CORAL: Binary classifications (active/inactive) for drug-induced liver injury

Alla P. Toropova*, Andrey A. Toropov

Department of Environmental Health Science, Laboratory of Environmental Chemistry and Toxicology, IRCCS-Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20156 Milano, Italy

Published version of this paper could be find here https://doi.org/10.1016/j.toxlet.2017.01.011

Toxicology Letters, Volume 268, 15 February 2017, Pages 51-57

Abstract

Introduction: The data on human hepatotoxicity (drug-induced liver injury) is extremely important information from point of view of drug discovery. Experimental clinical data on this endpoint is scarce. Experimental way to extend databases on this endpoint is extremely difficult. Quantitative structure - activity relationships (QSAR) is attractive alternative of the experimental approach.

Methods: Predictive models for human hepatotoxicity (drug-induced liver injury) have been built up by the Monte Carlo method with using of the CORAL software (http://www.insilico.eu/coral). These models are the binary classifications into active class and inactive class. These models are calculated with so-called “semi correlations” described in this work. The Mattews correlation coefficient of these models for external validation sets ranged from 0.52 to 0.62.

Results discussion: The approach has been checked up with a group of random splits into the training and validation sets. These stochastic experiments have shown the stability of results: predictability of the models for various splits. Thus, the attempt to build up the classification QSAR model by means of the Monte Carlo technique, based on representation of the molecular structure via simplified molecular input line entry systems (SMILES) and hydrogen suppressed graph (HSG) using the CORAL software (http://www.insilico.eu/coral) has shown ability of this approach to provide quite good prediction of the examined endpoint (drug-induced liver injury).

Keywords: QSAR; Drug-induced liver injury; Monte Carlo method; Predictive toxicology; Risk assessment

*) 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.

E-mail: alla.toropova@marionegri.it

1. Introduction

The computational prediction of biochemical endpoints without animal testing is encouraged by various social organizations, for instance, REACH (Registration, Evaluation, Authorisation and Restriction of Chemical Substances) (Pery et al., 2013) and OECD (Organisation for Economic Co-operation and Development) (Gobbi et al., 2016). However it should be noted, the animal testing remains the basis of development of the above computational methods. The collecting of data on therapeutic effects together with possible toxic effects of various substances upon human takes very long time (Mann, 2015). Consequently, the computational predictions of therapeutic and toxic effects upon human is a problem even more important than attempts to avoid (or at least to reduce) the animal testing.

In fact, each therapeutic agent has both positive and negative (or even dangerous) effects (Toropova and Toropov, 2014; Gobbi et al., 2016). Drug-induced liver injury is one of the most common drug-induced impact to human organism leading to life-threatening conditions such as acute liver failure (Persson et al., 2013; Chen et al., 2015; Xu et al., 2015; Zhu and Kruhlak, 2014; Kostadinova et al., 2013), in addition this is one of the leading causes of the termination of drug development projects (Chen et al., 2013). Preliminary estimation of a drug's potential to cause drug-induced liver injury in humans is a complex problem caused by absence of sensitive and reliable biomarkers able to indicate the dengerous therapeutic agent (Chen et al., 2015,2014). Recent reviews (Chen et al., 2015,2014) contains a large group of quantitative structure – activity relationships (QSARs) related to the drug-induced liver injury.

Thus, the QSAR analysis of drug-induced liver injury should be estimated as important task of the medicinal chemistry (Xu et al., 2015).

The CORAL software (http://www.insilico.eu/coral) is a tool to build up QSAR models for endpoints related to medicinal chemistry (Veselinović et al., 2013; Worachartcheewan et al., 2014; Li et al., 2014; Ghaedi, 2015; Nesmerak et al., 2015; Islam and Pillay, 2016). In addition, the classification models for anti-sarcoma activity (Toropov et al., 2012a) and for liver-related adverse effects of drugs (Toropov et al., 2012b) were suggested. Thus, further studies of the software can be useful from practical and theoretical points of view.

The aim of this work is to estimate the CORAL software as a tool to build up predictive models for the drug-induced liver injury by means of the Monte Carlo technique.

2. Method

2.1. Data

The collection of 2029 therapeutic agents with the drug-induced liver injury expressed in form active and inactive according to experimental data is available in the literature (Zhu and Kruhlak, 2014). A drug is labeled as hepatotoxic if (1) it is present in the US drug-induced liver injury network (DILIN) (Fontana et al., 2009), or (2) it is known to cause acute liver failure, or (3) it has been withdrawn or suspended in either the US or European markets. A drug is labeled as non hepatotoxic if it had been on the market for more than 5 years and if no publications (from 1970 to 2012, including databases) contain facts about hepotoxicy of this drug.

The above-mentioned data were split into the training (≈ 70%), calibration (≈ 15%), and validation (≈ 15%) sets, three times. The principles of the distribution of compounds into the training, calibration, and validation sets are the following: (i) these distributions should be random; (ii) the training set should contain majority of available compounds; (iii) these distributions should be significantly different (Table 1).

[Table 1 around here]

2.2. Optimal descriptor

The optimal descriptor of correlation weights (DCW) used in this work is calculated as the following:

$$\begin{aligned} DCW(T^*,N^*) = {} & \Sigma CW(s_k) + \Sigma CW(ss_k) + \Sigma CW(sss_k) + \\ & CW(HARD) + \\ & \Sigma CW(EC0_k) + \Sigma CW(EC1_k) + \Sigma CW(EC2_k) + \Sigma CW(NNC_k) + \\ & CW(C3) + CW(C5) + CW(C6) \end{aligned} \tag{1}$$

The optimal descriptor calculated with Eq. 1 is so-called hybrid descriptor (Toropova et al., 2012; Achary, 2014a,b; Fatemi and Malekzadeh, 2015). The hybrid descriptor is calculated from two representations of the molecular structure: (i) hydrogen suppressed graph (HSG); and (ii) simplified molecular input-line entry systems (SMILES).

Table 2 contains the scheme of registration of molecular features extracted from HSG (*$EC0_k$*, *$EC1_k$*, *$EC2_k$*, *NNC_k*, *C3*, *C5*, and *C6*). Table 3 contains the scheme of registration of molecular features extracted from SMILES (*S_k*, *SS_k*, *SSS_k*, and *HARD*). Each molecular feature mentioned above is represented by sequence of twelve symbols (Table 2 and Table 3).

[Table 2 around here]

[Table 3 around here]

The *$CW(S_k)$, $CW(SS_k)$, $CW(SSS_k)$, CW(HARD), $CW(EC0_k)$, $CW(EC1_k)$, $CW(EC2_k)$, $CW(NNC_k)$, CW(C3), CW(C5), and CW(C6)* are correlation weights for the above mentioned molecular features extracted from SMILES and HSG. The numerical data on the correlation weights are calculated by the

Monte Carlo method optimization (Toropova et al., 2011, 2012). The correlation coefficient between the optimal descriptor and an endpoint is the target function of the optimization.

The T is threshold, i.e. coefficient to separate molecular features into two categories: rare and not rare. For instance, if T=2, then features, which have prevalence less than 2 in the training set, are recognized as rare. The features, which are recognized as rare have correlation weights equal to zero, i.e. these features are not involved in building up a model. The N is the number of epochs of the Monte Carlo optimization. The T* and N* are such values of these parameters which give the best statistics for the calibration set (Toropova and Toropov, 2014).

The balance of correlation has been used to calculate the correlation weights (Toropova et al., 2011). The general scheme of the balance of correlations involve four steps.

Step 1. The available data distributed into four sets: training, invisible training, calibration, and validation sets. The training set is builder of the model: structures of this set provide correlation weights for molecular features in the Monte Carlo optimization. The invisible training set is utilized to check whether correlation between descriptor and endpoint is true. In other words, compounds of the invisible training set should confirm that even for them the correlation between descriptor and endpoint takes place. The calibration set is utilized to detect the number of iteration N_x, when the overtraining appears (improving of correlation for training set is accompanied by decrease of correlation coefficient for the calibration set). Therefore, the above-mentioned N* should be calculated as $N^*=N_x-1$.

Step 2. Building up model with T=T* and N=N*.

Step 3. Estimation of the predictive potential of the obtained model (using the T* and N*) with external validation set.

In this work, instead of the traditional correlations, the “semi correlations” have been built up for three different splits (Figure 1).

[Figure 1 around here]

In the case of the binary classification model (active=1, inactive=0) the following statistical characteristics are utilized (Toropov et al., 2012) Mattews correlation coefficient (MCC), sensitivity, specificity, and accuracy, which are calculated as the following:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{2}$$

$$Sensitivity = \frac{TP}{TP+FN} \tag{3}$$

$$Specificity = \frac{TN}{FP + TN} \tag{4}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{5}$$

The criteria (Eqs. 2-5) are measure for quality classification model. The MCC is the analogy of the traditional correlation coefficient for the case of binary classification. The sensitivity is measure of ability of the model correctly detect positive observations (i.e. active compounds). The specificity is measure of ability of the model correctly detect negative observation (i.e. inactive compounds). The accuracy is measure of the general ability of the model (i.e. quality of prediction for both positive and negative observations).

In addition, criteria suggested recently in the literature (Veselinović et al., 2016) to define the quality of splits and domain of applicability for the traditional correlations were used in this work for analysis of the "semi correlations".

The distribution into the "visible" training and calibration sets and "invisible" validation set has apparent influence upon the predictability of a model. A possible measure of the quality of the split is the following:

$$SA_{defect} = \sum_{active} \Sigma \left| P(SA) - P'(SA) \right| \tag{6}$$

where the probability of an structural attribute (SA) in the training set P(SA) and the probability of SA in the calibration set P'(SA) are calculated by

$$P(SA) = \frac{N_{set}(SA)}{N_{set}} \tag{7}$$

where $N_{set}(SA)$ is the number of SMILES which are containing SA and the N_{set} is the total number of SMILES in the set. The defect is calculated with only active (not blocked) SA (Veselinović et al., 2016). If the defect = 0, the split should be estimated as "ideal" one. However in fact, this situation is impossible. However, the value of the defect calculated with Eq. 6 gives possibility to compare various splits.

The sum of the SA*defect* of all active SMILES attributes can be a measure of quality (defect) of each SMILES:

$$SMILES_{defect} = \sum_{SA_{defect} \in SMILES} SA_{Defect} \qquad (8)$$

The sum of all SPLIT*defect* can be a measure of quality (defect) of the split into the visible training (calibration) sets and invisible validation set:

$$Split_{defect} = W\% \times \sum SMILES_{Defect} \qquad (9),$$

where W% is the percentage of molecular features which are present in the training set, in the invisible training set, and in the calibration set, simultaneously.

The probabilistic domain of applicability can be defined via inequality

$$SMILES_{defect} < 2 \times \overline{SMILES_{defect}} \qquad (10)$$

In other words, a SMILES characterized by the *SMILESdefect* which is lower than the doubled average value of the characteristics over compounds of the training set, the SMILES falls into the domain of applicability, otherwise the SMILES is out of the domain of applicability.

In addition, one can compare quality (defect) of different splits into the training, calibration, and validation sets: preferable split should be characterized by lower defect calculated with Eq. 9. Thus, internal selforganization of split based on the criterion calculated with Eq. 9 becomes available.

3. Results and Discussion

Table 4 contains the statistical characteristics of the binary classifications. The activity defined according to the formula

$$\text{Activity} = \begin{cases} 1, & \text{if } \; y > 0.5 \\ 0, & \text{if } \; y \le 0.5 \end{cases} \qquad (11)$$

where $y = C_0 + C_1 \times DCW(T^*, N^*)$

The split defect (Eq. 9) and the MCC (Eq. 2) for the validation set are correlated (Figure 2). Thus, the split defect calculated with Eq. 9 can be a criterion to compare expected predictive potential of models before building up these models. Three different splits examined in this work are represented in *Supplementary materials* section. It is to be noted, that models built up in this work obey OECD principles (Toropova and Toropov, 2014). These data can be used to reproduce the suggested model with the CORAL software available on the Internet (http://www.insilico.eu/coral).

In the recent review (Chen et al., 2014) a group of QSAR models for the drug-induced liver injury have been represented and compared. The ranges of sensitivity, specificity, and accuracy over the group of QSAR are 0.40-0.94; 0.65-0.92; and 0.46-0.82, respectively. Thus, the statistical characteristics of models built up in this work are satisfactory (Table 4).

[Table 4 around here]

Having results of a group of runs of the Monte Carlo optimization, one obtain three categories of the SMILES attributes: (i) attributes which have solely positive correlation weights. These are promoters of activity of compounds; (ii) attributes which have solely negative correlation weights. These are promoters of inactivity of compounds; and (iii) attributes which have in several runs of the Monte Carlo optimization both positive and negative correlation weights, the role of these attributes is not clear. Thus, the suggested approach gives possibility for mechanistic interpretation of a model (Toropov et al., 2012; Gobbi et al., 2016; Veselinović et al., 2016). The analysis of three models built up with different distributions into the training, invisible training, calibration, and validation set has shown: there are structural indicators of high probability of liver injury caused by impact of a drug-like substance. The structural alerts (SA) should be selected in accordance with two conditions: (i) the correlation weight of the SA must be positive for all models; and (ii) the prevalence of the SA should be significant. Table 5 contains a collection of structural alerts which obey these conditions. In addition, it was noted, the presence of double bonds is promoter of decrease of probability of liver injury (Table 5).

[Table 5 around here]

The CORAL software gives possibility of application of two representations of the molecular structure via SMILES (Gobbi et al., 2016) and via molecular graphs (Toropov et al., 2012). The integrated representation with involving both molecular features extracted from SMILES together with features extracted from graph also is available (Toropov et al., 2012). In the previous work (Toropov et al., 2012) the integrated list of molecular features extracted from SMILES and graph has been utilized to predict liver-related adverse effects of drugs. Attempts to involve similar integrated molecular features to build up a predictive model for hepatotoxicity have shown that hybrid descriptors gives better prediction in terms of sensitivity, specificity, accuracy, and MCC. It is possible, an approach based on solely graph can be equivalent or even better than approach suggested in this work, but such combinations of features extracted from graph or combinations of features extracted from SMILES were not found.

Important advantages of the suggested approach is the possibility to build up models solely from data on molecular architecture represented by SMILES and experimental data for compounds, without additional physicochemical descriptors or descriptors of quantum mechanics.

Previous works (Toropov at al., 2012a,b) have shown, that the “semi correlations” are able be a tool to build up the predictive classification model. There are two improvements used in this work and which were not available for the above-mentioned works (Toropov et al., 2012a,b). The first, in this work, new descriptors related to rings as well as the integrated descriptor HARD are involved in building up models. The second, balance of correlations done with taking into account new statistical characteristics of distributions into the training and validation sets (Eq. 9).

4. Conclusions

The suggested binary classifications are characterized by quite good values of the MCC, sensitivity, specificity, and accuracy. The inequality 10 gives possibility to define the domain of applicability of the models. The models give the mechanistic interpretation for the approach in terms of structural alerts, i.e. structural attributes, which should be positive for all models and which have significant prevalence in the training and calibration sets. Thus, the suggested models are built up according to OECD principles (Gobbi et al., 2016).

Acknowledgements

The authors are grateful for the contribution of the EC project PeptiCAPS (Project reference: 686141).

References

Achary, P.G.R., 2014a. QSPR modelling of dielectric constants of π-conjugated organic compounds by means of the CORAL software. SAR QSAR Environ. Res. 25 (6), 507-526.

Achary, P.G.R., 2014b. Simplified molecular input line entry system-based optimal descriptors: QSAR modelling for voltage-gated potassium channel subunit Kv7.2. SAR QSAR Environ. Res. 25 (1), 73-90.

Chen, M. , Hong, H., Fang, H., Kelly, R., Zhou, G., Borlak, J., Tong, W., 2013. Quantitative Structure-Activity Relationship models for predicting drug-induced liver injury based on FDA-approved drug labeling annotation and using a large collection of drugs. Toxicol. Sci. 136, 242–249.

Chen, M., Bisgin, H., Tong, L., Hong, H., Fang, H., Borlak, J., Tong, W., 2014. Toward predictive models for drug-induced liver injury in humans: are we there yet? Biomark. Med. 8, 201–213.

Chen, M., Suzuki, A., Borlak, J., Andrade, R.J., Lucena, M.I., 2015. Drug-induced liver injury: Interactions between drug properties and host factors. J. Hepatol. 63, 503–514.

Fatemi, M.H., Malekzadeh, H., 2015. CORAL: Predictions of retention indices of volatiles in cooking rice using representation of the molecular structure obtained by combination of SMILES and graph approaches. J. Iran. Chem. Soc. 12 (3), 405-412.

Fontana, R.J., Watkins, P.B., Bonkovsky, H.L., Chalasani, N., Davern, T., Serrano, J., Rochon, J., 2009. Drug-Induced Liver Injury Network (DILIN) prospective study: rationale, design and conduct. Drug Saf. 32 (1), 55–68.

Islam, M.A., Pillay, T.S., 2016. Simplified molecular input line entry system-based descriptors in QSAR modeling for HIV-protease inhibitors. Chemometr. Intell. Lab. Syst. 153, 67-74.

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.

Kostadinova R., Boess F., Applegate D., Suter L., Weiser T., Singer T., Naughton B., Roth A., 2013. A long-term three dimensional liver co-culture system for improved prediction of clinically relevant drug-induced hepatotoxicity. Toxicol. Appl. Pharmacol. 268, 1–16.

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.

Mann, D.A., 2015. Human induced pluripotent stem cell-derived hepatocytes for toxicology testing. Expert Opin. Drug Metab. Toxicol. 11, 1-5.

Nesměrák, K., Toropov, A.A., Toropova, A.P., Yildiz, I., Yalcin, I., Brozikova, M., Klimešová, V., Waisser, K., 2015. Prediction of retention characteristics of heterocyclic compounds. Anal. Bioanal. Chem. 407 (30), 9185-9189.

Persson, M., Løye, A.F., Mow, T., Hornberg, J.J., 2013. A high content screening assay to predict human drug-induced liver injury during drug discovery. J. Pharmacol. Toxicol. Methods 68, 302–313.

Péry, A.R.R., Brochot, C., Zeman, F.A., Mombelli, E., Desmots, S., Pavan, M., Fioravanzo, E., Zaldívar, J.-M., 2013. Prediction of dose-hepatotoxic response in humans based on toxicokinetic / toxicodynamic modeling with or without in vivo data: A case study with acetaminophen. Toxicol. Lett. 220, 26-34.

Toropov, A.A., Toropova, A.P., Benfenati, E., Gini, G., Leszczynska, D., Leszczynski, J., 2012a. CORAL: Classification model for predictions of anti-sarcoma activity. Curr. Top. Med. Chem. 12 (24), 2741-2744.

Toropov, A.A., Toropova, A.P., Rasulev, B.F., Benfenati, E., Gini, G., Leszczynska, D., Leszczynski J., 2012b. CORAL: binary classifications (active/inactive) for liver-related adverse effects of drugs. Curr. Drug Saf. 7, 257-261.

Toropova, A.P., Toropov, A.A., Benfenati, E., Gini, G., Leszczynska, D., Leszczynski, J., 2011. CORAL: quantitative structure–activity relationship models for estimatingtoxicity of organic compounds in rats. J. Comput. Chem. 32 (12), 2727–2733.

Toropova, A.P., Toropov, A.A., Martyanov, S.E., Benfenati, E., Gini, G., Leszczynska, D., Leszczynski, J., 2012. CORAL: QSAR modeling of toxicity of organic chemicals towards Daphnia magna. Chemometr. Intell. Lab. Syst., 110 (1), 177-181.

Toropova, A.P., Toropov, A.A., 2014. CORAL software: prediction of carcinogenicity of drugs by means of the Monte Carlo method. Eur. J. Pharm. Sci. 52, 21-25.

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. Pharmacol. 42, 135–145

Veselinović, A.M., Milosavljević, J.B., Toropov, A.A., Nikolić, G.M., 2013. SMILES-Based QSAR models for the calcium channel-antagonistic effect of 1,4-dihydropyridines. Arch. Pharm., 346 (2), 134-139.

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.

Zhu, X., Kruhlak, N.L., 2014. Construction and analysis of a human hepatotoxicity data base suitable for QSAR modeling using post-market safety data. Toxicology 321, 62–72.

Xu, Y., Dai, Z., Chen, F., Gao, Sh., Pei, J., Lai, L., 2015. Deep learning for drug-induced liver injury. J. Chem. Inf. Model. 55, 2085−2093.

Worachartcheewan, A., Nantasenamat, C., Isarankura-Na-Ayudhya, C., Prachayasittikul, V., 2014. QSAR study of H1N1 neuraminidase inhibitors from influenza a virus. Lett. Drug Des. Discov., 11 (4), 420-427.

Table 1

The measure of non-identity of splits into the training, invisible training, calibration, and validation sets, which are examined in this work

split	**Set**	**Split 1**	**Split 2**	**Split 3**
1	Training	100*	97.6	37.1
	Invisible training	100	98.2	39.1
	Calibration	100	4.8	24.0
	Validation	100	6.1	20.9
2	Training		100	37.7
	Invisible training		100	39.0
	Calibration		100	20.5
	Validation		100	15.2
3	Training			100
	Invisible training			100
	Calibration			100
	Validation			100

$$Identity\,(\%) = \frac{N_{i,j}}{0.5 * (N_i + N_j)} \times 100$$

where

$N_{i,j}$ is the number of substances which are distributed into the same set for both i-th split and j-th split (set =sub-training, calibration, test, validation) ;

N_i is the number of substances which are distributed into the set for i-th split;

N_j is the number of substances which are distributed into the set for j-th split.

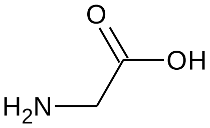
Table 2
Molecular features extracted from HSG for building up models

<table>
<tr><th>ID</th><th>Comment</th></tr>
<tr><td><i>EC0_k</i></td><td>Vertex degree for k-th vertex (the number of neighbors which are not hydrogen atoms). For instance, carbon vertex with vertex degree equal to three is represented by the following twelve symbols
<table>
<tr><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td><td>9</td><td>10</td><td>11</td><td>12</td></tr>
<tr><td><b>E</b></td><td><b>C</b></td><td><b>0</b></td><td><b>-</b></td><td><b>C</b></td><td><b>.</b></td><td><b>.</b></td><td><b>.</b></td><td><b>3</b></td><td><b>.</b></td><td><b>.</b></td><td><b>.</b></td></tr>
</table></td></tr>
<tr><td><i>EC1_k</i></td><td>Extended connectivity of the first order (Toropova et al., 2011; 2016)
$$EC1_k = \sum_{Edge(k,j)} EC0_j$$
For instance, nitrogen vertex with EC1=10 is represented by the following twelve symbols:
<table>
<tr><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td><td>9</td><td>10</td><td>11</td><td>12</td></tr>
<tr><td><b>E</b></td><td><b>C</b></td><td><b>1</b></td><td><b>-</b></td><td><b>N</b></td><td><b>.</b></td><td><b>.</b></td><td><b>.</b></td><td><b>1</b></td><td><b>0</b></td><td><b>.</b></td><td><b>.</b></td></tr>
</table></td></tr>
<tr><td><i>EC2_k</i></td><td>Extended connectivity of the second order (Toropova et al., 2011; 2016)
$$EC2_k = \sum_{Edge(k,j)} EC1_j$$
For instance, oxygen vertex with EC2=17 is represented by the following twelve symbols:</td></tr>
<tr><td></td><td><table>
<tr><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td><td>9</td><td>10</td><td>11</td><td>12</td></tr>
<tr><td><b>E</b></td><td><b>C</b></td><td><b>2</b></td><td><b>-</b></td><td><b>O</b></td><td><b>.</b></td><td><b>.</b></td><td><b>.</b></td><td><b>1</b></td><td><b>7</b></td><td><b>.</b></td><td><b>.</b></td></tr>
</table></td></tr>
<tr><td><i>NNC_k</i></td><td>
$$NNC_k = 100 \times N_{Total} + 10 \times N_C + N_{A \neq C}$$
N_{Total} is the total number of neighbors for k-th vertex
N_C is the number of neighbors which are carbon vertexes
$N_{A \neq C}$ is the number of neighbors which are not carbon vertexes
For example, NNCk=211 for nitrogen vertex is represented by the following twelve symbols
<table>
<tr><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td><td>9</td><td>10</td><td>11</td><td>12</td></tr>
<tr><td><b>N</b></td><td><b>N</b></td><td><b>C</b></td><td><b>-</b></td><td><b>N</b></td><td><b>.</b></td><td><b>.</b></td><td><b>.</b></td><td><b>2</b></td><td><b>1</b></td><td><b>1</b></td><td><b>.</b></td></tr>
</table></td></tr>
<tr><td><i>C3</i></td><td>Descriptor for three-members rings, reflects their number (0, 1, 2, …); presence of heteroatoms (H). For instance, below the version of C3 where recorded the following situation: (i) there are two three-members rings; and (ii) at least one of them contains heteroatom
<table>
<tr><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td><td>9</td><td>10</td><td>11</td><td>12</td></tr>
<tr><td><b>C</b></td><td><b>3</b></td><td><b>.</b></td><td><b>.</b></td><td><b>.</b></td><td><b>.</b></td><td><b>H</b></td><td><b>.</b></td><td><b>2</b></td><td><b>.</b></td><td><b>.</b></td><td><b>.</b></td></tr>
</table></td></tr>
<tr><td><i>C5</i></td><td>The analogy of C3 for five-members rings. For instance, below the version of C5 where recorded the following situation: (i) there are three five-members rings; and (ii) these rings have not heteroatoms
<table>
<tr><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td><td>9</td><td>10</td><td>11</td><td>12</td></tr>
<tr><td><b>C</b></td><td><b>5</b></td><td><b>.</b></td><td><b>.</b></td><td><b>.</b></td><td><b>.</b></td><td><b>.</b></td><td><b>.</b></td><td><b>3</b></td><td><b>.</b></td><td><b>.</b></td><td><b>.</b></td></tr>
</table></td></tr>
<tr><td><i>C6</i></td><td>The analogy of C3 (or C5) for six-members rings. For instance, below the version of C6 where recorded the following situation: (i) there are three six-members rings; and (ii) at least one of them contains heteroatom
<table>
<tr><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td><td>9</td><td>10</td><td>11</td><td>12</td></tr>
<tr><td><b>C</b></td><td><b>6</b></td><td><b>.</b></td><td><b>.</b></td><td><b>.</b></td><td><b>.</b></td><td><b>H</b></td><td><b>.</b></td><td><b>3</b></td><td><b>.</b></td><td><b>.</b></td><td><b>.</b></td></tr>
</table></td></tr>
</table>

Table 3

Molecular features extracted from SMILES for building up models

<table>
<tr><th>ID</th><th>Comment</th></tr>
<tr><td>S_k

SS_k
SSS_k</td><td>SMILES-atom, i.e. one character or two characters which cannot examined separately, e.g. ‘Cl’, ‘Br’, etc.
Two SMILES-atoms
Three SMILES-atoms
For example, if SMILES is sequence of symbols ‘ABCDE’, than
S_k = {‘A’,’B’,’C’,’D’, and ’E’}
SS_k = {‘AB’,’BC’,’CD’, and ’DE’}
SSS_k = {‘ABC’,’BCD’, and ’CDE’}</td></tr>
<tr><td></td><td>Example of registration for S_k (‘A’)
<table>
<tr><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td><td>9</td><td>10</td><td>11</td><td>12</td></tr>
<tr><td><b>A</b></td><td>.</td><td>.</td><td>.</td><td>.</td><td>.</td><td>.</td><td>.</td><td>.</td><td>.</td><td>.</td><td>.</td></tr>
</table></td></tr>
<tr><td></td><td>Example of registration for SS_k (‘BC’)
<table>
<tr><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td><td>9</td><td>10</td><td>11</td><td>12</td></tr>
<tr><td><b>B</b></td><td>.</td><td>.</td><td>.</td><td><b>C</b></td><td>.</td><td>.</td><td>.</td><td>.</td><td>.</td><td>.</td><td>.</td></tr>
</table></td></tr>
<tr><td></td><td>Example of registration for SSS_k (‘CDE’)
<table>
<tr><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td><td>9</td><td>10</td><td>11</td><td>12</td></tr>
<tr><td><b>C</b></td><td>.</td><td>.</td><td>.</td><td><b>D</b></td><td>.</td><td>.</td><td>.</td><td><b>E</b></td><td>.</td><td>.</td><td>.</td></tr>
</table></td></tr>
<tr><td>HARD</td><td>
<table>
<tr><td>SMILES</td><td>NCC(O)=O</td></tr>
<tr><td>Structure</td><td></td></tr>
</table>
For this structure, HARD is registrated by the following twelve symbols*
<table>
<tr><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td><td>9</td><td>10</td><td>11</td><td>12</td></tr>
<tr><td></td><td>=</td><td>#</td><td>@</td><td>N</td><td>O</td><td>S</td><td>P</td><td>F</td><td>Cl</td><td>Br</td><td>I</td></tr>
<tr><td><b>$</b></td><td><b>1</b></td><td><b>0</b></td><td><b>0</b></td><td><b>1</b></td><td><b>1</b></td><td><b>0</b></td><td><b>0</b></td><td><b>0</b></td><td><b>0</b></td><td><b>0</b></td><td><b>0</b></td></tr>
</table></td></tr>
</table>

*) The ‘=’, ‘#’, and ‘@’ are indicators of chemical bonds (double, triple, and stereo chemical); N, O, S, P, F, Cl, Br, and I are names of chemical elements. If structure contains an object, then the corresponding symbol is 1, if the object is absent the corresponding symbol is 0.

Table 4

The statistical characteristics of models for potential ability of compounds to lead to liver injury

Split	Set	TP*	TN	FP	FN	N	Sensitivity	Specificity	Accuracy	MCC
y = -0.0067297 (± 0.0006349) + 0.0227289 (± 0.0000260) * DCW(1,29)										
1	Training	152	391	44	103	690	0.5961	0.8989	0.7870	0.530
	Invisible training	183	361	60	98	702	0.6512	0.8575	0.7749	0.524
	Calibration	41	227	30	21	319	0.6613	0.8833	0.8401	0.518
	Validation	46	217	37	18	318	0.7188	0.8543	0.8270	0.523
y = -0.2386871 (± 0.0008047) + 0.0177597 (± 0.0000205) * DCW(1,21)										
2	Training	159	397	42	110	708	0.5911	0.9043	0.7853	0.533
	Invisible training	185	355	66	97	703	0.6560	0.8432	0.7681	0.511
	Calibration	47	199	47	16	309	0.7460	0.8089	0.7961	0.486
	Validation	42	226	35	6	309	0.8750	0.8659	0.8673	0.621
y = -0.1390926 (± 0.0006896) + 0.0163083 (± 0.0000188) * DCW(1,27)										
3	Training	205	399	64	98	766	0.6766	0.8618	0.7885	0.551
	Invisible training	202	393	61	94	750	0.6824	0.8656	0.7933	0.561
	Calibration	34	188	34	0	256	1.0000	0.8468	0.8672	0.651
	Validation	29	194	34	0	257	1.0000	0.8509	0.8677	0.625

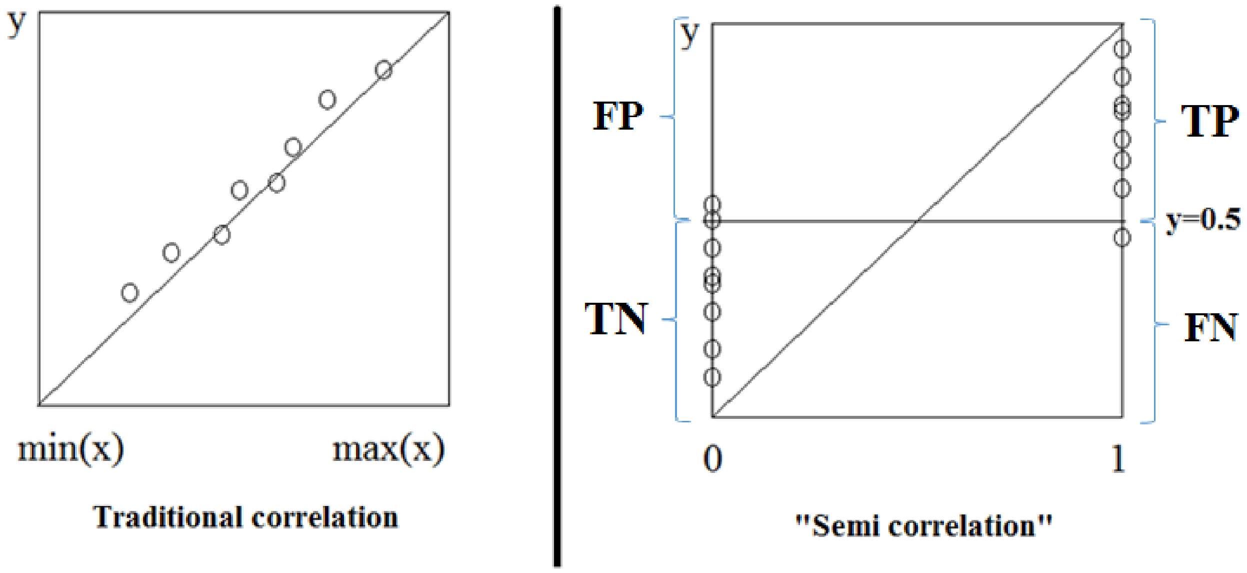
*) TP = true positive; TN = true negative; FP = false positive; and FN = false negative.

Table 5

Structural attributes with significant prevalence, which are indicator of dangerous potential of a compounds in aspect of the drug-induced liver injury

Split	Structural attribute, *SA*	*CW(SA)*	N1*	N2	N3	SA_{Defect}
	Promoter of toxicity increase					
1	`C6......1...`	5.69241	129	108	59	0.0000
2	`C6......1...`	8.30911	131	108	71	0.0002
3	`C6......1...`	6.25372	125	136	58	0.0003
1	`$10011000000`	1.74667	135	137	102	0.0005
2	`$10011000000`	3.80956	142	139	81	0.0003
3	`$10011000000`	3.62303	161	152	73	0.0003
1	`C6....H.3...`	2.49876	90	102	37	0.0001
2	`C6....H.3...`	4.75157	94	101	40	0.0000
3	`C6....H.3...`	5.68933	110	95	35	0.0000
1	`EC2-N...12..`	0.06200	133	149	57	0.0001
2	`EC2-N...12..`	0.62447	140	147	55	0.0001
3	`EC2-N...12..`	3.81259	144	161	49	0.0000
1	`NNC-C...202.`	2.68467	153	149	54	0.0003
2	`NNC-C...202.`	2.69133	153	150	56	0.0002
3	`NNC-C...202.`	3.24763	156	175	42	0.0002
	Promoter of toxicity decrease					
1	`=...........`	-0.12780	669	673	308	0.0000
2	`=...........`	-2.18932	688	674	301	0.0000
3	`=...........`	-3.49799	739	726	247	0.0000

*) The N1, N2, and N3 are the numbers of SA in the training, invisible training, and calibration sets, respectively.



$$y = C_0 + C_1 * DCW(T^*, N^*)$$

Figure 1

The general representations of the traditional correlation vs. “semi correlation”

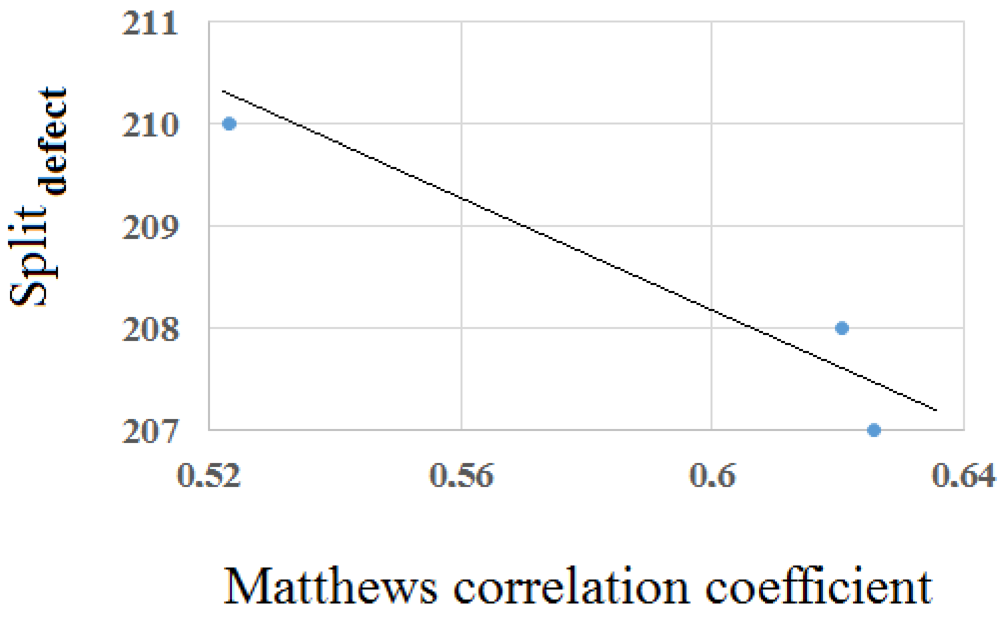


Figure 2

The diagram of the correlation between the Matthews correlation coefficient for external validation set and the split defect, that is calculated with Eq. 9.