0.6691
*	+	#	90-15-3	Oc2cccc1cccc12	-0.1140	0.7445	0.2945	0.4339
+	+	+	2835-99-6	Cc1cc(O)ccc1N	-0.1610	0.9516	0.2203	0.4590
-	+	*	2871-01-4	[O-][N+](=O)c1cc(N)ccc1NCCO	-0.3420	0.1980	0.6233	0.5075
*	+	#	591-27-5	Nc1cccc(O)c1	-0.5050	0.0455	-0.1759	-0.1767
#	+	+	154750-20-6	BrCC1CC(C)(C)OC1=O	-0.5560	-0.2288	-0.1505	-0.3233
-	#	+	93-51-6	Oc1ccc(C)cc1OC	-0.7630	-0.2548	-0.2931	-0.3317
#	+	*	105-13-5	COc1ccc(CO)cc1	-0.7710	-0.9695	-0.5446	-0.7151
-	*	#	2785-87-7	Oc1ccc(cc1OC)CCC	-0.8330	-0.0972	-0.3708	-0.3788
#	+	+	29705-39-3	O=[N+][O-]c1cc(ccc1N)N(CCO)CCO	-0.9140	-0.7413	-0.5636	-0.5391
+	+	+	97-53-0	Oc1ccc(cc1OC)CC=C	-1.1140	-0.7143	-0.4863	-0.6102
-	+	+	186743-25-9	Oc1cc(C)c(cc1OC)CC=C	-1.1140	-0.7465	-0.6799	-0.8588
+	*	*	186743-24-8	Oc1c(C)cc(cc1OC)CC=C	-1.2300	-1.3321	-0.9149	-1.3069
#	*	+	140-67-0	COc1ccc(CC=C)cc1	-1.2550	-2.0902	-0.9644	-1.4354
+	+	+	186743-26-0	COc1c(C)c(ccc1O)CC=C	-1.5050	-1.1625	-1.1797	-1.2517
+	+	+	111-30-8	O=CCCCC=O	1.0000	0.4472	0.0264	0.0087
*	+	+	57074-21-2	Clc1c(C)cc(O)c(C=O)c1O	0.3980	-0.4980	0.0491	-0.4647
#	+	+	50-00-0	C=O	0.2150	0.2361	0.4540	0.2066
-	+	#	579-07-7	O=C(C(C)=O)c1cccc1	-0.1140	-0.7503	-0.6588	-0.3592
*	+	+	107-22-2	O=CC=O	-0.1460	0.1867	0.0346	-0.2066
+	+	+	122-78-1	O=CCc1cccc1	-0.4770	-1.1398	-0.8897	-0.9402
*	+	*	93-53-8	CC(C=O)c1cccc1	-0.7990	-1.2930	-1.0798	-0.7144
*	*	*	112-45-8	C=CCCCCCCCC=O	-0.8330	-0.8979	-0.6882	-0.6799
+	#	-	167998-73-4	Cc1c(cc(C)c(C)c1C)C(=O)CC(C)=O	-0.9190	-1.2671	-1.6142	-1.4040
*	+	*	167998-76-7	CCc1ccc(cc1C(=O)CC(C)=O)CC	-0.9820	-0.8240	-1.3169	-1.6854
-	+	+	465-29-2	CC2(C)C1CCC2(C)C(=O)C1=O	-1.0000	-1.0456	-0.9420	-1.0633
+	+	+	110-41-8	CC(CCCCCCCC)C=O	-1.0000	-2.1593	-1.7749	-1.7576
*	+	-	55846-68-1	O=C(CC(=O)CCCC)c1cccc1	-1.0410	-0.1594	-0.6558	-0.2554
+	+	+	502-67-0	C/C(C)=C\CCC(\C)=C\CCC(\C)=C\C=O	-1.0790	-1.5868	-1.4361	-1.4939
*	#	#	5392-40-5	C\C(C)=C\CCC(\C)=C\C=O	-1.1140	-1.1852	-1.2277	-1.0547
#	#	#	56290-55-2	Cc1ccc(C)cc1C(=O)CC(C)=O	-1.1140	-0.7534	-0.5728	-0.8690
-	*	*	5406-12-2	Cc1ccc(CCC=O)cc1	-1.1460	-0.4056	-0.3816	-0.1440
#	#	*	31906-04-4	O=CC1CC=C(CCCC(C)(C)O)CC1	-1.2300	-0.8794	-1.4520	-0.8135
*	+	#	80-54-6	CC(C)(C)c1ccc(CC(C)C=O)cc1	-1.2790	-1.7377	-1.1892	-1.0744
+	+	+	613-29-6	CCCCN(CCCC)c1cccc1	-1.2920	-3.0092	-1.6297	-1.6630
*	+	+	326-06-7	O=C(CC(=O)C(F)(F)F)c1cccc1	-1.3010	-1.2779	-1.3061	-1.3632
+	+	*	2277-19-2	CC\C=C\CCCCC=O	-1.3620	-0.7042	-0.5026	-0.4000
*	+	-	13706-86-0	CC(=O)C(=O)CC(C)C	-1.4150	-0.8299	-0.9698	-1.1509
-	+	+	1118-71-4	CC(C)(C)C(=O)CC(=O)C(C)(C)C	-1.4310	-1.2464	-1.1899	-1.3241
+	+	*	6668-24-2	O=C(C(C)C(C)=O)c1cccc1	-1.4620	-1.3293	-1.3108	-0.8818
-	+	*	170928-69-5	Cc1c(cc(C)c(C)c1C)C(=O)CC(=O)OCC	-1.5190	-1.7942	-1.8179	-1.3823
-	+	+	107-75-5	O=CCC(C)CCCC(C)(C)O	-1.5190	-1.9562	-1.9550	-1.9255
-	+	+	97-96-1	CCC(CC)C=O	-1.8810	-1.7752	-1.5785	-1.8132
-	+	+	109-55-7	NCCCN(C)C	-0.3420	-0.2902	-0.4775	-0.2196
+	+	+	107-15-3	NCCN	-0.3420	-0.6696	-0.6187	-0.4021
*	+	*	111-40-0	NCCNCCN	-0.7630	-0.5610	-0.3371	0.3627
#	+	#	89-25-8	O=C2CC(C)=NN2c1cccc1	-0.9290	-1.0923	-0.8471	-0.6355
+	#	#	106-24-1	C\C(C)=C\CCC(\C)=C\CO	-1.4150	-1.3266	-1.3816	-1.1789

+	+	*	1086-00-6	ClCc4ccc1ccc2cccc3ccc4c1c23	2.3010	6.1748	3.1322	4.5580
+	+	+	26172-55-4	ClC1=CC(=O)N(C)S1	2.0460	2.0881	2.1117	1.7824
#	#	+	70-25-7	N=C(N[N+])(=O)[O-])N(C)N=O	1.5230	1.4077	1.1486	1.5407
+	#	+	684-93-5	NC(=O)N(C)N=O	1.3010	1.3216	1.0296	1.4946
+	+	+	57-57-8	O=C1CCO1	0.8240	-0.3238	-0.4000	-0.3627
-	+	+	100-39-0	BrCc1ccccc1	0.6990	0.8599	0.1895	0.6381
+	+	+	2374-65-4	O=S(=O)(CCCCCCCCCCCC)OC	0.4090	0.4432	0.0468	0.3802
#	+	+	55406-53-6	O=C(NCCCC)OCC#Cl	0.0460	-0.0027	0.0583	0.1495
-	+	+	759-73-9	NC(=O)N(CC)N=O	-0.0410	-0.1974	0.0079	-0.2079
-	+	+	1675-54-3	CC(C)(c2ccc(OCC1CO1)cc2)c4ccc(OCC3CO3)cc4	-0.1760	-0.1309	-0.4466	-0.4945
#	+	+	2634-33-5	O=C2NSc1ccccc12	-0.3620	-0.3743	-0.5068	-0.4722
#	+	+	112-82-3	BrCCCCCCCCCCCCCCCCC	-0.3620	-1.2453	-0.9885	-0.8732
#	+	+	118-58-1	Oc2cccc2C(=O)OCc1ccccc1	-0.4620	-0.0715	-0.5535	-0.1662
+	+	+	64-67-5	O=S(=O)(OCC)OCC	-0.5190	-0.3357	-0.2340	-0.3472
-	*	+	3508-00-7	BrCCCCCCCCCCCCCCCCC	-0.6810	-1.3033	-1.0426	-0.8946
+	*	+	629-72-1	BrCCCCCCCCCCCCCCCCC	-0.7080	-1.1873	-0.9343	-0.8518
-	+	+	137-26-8	CN(C)C(=S)SSC(=S)N(C)C	-0.7160	-0.7376	-0.8237	-0.8274
+	#	+	4276-49-7	BrCCCCCCCCCCCCCCCCCCCC	-0.7850	-1.4773	-1.2051	-0.9590
#	+	+	103-63-9	BrCCc1ccccc1	-0.7920	-1.0559	-0.7652	-0.9021
+	+	+	66-27-3	CS(=O)(=O)OC	-0.9080	-1.5961	-0.9714	-1.0090
+	#	+	6938-66-5	BrCCCCCCCCCCCCCCCCCCCCC CC	-0.9190	-1.5933	-1.3134	-1.0019
#	*	+	51323-71-8	CS(=O)(=O)OCCCCCCCCCCCCC	-0.9440	-1.2393	-1.1707	-0.8067
*	#	+	4860-03-1	CCCCCCCCCCCCCCCCCCCCCl	-0.9590	-1.1882	-1.3479	-1.0354
-	*	+	112-71-0	BrCCCCCCCCCCCCCCCCC	-0.9640	-1.1293	-0.8802	-0.8303
-	+	+	111-25-1	BrCCCCC	-1.0000	-0.6654	-0.4469	-0.6588
+	#	+	765-09-3	BrCCCCCCCCCCCCC	-1.0000	-1.0714	-0.8260	-0.8089
+	#	+	4292-19-7	CCCCCCCCCCCCCCI	-1.1140	-1.2060	-1.2998	-1.1700
#	+	+	3386-33-2	CCCCCCCCCCCCCCCCCCCCCl	-1.2040	-1.3042	-1.4562	-1.0782
-	*	#	120-51-4	O=C(OCc1ccccc1)c2ccccc2	-1.2300	-0.9765	-0.9812	-0.4299
#	+	+	73367-80-3	O=C(O)CCCCCCCCCCCCBr	-1.2550	-1.1212	-0.4902	-0.9227
*	+	+	544-77-4	CCCCCCCCCCCCCCCCCCI	-1.2790	-1.4380	-1.5165	-1.2557
+	#	#	693-67-4	BrCCCCCCCCCCCCC	-1.3010	-0.9554	-0.7177	-0.7660
+	+	*	2425-54-9	CCCCCCCCCCCCCCCCCCI	-1.3010	-1.0723	-1.2395	-0.9925
-	+	+	74036-97-8	CCCCCCCC(Br)CCCCC	-1.3220	-1.9829	-1.9584	-2.1401
-	+	*	35709-09-2	CS(=O)(=O)OCCCCCCCC\C=C\C CCCCCCC	-1.3980	-1.3682	-1.2644	-0.5742
+	+	+	2426-08-6	CCCCOCC1CO1	-1.4910	-0.8828	-0.8020	-0.8908
+	-	+	50-32-8	c1cc2c3cccc4cccc5ccc(cc2cc1)c3c4 5	3.0460	3.0263	3.1284	3.1031
*	+	+	57-97-6	Cc2c1c4ccccc4ccc1c(C)c3ccccc23	2.2220	2.2344	1.7213	2.2719
+	+	+	15646-46-5	O=C1OC(=N/C1=C\OCC)c2ccccc2	2.5230	2.3031	1.3807	1.9377
+	+	+	3326-32-7	Oc4cc5Oc1cc(O)ccc1C3(OC(=O)c 2cc(\N=C=S)ccc23)c5cc4	0.8540	0.8355	0.8521	0.7959
*	+	+	525-76-8	CC2=Nc1ccccc1C(=O)O2	0.1550	-0.1332	-0.1795	-0.2520
*	+	+	176665-02-4	CC1(C)N=C(CCCCC)OC1=O	-0.1140	-0.5822	-0.4943	-0.4701
*	+	+	149-30-4	Sc1nc2ccccc2s1	-0.2300	-0.1665	-0.1685	-0.1970
-	+	#	176664-99-6	CC1(C)N=C(CCCC)OC1=O	-0.2550	-0.2074	-0.1703	-0.1464
+	+	+	764-85-2	ClC(=O)CCCCCCCC	-0.2550	-0.2097	-0.2911	-0.5488
#	+	*	NA-2	OS(=O)(=O)c1ccccc1OC(=O)C(C) CCCCCCCCCCCCCCCCC	-0.3010	-0.3146	-0.0531	1.0743
-	+	+	36727-29-4	CC(CC(C)(C)C)CC(Cl)=O	-0.4310	-0.8487	-1.0073	-1.1885
#	+	#	176665-04-6	CC1(C)N=C(CCCCCCCCC)OC1= O	-0.4470	-0.7561	-0.6567	-0.5344
+	*	-	17369-59-4	CC\C=C1/OC(=O)c2ccccc12	-0.5680	-0.3681	0.3439	-0.2083
#	-	+	119-84-6	O=C1CCc2ccccc2O1	-0.7480	-0.4905	-0.3914	-0.3632
+	+	+	552-30-7	O=C(O)c1ccc2C(=O)OC(=O)c2c1	-0.9640	-0.8788	-0.3265	-0.9308
#	+	+	176665-06-8	CC1(C)N=C(CCCCCCCCCC)OC	-1.2040	-0.8721	-0.7650	-0.5773

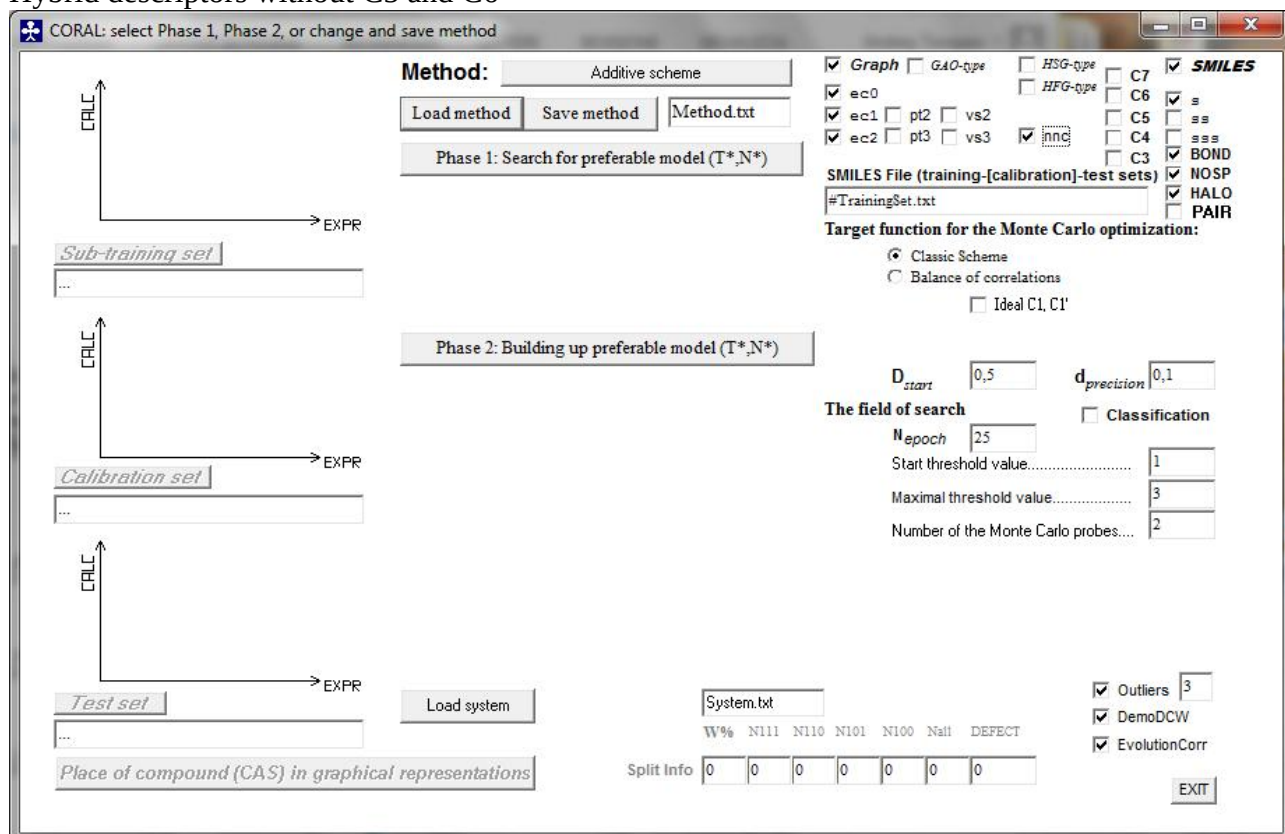
				1=O				
*	#	#	176665-09-1	CC1(C)N=C(CCCCCCCCCCCCCC CC)OC1=O	-1.2550	-1.1041	-0.9817	-0.6631
+	+	*	176665-11-5	CC1(C)N=C(CCCCCCCCCCCCCC CCCC)OC1=O	-1.2790	-1.2201	-1.0900	-0.7059
+	+	#	93-99-2	O=C(Oc1ccccc1)c2ccccc2	-1.3010	-1.4324	-0.8317	-0.6620
-	+	+	39236-46-9	O=C2NC(NC(=O)NCNC(=O)NC1 NC(=O)N(CO)C1=O)C(=O)N2CO	-1.3800	-1.4619	-1.2950	-1.3815
+	+	+	61-33-6	OC(=O)C2N3C(=O)C(NC(=O)Cc1 ccccc1)C3SC2(C)C	-1.4770	-1.5612	-1.4555	-1.4719
#	*	+	4638-48-6	Oc2ccc(Cl)cc2C(=O)Nc1ccccc1	-0.6990	-0.7760	0.3014	-0.9981
+	+	+	99-83-2	CC=1C=CC(CC=1)C(C)C	-0.7320	-0.7951	-0.8833	-0.6850
*	*	*	NA-3	CC1=CC[C@H](CC1=C)C(C)=C	-0.8630	-1.2402	-0.6969	-1.0961
*	#	+	2136-71-2	OCCOCCCCCCCCCCCCCCCC	-0.9440	-1.5781	-1.1168	-1.1114
+	*	+	99-86-5	CC(C)C1=CC=C(C)CC1	-0.9490	-0.9365	-1.0373	-0.8092
*	+	+	32388-55-9	CC(=O)C=1C[C@]23C[C@@H](C =1C)C(C)(C)[C@@H]3CC[C@H] 2C	-1.1430	-2.3300	-1.2201	-1.1483
*	+	*	514-10-3	CC(C)C=3CC[C@H]1C(=CC[C@ H]2[C@@](C)(CCC[C@]12C)C(= O)O)C=3	-1.1760	-1.5073	-1.2023	-2.6899
-	*	#	78-70-6	CC(O)(C=C)CC\C=C(\C)C	-1.4770	-1.5682	-1.7102	-1.3171
-	+	+	5989-27-5	CC1=CCC(CC1)C(C)=C	-1.8390	-0.9365	-1.0373	-0.8092
#	+	+	1897-45-6	Clc1c(C#N)c(Cl)c(C#N)c(Cl)c1Cl	2.3980	2.4727	2.4187	2.0047
#	+	+	97-00-7	O=[N+][([O-])]c1cc(ccc1Cl)[N+][([O-])=O	1.3010	0.4163	0.3058	0.6180
*	+	#	108-77-0	Clc1nc(Cl)nc(Cl)n1	1.0460	1.1991	1.3356	1.6223
+	+	+	87-86-5	Clc1c(O)c(Cl)c(Cl)c(Cl)c1Cl	-1.3010	-1.3295	-1.2682	-1.2746
+	+	+	23593-75-1	Clc1ccccc1C(c2ccccc2)(c3ccccc3)n 4ccnc4	-0.6810	-0.7531	-0.6920	-0.7860
-	+	*	106-22-9	C\C(C)=C\CCC(C)CCO	-1.6380	-1.5457	-1.6128	-1.5401
+	+	+	106-51-4	O=C1C=CC(=O)C=C1	2.0040	1.0116	0.6318	0.7250
+	+	+	62-53-3	Nc1ccccc1	-1.9490	-0.5475	-0.4680	-0.5275
#	#	+	555-10-2	C=C1C=CC(CC1)C(C)C	-0.7480	-0.9365	-1.0373	-0.8092
+	*	*	NA-4	CC1(C)N=C(CCCCCCCCCCCCCC CCCCC)OC1=O	-1.4150	-1.3361	-1.1983	-0.7488
#	+	+	112-67-4	ClC(=O)CCCCCCCCCCCCCCCC	-0.9440	-0.6157	-0.6702	-0.8098
-	+	#	57077-36-8	ClC(=O)CCCCC(C)C	-0.4310	-0.3503	-0.8159	-1.1752
+	+	+	1154-59-2	Clc2ccc(NC(=O)c1cc(Cl)cc(Cl)c1O)cc2Cl	1.3980	1.3003	1.1275	0.7116
#	+	#	4282-42-2	CCCCCCCCCI	-1.3800	-1.0321	-1.1374	-1.1056
#	#	#	143-15-7	BrCCCCCCCCCCCCC	-1.2550	-1.0134	-0.7718	-0.7874
-	#	+	19218-94-1	CCCCCCCCCCCCCCI	-1.1460	-1.3220	-1.4082	-1.2129
-	+	+	112-89-0	BrCCCCCCCCCCCCCCCC	-1.1760	-1.3613	-1.0968	-0.9161
*	*	+	3344-77-2	BrCCCCCCCCCCCCCO	-0.8390	-0.6464	-0.3902	-0.5298
-	#	+	10520-81-7	O=C(O)C(Br)CCCCCCCCCCCC	-0.5310	-1.0071	-0.9681	-0.4644
+	+	+	2682-20-4	O=C1C=CSN1C	-0.2790	-0.2922	0.0702	-0.1905
+	+	+	100-11-8	O=[N+][([O-])c1ccc(CBr)cc1	1.3010	1.0309	0.7036	1.0142
+	+	+	77-78-1	O=S(=O)(OC)OC	0.7210	0.6535	0.5376	0.5926
-	#	#	620159-84-4	CCC(C)(C)C1CCC(CC1)CC=O	-1.5680	-1.8849	-2.2633	-2.3713
+	+	*	1205-17-0	O=CC(C)Cc1ccc2OCOc2c1	-1.2150	-1.2677	-1.2837	-1.2646
-	+	+	35578-47-3	O=C(c1ccc(Br)cc1)C(=O)c2ccc(Br) cc2	-1.3120	-0.9378	-1.0298	-0.9460
#	*	*	103-95-7	CC(C)c1ccc(CC(C)C=O)cc1	-1.3420	-1.4348	-1.4117	-0.8986
+	+	+	431-03-8	CC(=O)C(C)=O	-1.0410	-0.5577	-0.5596	-0.2716
-	#	+	600-22-6	CC(=O)C(=O)OC	-0.3800	-0.5831	-0.1351	-0.5854
#	+	-	97-54-1	Oc1ccc(cc1OC)\C=C\C	-0.0790	-0.6366	-0.4089	-0.3732
-	+	+	818-61-1	C=CC(=O)OCCO	-0.1460	0.1998	-0.3123	-0.1750
+	+	+	104-46-1	COc1ccc(/C=C/C)cc1	-0.3620	-1.4733	-0.4096	-1.0797
+	+	+	2050-14-8	Oc2ccccc2/N=N/c1ccccc1O	-1.4460	-1.4489	-0.7540	-1.2518

-	#	+	526-37-4	<chem>O=Cc1c(cc(C)cc1O)O</chem>	0.2220	-0.2621	-0.6686	-0.1716
-	+	+	6358-09-4	<chem>Nc1cc(cc(Cl)c1O)[N+](=[O-])=O</chem>	-0.8360	0.5749	0.7836	0.5694
-	*	+	2835-95-2	<chem>Nc1cc(O)c(C)cc1</chem>	-0.5310	0.0336	-0.1389	0.0819
+	+	+	1166-52-5	<chem>Oc1cc(cc(O)c1O)C(=O)OCCCCC</chem> <chem>CCCCC</chem>	0.5230	0.2053	-0.5489	0.0829
#	#	#	186743-29-3	<chem>COc1c(C)c(ccc1O)\C=C\C</chem>	-0.5560	-1.0848	-1.1023	-1.0147
+	*	#	96-33-3	<chem>C=CC(=O)OC</chem>	-1.3010	-1.7153	-0.9876	-1.4147
+	#	+	5910-85-0	<chem>CC\C=C/C=C\C=O</chem>	-0.6020	-0.1383	0.0371	-0.0551
+	+	+	104-27-8	<chem>COc1ccc(/C=C/C(=O)CC)cc1</chem>	-0.9680	-0.0969	0.0450	-0.3085

The CORAL methods utilized to build up QSAR models for skin sensitization ($pEC3$) Hybrid descriptors (model for distribution 1) calculated with Eqs. 1, 2, and 3



Hybrid descriptors without C5 and C6



Super HSG descriptors

CORAL: select Phase 1, Phase 2, or change and save method

Method: Additive scheme

Load method Save method Method.txt

Phase 1: Search for preferable model (T*,N*)

Phase 2: Building up preferable model (T*,N*)

Sub-training set

Calibration set

Test set

Place of compound (CAS) in graphical representations

Load system

System.txt

W% N111 N110 N101 N100 Nall DEFECT

Split Info

Graph ☒ GAO-type ☐ HSG-type ☒ HFG-type ☐ C7 ☐ SMILES

☒ ec0 ☐ pt2 ☒ vs2 ☐ C6 ☐ s

☒ ec1 ☒ pt3 ☒ vs3 ☐ C5 ☐ ss

☒ ec2 ☐ nnc ☐ C4 ☐ sss

☐ C3 ☐ BOND

SMILES File (training-[calibration]-test sets)

#TrainingSet.txt

Target function for the Monte Carlo optimization:

☒ Classic Scheme

☐ Balance of correlations

☐ Ideal C1, C1'

D_{start} 0,5 d_{precision} 0,1

The field of search

N_{epoch} 25

Start threshold value..... 1

Maximal threshold value..... 3

Number of the Monte Carlo probes.... 2

☐ Classification

☒ Outliers 3

☒ DemoDCW

☒ EvolutionCorr

EXIT

Super SMILES descriptors

CORAL: select Phase 1, Phase 2, or change and save method

Method: Additive scheme

Load method Save method Method.txt

Phase 1: Search for preferable model (T*,N*)

Phase 2: Building up preferable model (T*,N*)

Sub-training set

Calibration set

Test set

Place of compound (CAS) in graphical representations

Load system

System.txt

W% N111 N110 N101 N100 Nall DEFECT

Split Info

Graph ☐ GAO-type ☐ HSG-type ☒ SMILES

☐ ec0 ☐ pt2 ☐ vs2 ☐ C6 ☒ s

☐ ec1 ☐ pt3 ☐ vs3 ☐ C5 ☒ ss

☐ ec2 ☐ nnc ☐ C4 ☒ sss

☐ C3 ☒ BOND

SMILES File (training-[calibration]-test sets)

#TrainingSet.txt

Target function for the Monte Carlo optimization:

☒ Classic Scheme

☐ Balance of correlations

☐ Ideal C1, C1'

D_{start} 0,5 d_{precision} 0,1

The field of search

N_{epoch} 25

Start threshold value..... 1

Maximal threshold value..... 3

Number of the Monte Carlo probes.... 2

☐ Classification

☒ Outliers 3

☒ DemoDCW

☒ EvolutionCorr

EXIT

SM-3

Correlation weights for hybrid descriptors utilized in the case of distribution 1

SMILES attributes and graph invariants SGk	Correlation weight of SGk	Prevalence in training set	Prevalence in calibration set	D(SG)
#.....	8.68360	4	0	1.0000
(.....	-1.99768	118	19	0.0009
+.....	-1.06515	10	0	1.0000
-.....	0.49699	10	0	1.0000
/.....	2.99994	16	3	0.0001
1.....	0.80898	98	19	0.0001
2.....	-0.50312	27	7	0.0020
3.....	-1.93946	7	3	0.0060
4.....	5.62601	5	1	0.0003
5.....	1.05781	2	0	1.0000
=.....	1.37538	114	17	0.0013
@.....	-0.74818	3	2	0.0102
@@.....	0.0	0	2	0.0000
C.....	0.12043	132	27	0.0004
C5.....0...	10.56604	126	25	0.0002
C5.....2...	1.94000	1	1	0.0145
C5....H.1...	14.30947	12	2	0.0007
C5....H.2...	3.75314	1	0	1.0000
C5...AH.1...	-7.50259	7	0	1.0000
C6.....0...	4.93271	61	11	0.0003
C6.....1...	4.12627	13	1	0.0038
C6.....3...	0.0	0	1	0.0000
C6...A..1...	4.06649	54	11	0.0004
C6...A..2...	1.87710	10	3	0.0030
C6...A..3...	-6.50360	2	0	1.0000
C6...A..4...	8.37627	1	1	0.0145
C6...A..5...	5.75267	1	0	1.0000
C6...AH.1...	14.25272	2	0	1.0000
C6...AH.2...	4.37070	2	0	1.0000
C6...AH.4...	-0.25145	1	0	1.0000
BOND00000000	5.75211	31	11	0.0043
BOND01000000	5.37661	2	0	1.0000
BOND10000000	-1.62488	109	15	0.0017
BOND10100000	-2.25086	3	2	0.0102
BOND11000000	11.87461	2	0	1.0000
F.....	-0.93446	1	0	1.0000
EC0-C...1...	-1.93468	103	20	0.0001
EC0-C...2...	-0.56262	138	27	0.0002
EC0-C...3...	-1.25261	123	22	0.0004
EC0-C...4...	1.68826	22	5	0.0011
EC0-F...1...	-0.62671	1	0	1.0000
EC0-Br..1...	-3.18295	10	6	0.0091
EC0-I...1...	2.00293	4	0	1.0000
EC0-Cl..1...	-0.56471	15	1	0.0041

EC0-N...1...	-0.24808	19	2	0.0028
EC0-N...2...	-1.87050	29	2	0.0041
EC0-N...3...	-2.37715	19	2	0.0028
EC0-O...1...	0.49885	112	22	0.0002
EC0-O...2...	-0.80941	45	6	0.0018
EC0-S...1...	4.62245	3	0	1.0000
EC0-S...2...	-0.25264	5	0	1.0000
EC0-S...4...	5.30831	6	1	0.0007
EC0-s...2...	1.19019	1	0	1.0000
EC1-C...1...	2.62324	1	0	1.0000
EC1-C...10..	2.62483	2	2	0.0145
EC1-C...11..	-0.99640	1	0	1.0000
EC1-C...12..	5.25338	1	0	1.0000
EC1-C...2...	-3.24692	55	13	0.0013
EC1-C...3...	-0.93405	101	22	0.0008
EC1-C...4...	-0.49823	105	22	0.0006
EC1-C...5...	0.87324	113	22	0.0001
EC1-C...6...	-0.50395	91	17	0.0001
EC1-C...7...	1.25465	67	13	0.0001
EC1-C...8...	-1.99549	29	7	0.0015
EC1-C...9...	0.50122	7	3	0.0060
EC1-F...4...	-0.05805	1	0	1.0000
EC1-Br..2...	6.68761	8	5	0.0095
EC1-Br..3...	0.93660	2	1	0.0074
EC1-Br..4...	-0.56241	1	0	1.0000
EC1-I...2...	1.12169	4	0	1.0000
EC1-Cl..2...	-0.18537	2	1	0.0074
EC1-Cl..3...	0.87553	13	0	1.0000
EC1-N...2...	1.37523	5	0	1.0000
EC1-N...3...	3.00430	14	2	0.0015
EC1-N...4...	6.75413	8	0	1.0000
EC1-N...5...	-4.81506	16	0	1.0000
EC1-N...6...	1.06748	13	1	0.0038
EC1-N...7...	-4.12814	8	2	0.0017
EC1-N...8...	3.00105	2	1	0.0074
EC1-N...9...	-3.87292	1	0	1.0000
EC1-O...1...	3.74586	1	0	1.0000
EC1-O...2...	0.68382	34	7	0.0005
EC1-O...3...	1.68400	81	14	0.0005
EC1-O...4...	-1.62068	23	5	0.0008
EC1-O...5...	3.81601	16	1	0.0043
EC1-O...6...	3.19210	15	2	0.0018
EC1-O...7...	1.62134	3	0	1.0000
EC1-S...2...	1.49750	1	0	1.0000
EC1-S...3...	4.87248	2	0	1.0000
EC1-S...5...	-1.37919	4	1	0.0017
EC1-S...6...	8.37061	6	0	1.0000
EC1-S...7...	-0.62374	1	0	1.0000
EC1-s...6...	-0.74922	1	0	1.0000
EC2-C...1...	2.25216	1	0	1.0000
EC2-C...10..	0.75335	69	14	0.0004
EC2-C...11..	2.00386	66	15	0.0011
EC2-C...12..	3.68993	54	10	0.0002
EC2-C...13..	-0.31278	56	6	0.0027
EC2-C...14..	-1.94007	35	8	0.0011

EC2-C...15..	0.50434	54	8	0.0013
EC2-C...16..	4.49877	46	9	0.0002
EC2-C...17..	0.62320	32	5	0.0011
EC2-C...18..	2.18631	32	6	0.0001
EC2-C...19..	5.25209	16	5	0.0033
EC2-C...20..	1.00187	9	3	0.0038
EC2-C...21..	5.74599	5	2	0.0053
EC2-C...22..	-2.56012	9	2	0.0009
EC2-C...23..	-0.87604	5	3	0.0091
EC2-C...24..	-3.62429	5	1	0.0003
EC2-C...25..	-0.75064	1	0	1.0000
EC2-C...26..	0.0	0	1	0.0000
EC2-C...27..	1.43410	1	1	0.0145
EC2-C...3...	3.81106	46	9	0.0002
EC2-C...29..	0.0	0	1	0.0000
EC2-C...31..	0.0	0	1	0.0000
EC2-C...33..	5.49916	1	0	1.0000
EC2-C...34..	1.49722	1	0	1.0000
EC2-C...4...	2.75124	29	4	0.0016
EC2-C...5...	-0.43415	34	10	0.0029
EC2-C...6...	2.62413	64	12	0.0001
EC2-C...7...	-0.50354	63	14	0.0009
EC2-C...8...	0.44209	70	17	0.0015
EC2-C...9...	-1.12910	62	16	0.0019
EC2-F...6...	-0.37381	1	0	1.0000
EC2-Br...3...	-5.49558	5	4	0.0121
EC2-Br...4...	13.43618	2	1	0.0074
EC2-Br...5...	-2.62652	3	0	1.0000
EC2-Br...6...	0.0	0	1	0.0000
EC2-Br...7...	1.00225	1	0	1.0000
EC2-I...3...	-0.80874	4	0	1.0000
EC2-Cl...3...	2.12185	2	0	1.0000
EC2-Cl...4...	13.56400	3	1	0.0038
EC2-Cl...5...	2.00383	4	0	1.0000
EC2-Cl...6...	6.55805	4	0	1.0000
EC2-Cl...7...	-1.18316	3	0	1.0000
EC2-N...10..	-0.50369	3	0	1.0000
EC2-N...11..	7.24871	4	0	1.0000
EC2-N...12..	-1.87161	10	0	1.0000
EC2-N...13..	-2.12723	14	2	0.0015
EC2-N...14..	3.81162	9	0	1.0000
EC2-N...15..	5.25422	1	0	1.0000
EC2-N...17..	-6.62125	2	1	0.0074
EC2-N...19..	0.0	0	1	0.0000
EC2-N...22..	5.37290	1	0	1.0000
EC2-N...25..	-0.56255	1	0	1.0000
EC2-N...3...	0.62062	4	0	1.0000
EC2-N...4...	8.74679	1	0	1.0000
EC2-N...5...	1.12740	11	1	0.0033
EC2-N...6...	5.18280	7	1	0.0015
EC2-N...8...	1.12672	4	0	1.0000
EC2-N...9...	3.43737	6	0	1.0000
EC2-O...1...	2.87486	1	0	1.0000
EC2-O...10..	1.81108	4	1	0.0017
EC2-O...11..	3.37183	8	0	1.0000

EC2-0...12..	-7.99722	4	1	0.0017
EC2-0...13..	11.18761	8	1	0.0021
EC2-0...14..	2.25325	1	0	1.0000
EC2-0...17..	-0.74774	1	0	1.0000
EC2-0...3...	5.19197	18	3	0.0007
EC2-0...4...	0.12780	23	5	0.0008
EC2-0...5...	1.31719	39	4	0.0028
EC2-0...6...	0.44021	40	11	0.0024
EC2-0...7...	-1.12459	21	2	0.0031
EC2-0...8...	1.43441	11	0	1.0000
EC2-0...9...	1.74821	11	2	0.0003
EC2-S...10..	-3.00029	1	0	1.0000
EC2-S...11..	10.62986	2	0	1.0000
EC2-S...12..	10.00012	1	0	1.0000
EC2-S...15..	1.43421	1	0	1.0000
EC2-S...17..	0.0	0	1	0.0000
EC2-S...18..	5.37712	2	0	1.0000
EC2-S...19..	15.87828	1	0	1.0000
EC2-S...20..	-3.49643	1	0	1.0000
EC2-S...21..	-2.75343	1	0	1.0000
EC2-S...22..	10.00136	1	0	1.0000
EC2-S...3...	-1.43492	1	0	1.0000
EC2-S...5...	0.94173	1	0	1.0000
EC2-S...6...	2.44086	1	0	1.0000
EC2-s...12..	-0.62801	1	0	1.0000
H.....	-2.00033	3	2	0.0102
Br.....	-1.62784	10	6	0.0091
I.....	2.43695	4	0	1.0000
Cl.....	0.31667	15	1	0.0041
HAL000000000	2.99720	117	21	0.0003
HAL000010000	-1.49747	4	0	1.0000
HAL000100000	10.87734	10	6	0.0091
HAL001000000	1.37143	15	1	0.0041
HAL010000000	2.37018	1	0	1.0000
N.....	-2.05819	44	5	0.0025
O.....	1.62937	115	22	0.0000
NNC-C...101.	-0.74998	23	4	0.0005
NNC-C...110.	-0.74942	92	19	0.0005
NNC-C...202.	2.99917	7	0	1.0000
NNC-C...211.	-3.37491	67	13	0.0001
NNC-C...220.	-0.06567	129	26	0.0003
NNC-C...303.	-1.56198	11	0	1.0000
NNC-C...312.	-1.30966	34	5	0.0014
NNC-C...321.	0.25150	69	14	0.0004
NNC-C...330.	1.62073	74	17	0.0011
NNC-C...413.	-0.50343	1	0	1.0000
NNC-C...431.	-0.31016	14	2	0.0015
NNC-C...440.	0.50213	7	3	0.0060
NNC-F...110.	-1.87602	1	0	1.0000
NNC-Br...110.	-4.00461	10	6	0.0091
NNC-I...110.	-1.87521	4	0	1.0000
NNC-Cl...110.	0.74924	12	1	0.0035
NNC-N...110.	2.74860	18	2	0.0026
NNC-N...202.	7.93582	3	0	1.0000
NNC-N...211.	-2.93642	7	1	0.0015

NNC-N...220.	0.62527	20	1	0.0048
NNC-N...303.	9.56694	2	0	1.0000
NNC-N...312.	-0.37482	9	0	1.0000
NNC-N...321.	2.00020	5	1	0.0003
NNC-N...330.	-3.12734	5	1	0.0003
NNC-O...101.	0.31619	26	1	0.0052
NNC-O...110.	3.56722	97	21	0.0008
NNC-O...211.	-0.31501	4	1	0.0017
NNC-O...220.	-1.30925	41	5	0.0022
NNC-S...110.	5.99876	3	0	1.0000
NNC-S...211.	4.50215	3	0	1.0000
NNC-S...220.	-1.37737	1	0	1.0000
NNC-S...404.	4.56384	3	0	1.0000
NNC-S...413.	1.62301	3	1	0.0038
NNC-s...220.	-1.18633	1	0	1.0000
NOSP00000000	7.37563	21	5	0.0014
NOSP01000000	1.24673	77	17	0.0009
NOSP01100000	0.25158	5	1	0.0003
NOSP10000000	6.55834	10	1	0.0029
NOSP10100000	4.87158	1	0	1.0000
NOSP11000000	6.87659	28	4	0.0015
NOSP11100000	0.93902	5	0	1.0000
S.....	1.12978	13	1	0.0038
[.....	0.31498	13	2	0.0011
\.....	0.37773	21	5	0.0014
c.....	-0.37139	73	15	0.0004
n.....	2.62300	4	0	1.0000
s.....	-0.69062	1	0	1.0000

SM-4

Distribution 1 into the training (+), calibration (#), and validation (*) sets: format of each line is (i) Type of set (i.e. '+', '#', '*'); (ii) CAS number; (iii) SMILES; (iv) *pEC3* value.

One can use these lines as input for the CORAL software.

```
+2657-25-2 O=C(/C=C/c1cccc1)c2ccc(O)cc2 2.699
*1482-74-2 Oc2ccc(C(=O)/C=C/c1cccc1)c(O)c2O 0.959
+111-12-6 O=C(C#CCCCC)OC 0.347
*1776-30-3 O=C(/C=C/c1cccc1)c2ccc(O)cc2O 0.252
+2953-00-7 CC(C)Oc1cc(ccc1O)\C=C\C 0.222
#1210-39-5 O=C\C=C(/c1cccc1)c2cccc2 0.222
+3775-21-1 CC2=NS(=O)(=O)N=C2c1cccc1 -0.146
+13041-12-8 Oc1c(C)cc(cc1OC)\C=C\C -0.204
+100-43-6 C=Cc1ccncc1 -0.204
+29043-97-8 O=C1OC(C)(C)CC1=C -0.255
*3913-71-1 CCCCCC\C=C\C=O -0.398
*111-80-8 O=C(C#CCCCC)OC -0.398
+577-71-9 O=[N+](O-)[c1cc(O)ccc1[N+](O-)]O -0.415
*104-55-2 O=C\C=C/c1cccc1 -0.477
*142-83-6 C\C=C\C=C\C=O -0.544
+122-57-6 CC(=O)/C=C/c1cccc1 -0.568
```

#533-75-5 OC1=CC=CC=CC1=O -0.633
 +21834-92-4 O=C\C(=C\CC(C)C)c1cccc1 -0.643
 #101-39-3 C\C(=C/c1cccc1)C=O -0.653
 +6728-26-3 CCC\C=C\C=O -0.740
 +141-05-9 O=C(\C=C/C(=O)OCC)OCC -0.763
 *116-26-7 CC=1C=CCC(C)(C)C=1C=O -0.875
 +2111-75-3 CC(=C)C1CC=C(C=O)CC1 -0.908
 #NA-1 CC(O)(C=C)CC\C=C(/C)C=O -0.978
 +103-11-7 O=C(C=C)OCC(CCCC)CC -1.000
 +122-40-7 O=CC(\CCCC)=C/c1cccc1 -1.041
 *7492-44-6 O=CC(\CCCC)=C/c1cccc1 -1.041
 +101-86-0 O=CC(\CCCC)=C/c1cccc1 -1.041
 +141-32-2 C=CC(=O)OCCCC -1.041
 +6485-40-1 CC1=CC[C@H](CC1=O)C(C)=C -1.111
 +103-41-3 O=C(OCc1cccc1)/C=C/c2cccc2 -1.265
 +104-54-1 OC\C=C\c1cccc1 -1.322
 +127-51-5 CC1(C)CCC=C(C)C1\C=C(/C)C(C)=O -1.338
 +140-88-5 C=CC(=O)OCC -1.447
 *97-90-5 O=C(OCCOC(=O)C(=C)C)C(=C)C -1.447
 +1565-94-2 C=C(C)C(=O)OCC(O)COc1ccc(cc1)C(C)(C)c2ccc(OCC(O)COC(=O)C(=C)C)cc2 -1.653
 +80-62-6 C=C(C)C(=O)OC -1.954
 +20048-27-5 Nc1ccc(cc1)/N=C3/C=C(N)C(=N\c2ccc(N)cc2)/C=C3N 1.398
 +99-56-9 Nc1ccc(cc1N)[N+](O)=O 1.301
 +65235-31-6 O=[N+](O)c1cc(O)ccc1NCCO 1.155
 +123-31-9 Oc1ccc(O)cc1 0.959
 +106-50-3 Nc1ccc(N)cc1 0.796
 +95-70-5 Nc1ccc(N)c(C)c1 0.699
 *610-81-1 O=[N+](O)c1cc(O)ccc1N 0.699
 *95-55-6 Nc1cccc1O 0.398
 +55302-96-0 Cc1ccc(NCCO)cc1O 0.398
 +5307-14-2 O=[N+](O)c1cc(N)ccc1N 0.398
 +108-45-2 Nc1cccc(N)c1 0.310
 +55658-55-4 CC1=CC[C@H](C/C1=N\O)C(C)=C 0.222
 *10897-60-1 Oc1ccc(CCO)cc1O 0.222
 +35691-65-7 N#CCCC(Br)(CBr)C#N 0.046
 #90-15-3 Oc2cccc1cccc12 -0.114
 #2835-99-6 Cc1cc(O)ccc1N -0.161
 +2871-01-4 [O-][N+](=O)c1cc(N)ccc1NCCO -0.342
 +591-27-5 Nc1cccc(O)c1 -0.505
 +154750-20-6 BrCC1CC(C)(C)OC1=O -0.556
 #93-51-6 Oc1ccc(C)cc1OC -0.763
 +105-13-5 COc1ccc(CO)cc1 -0.771
 +2785-87-7 Oc1ccc(cc1OC)CCC -0.833
 +29705-39-3 O=[N+](O)c1cc(ccc1N)N(CCO)CCO -0.914
 +97-53-0 Oc1ccc(cc1OC)CC=C -1.114
 +186743-25-9 Oc1cc(C)c(cc1OC)CC=C -1.114
 +186743-24-8 Oc1c(C)cc(cc1OC)CC=C -1.230
 *140-67-0 COc1ccc(CC=C)cc1 -1.255
 +186743-26-0 COc1c(C)c(ccc1O)CC=C -1.505
 +111-30-8 O=CCCCC=O 1.000

+57074-21-2 Clc1c(C)cc(O)c(C=O)c1O 0.398
+50-00-0 C=O 0.215
+579-07-7 O=C(C(C)=O)c1ccccc1 -0.114
#107-22-2 O=CC=O -0.146
+122-78-1 O=CCc1ccccc1 -0.477
*93-53-8 CC(C=O)c1ccccc1 -0.799
+112-45-8 C=CCCCCCCCC=O -0.833
+167998-73-4 Cc1c(cc(C)c(C)c1C)C(=O)CC(C)=O -0.919
#167998-76-7 CCc1ccc(cc1C(=O)CC(C)=O)CC -0.982
+465-29-2 CC2(C)C1CCC2(C)C(=O)C1=O -1.000
#110-41-8 CC(CCCCCCCC)C=O -1.000
+55846-68-1 O=C(CC(=O)CCCC)c1ccccc1 -1.041
+502-67-0 C/C(C)=C\CCC(\C)=C\CCC(\C)=C\C=O -1.079
*5392-40-5 C\C(C)=C\CCC(\C)=C\C=O -1.114
+56290-55-2 Cc1ccc(C)cc1C(=O)CC(C)=O -1.114
+5406-12-2 Cc1ccc(CCC=O)cc1 -1.146
+31906-04-4 O=CC1CC=C(CCCC(C)(C)O)CC1 -1.230
#80-54-6 CC(C)(C)c1ccc(CC(C)C=O)cc1 -1.279
#613-29-6 CCCCCN(CCCC)c1ccccc1 -1.292
+326-06-7 O=C(CC(=O)C(F)(F)F)c1ccccc1 -1.301
+2277-19-2 CC\C=C\C\CCCC=O -1.362
+13706-86-0 CC(=O)C(=O)CC(C)C -1.415
+1118-71-4 CC(C)(C)C(=O)CC(=O)C(C)(C)C -1.431
+6668-24-2 O=C(C(C)C(C)=O)c1ccccc1 -1.462
*170928-69-5 Cc1c(cc(C)c(C)c1C)C(=O)CC(=O)OCC -1.519
+107-75-5 O=CCC(C)CCCC(C)(C)O -1.519
+97-96-1 CCC(CC)C=O -1.881
+109-55-7 NCCCN(C)C -0.342
+107-15-3 NCCN -0.342
+111-40-0 NCCNCCN -0.763
#89-25-8 O=C2CC(C)=NN2c1ccccc1 -0.929
+106-24-1 C\C(C)=C\CCC(\C)=C\CO -1.415
#1086-00-6 ClCc4ccc1ccc2cccc3ccc4c1c23 2.301
+26172-55-4 ClC1=CC(=O)N(C)S1 2.046
+70-25-7 N=C(N[N+](=O)[O-])N(C)N=O 1.523
+684-93-5 NC(=O)N(C)N=O 1.301
+57-57-8 O=C1CCO1 0.824
#100-39-0 BrCc1ccccc1 0.699
+2374-65-4 O=S(=O)(CCCCCCCCCCCC)OC 0.409
+55406-53-6 O=C(NCCCC)OCC#CI 0.046
+759-73-9 NC(=O)N(CC)N=O -0.041
+1675-54-3 CC(C)(c2ccc(OCC1CO1)cc2)c4ccc(OCC3CO3)cc4 -0.176
+2634-33-5 O=C2NSc1ccccc12 -0.362
#112-82-3 BrCCCCCCCCCCCCCCCCC -0.362
+118-58-1 Oc2ccccc2C(=O)OCc1ccccc1 -0.462
+64-67-5 O=S(=O)(OCC)OCC -0.519
+3508-00-7 BrCCCCCCCCCCCCCCCCC -0.681
*629-72-1 BrCCCCCCCCCCCCCCCCC -0.708
+137-26-8 CN(C)C(=S)SSC(=S)N(C)C -0.716
#4276-49-7 BrCCCCCCCCCCCCCCCCCCCCC -0.785
*103-63-9 BrCCc1ccccc1 -0.792

#66-27-3 CS(=O)(=O)OC -0.908
 #6938-66-5 BrCCCCCCCCCCCCCCCCCCCCC -0.919
 *51323-71-8 CS(=O)(=O)OCCCCCCCCCCCCC -0.944
 +4860-03-1 CCCCCCCCCCCCCCCCCCl -0.959
 *112-71-0 BrCCCCCCCCCCCCCCCC -0.964
 *111-25-1 BrCCCCCCC -1.000
 *765-09-3 BrCCCCCCCCCCCCCCCC -1.000
 +4292-19-7 CCCCCCCCCCCCCCl -1.114
 *3386-33-2 CCCCCCCCCCCCCCCCCCl -1.204
 +120-51-4 O=C(OCc1ccccc1)c2ccccc2 -1.230
 +73367-80-3 O=C(O)CCCCCCCCCCCCBr -1.255
 *544-77-4 CCCCCCCCCCCCCCCCCCl -1.279
 +693-67-4 BrCCCCCCCCCCCC -1.301
 +2425-54-9 CCCCCCCCCCCCCCCCCCl -1.301
 +74036-97-8 CCCCCCCC(Br)CCCCC -1.322
 +35709-09-2 CS(=O)(=O)OCCCCCCCCC\C=C\CCCCCCCC -1.398
 +2426-08-6 CCCCOC1CO1 -1.491
 +50-32-8 c1cc2c3ccc4cccc5ccc(cc2cc1)c3c45 3.046
 +57-97-6 Cc2c1c4cccc4ccc1c(C)c3ccccc23 2.222
 +15646-46-5 O=C1OC(=N/C1=C\OCC)c2ccccc2 2.523
 +3326-32-7 Oc4cc5Oc1cc(O)ccc1C3(OC(=O)c2cc(\N=C=S)ccc23)c5cc4 0.854
 +525-76-8 CC2=Nc1ccccc1C(=O)O2 0.155
 +176665-02-4 CC1(C)N=C(CCCCC)OC1=O -0.114
 +149-30-4 Sc1nc2ccccc2s1 -0.230
 +176664-99-6 CC1(C)N=C(CCCC)OC1=O -0.255
 *764-85-2 ClC(=O)CCCCCCCC -0.255
 +NA-2 OS(=O)(=O)c1ccccc1OC(=O)C(C)CCCCCCCCCCCCCCCCC -0.301
 +36727-29-4 CC(CC(C)(C)C)CC(Cl)=O -0.431
 #176665-04-6 CC1(C)N=C(CCCCCCCCC)OC1=O -0.447
 +17369-59-4 CC\C=C1/OC(=O)c2ccccc12 -0.568
 +119-84-6 O=C1CCc2ccccc2O1 -0.748
 +552-30-7 O=C(O)c1ccc2C(=O)OC(=O)c2c1 -0.964
 +176665-06-8 CC1(C)N=C(CCCCCCCCCC)OC1=O -1.204
 +176665-09-1 CC1(C)N=C(CCCCCCCCCCCCCC)OC1=O -1.255
 *176665-11-5 CC1(C)N=C(CCCCCCCCCCCCCC)OC1=O -1.279
 #93-99-2 O=C(Oc1ccccc1)c2ccccc2 -1.301
 +39236-46-9 O=C2NC(NC(=O)NCNC(=O)NC1NC(=O)N(CO)C1=O)C(=O)N2CO -1.380
 +61-33-6 OC(=O)C2N3C(=O)C(NC(=O)Cc1ccccc1)C3SC2(C)C -1.477
 +4638-48-6 Oc2ccc(Cl)cc2C(=O)Nc1ccccc1 -0.699
 +99-83-2 CC=1C=CC(CC=1)C(C)C -0.732
 +NA-3 CC1=CC[C@H](CC1=C)C(C)=C -0.863
 +2136-71-2 OCCOCCCCCCCCCCCCCCCC -0.944
 +99-86-5 CC(C)C1=CC=C(C)CC1 -0.949
 #32388-55-9 CC(=O)C=1C[C@]23C[C@@H](C=1C)C(C)(C)[C@@H]3CC[C@H]2C -1.143
 #514-10-3 CC(C)C=3CC[C@H]1C(=CC[C@H]2[C@@](C)(CCC[C@]12C)C(=O)O)C=3 -1.176
 +78-70-6 CC(O)(C=C)CC\C=C(\C)C -1.477
 +5989-27-5 CC1=CCC(CC1)C(C)=C -1.839
 +1897-45-6 Clc1c(C#N)c(Cl)c(C#N)c(Cl)c1Cl 2.398
 +97-00-7 O=[N+](O-)[O-]c1cc(ccc1Cl)[N+](O-)=O 1.301
 +108-77-0 Clc1nc(Cl)nc(Cl)n1 1.046
 +87-86-5 Clc1c(O)c(Cl)c(Cl)c(Cl)c1Cl -1.301

+23593-75-1 Clc1cccc1C(c2cccc2)(c3cccc3)n4ccnc4 -0.681
+106-22-9 C\C(C)=C\CCC(C)CCO -1.638
+106-51-4 O=C1C=CC(=O)C=C1 2.004
+62-53-3 Nc1cccc1 -1.949
+555-10-2 C=C1C=CC(CC1)C(C)C -0.748
+NA-4 CC1(C)N=C(CCCCCCCCCCCCCCCCCC)OC1=O -1.415
+112-67-4 ClC(=O)CCCCCCCCCCCCC -0.944
+57077-36-8 ClC(=O)CCCCC(C)C -0.431
+1154-59-2 Clc2ccc(NC(=O)c1cc(Cl)cc(Cl)c1O)cc2Cl 1.398
+4282-42-2 CCCCCCCCCI -1.380
+143-15-7 BrCCCCCCCCCCCCC -1.255
+19218-94-1 CCCCCCCCCCCCCCI -1.146
+112-89-0 BrCCCCCCCCCCCCCCCCC -1.176
#3344-77-2 BrCCCCCCCCCCCCCO -0.839
#10520-81-7 O=C(O)C(Br)CCCCCCCCCCCC -0.531
+2682-20-4 O=C1C=CSN1C -0.279
+100-11-8 O=[N+](O-)[O-]c1ccc(CBr)cc1 1.301
+77-78-1 O=S(=O)(OC)OC 0.721
+620159-84-4 CCC(C)(C)C1CCC(CC1)CC=O -1.568
+1205-17-0 O=CC(C)Cc1ccc2OCOc2c1 -1.215
+35578-47-3 O=C(c1ccc(Br)cc1)C(=O)c2ccc(Br)cc2 -1.312
+103-95-7 CC(C)c1ccc(CC(C)C=O)cc1 -1.342
+431-03-8 CC(=O)C(C)=O -1.041
*600-22-6 CC(=O)C(=O)OC -0.380
#97-54-1 Oc1ccc(cc1OC)\C=C\C -0.079
*818-61-1 C=CC(=O)OCCO -0.146
+104-46-1 COc1ccc(/C=C/C)cc1 -0.362
+2050-14-8 Oc2cccc2/N=N/c1cccc1O -1.446
+526-37-4 O=Cc1c(cc(C)cc1O)O 0.222
+6358-09-4 Nc1cc(cc(Cl)c1O)[N+](O-)=O -0.836
#2835-95-2 Nc1cc(O)c(C)cc1 -0.531
+1166-52-5 Oc1cc(cc(O)c1O)C(=O)OCCCCCCCCCCCC 0.523
#186743-29-3 COc1c(C)c(ccc1O)\C=C\C -0.556
*96-33-3 C=CC(=O)OC -1.301
+5910-85-0 CC\C=C/C=C\C=O -0.602
+104-27-8 COc1ccc(/C=C/C(=O)CC)cc1 -0.968