**CORAL software: analysis of impacts of pharmaceutical agents upon metabolism via**

**the optimal descriptors**

Mariya A. Toropova1, Ivan Raska Jr2, Alla P. Toropova3, \*, Maria Raskova2

**Published version of this paper can be found here:**

**DOI : 10.2174/1389200218666170301105916**

[**http://www.eurekaselect.com/150558**](http://www.eurekaselect.com/150558)

**In-House Processing From**

Please fill this form properly and enclose with each manuscript before submitting to the composer.

I certify that the submitted manuscript has not been published in any Bentham Science Journal.

**JMS CODE = BSP-CDM-2016-HT11-1**

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No.** | **Description** | **Y** | **N** |
| 1. | Soft copy of the manuscript | y |  |
| 2. | Telephone, Fax and E-mail address of Corresponding author | y |  |
| 3. | Photograph of the main author | y |  |
| 4. | Graphical Abstract | y |  |
| 5. | Abstract / Structures | y |  |
| 6. | Key Words | y |  |
| 7. | Copyright Letter | y |  |
| 8. | References in the text and References section are according to the journal’s format | y |  |
| 9. | All figures and tables are properly cited within the text | y |  |
| 10. | Soft copies of figures provided (especially structures drawn in ChemDraw Software) | y |  |
| 11. | Acknowledgements section available | y |  |
| 12. | Conflict of Interest present | y |  |
| 13. | Supplementary Material provided | y |  |

**Name of Journal: \_\_**CDM**\_\_\_\_ Vol. & Issue No:** \_\_\_18-6**\_\_\_\_\_\_\_\_\_\_\_**

**Title of the Article:** CORAL software: analysis of impacts of pharmaceutical agents upon metabolism via the optimal descriptors\_\_\_\_\_\_

**Corresponding Author:**  Dr. Alla P. Toropova \_\_\_\_\_\_\_\_ **Pub Med ID:\_\_\_\_\_\_** **\_\_\_\_\_\_\_\_\_**

**MS Date of Receipt: \_\_** **9/29/2016 \_\_MS Date of Revision: \_\_ 10/17/2016 \_\_MS Date of Acceptance: \_** **1/17/2017 \_**

**Name of Manager: \_**Nida Badar\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ **Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

|  |  |
| --- | --- |
| **For Copy Editors:**  Name of Copy Editor: \_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Manuscript sent to copy editor on: \_\_\_\_\_\_\_\_\_\_\_\_\_\_  Edited version sent to manager on: \_\_\_\_\_\_\_\_\_\_\_\_\_\_  Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Remarks: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | **For Composers:**  Name of Composer: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Manuscript sent to composer on: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Compsoed version sent to manager on: \_\_\_\_\_\_\_\_\_\_\_\_\_  Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Remarks: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

**CORAL software: analysis of impacts of pharmaceutical agents upon metabolism via**

**the optimal descriptors**

Mariya A. Toropova1, Ivan Raska Jr2, Alla P. Toropova3, \*, Maria Raskova2

*1 Dipartimento di Scienze Farmaceutiche, Universita‘ degli Studi di Milano,Via L. Mangiagalli, 25, 20133 Milan, Italy*

*2 3rd Department of Medicine-Department of Endocrinology and Metabolism, First Faculty of Medicine, Charles University in Prague and General University Hospital in Prague, UNemocnice1, 12808 Prague 2, Czech Republic*

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

**Abstract:**

**Backgrounds.** The CORAL software has been developed as a tool to build up quantitative structure–activity relationships (QSAR) for various endpoints.

**Objective**. The task of the present work was to estimate and to compare QSAR models for biochemical activity of various therapeutic agents, which are built up by the CORAL software.

**Method.** The Monte Carlo technique gives possibility to build up predictive model of an endpoint by means of selection of so-called correlation weights of various molecular features extracted from simplified molecular input-line entry system (SMILES). Descriptors calculated with these weights are basis for building up correlations "structure – endpoint".

**Results**. Optimal descriptors, which are aimed to predict values of endpoints with apparent influence upon metabolism are crytically compared in aspect of their robustness and heuristic potential. Arguments which are confirming the necessity of reformulation of basics of QSARs are listed: (i) each QSAR model is stochastic experiment. The result of this experiment is defined by distribution into the training set and validation set; (ii) predictive potential of a model should be checked up with a group of different splits; and(iii) only model stochastically stable for a group of splits can be estimated as a reliable tool for the prediction. Examples of the improvement of the models previously suggested are demonstrated.

**Conclusions.** The current version of the CORAL software remains a convenient tool to build up predictive models. The Monte Carlo technique involved for the software confirms the principle “QSAR is a random event” is important paradigm for the QSPR/QSAR analyses.

**Keywords:** Metabolism, toxicology, pharmacology, QSAR, CORAL, validation, OECD principle

\*) Corresponding author

Alla P. Toropova

Laboratory of Environmental Chemistry and Toxicology,

IRCCS - Istituto di Ricerche FarmacologicheMario Negri,

Via La Masa 19, 20156 Milano, Italy

Tel: +39 02 3901 4595

Fax: +39 02 3901 4735

Email: [alla.toropova@marionegri.it](mailto:alla.toropova@marionegri.it)

**OPTIMAL DESCRIPTORS:**

**A NEW GENERATION OF DESCRIPTORS FOR QSPR/QSAR**

In this review, the applications of the CORAL free software is analyzed and possible ways to improve the system of building up predictive models are discussed.

Optimal descriptors have become one of the most interesting topics of researches in the academia as well as the chemical industry in recent times. In this section, we have attempted to define the optimal descriptors considering their brief history, representation of features of the molecular structure, mechanistic interpretation, and applicability domain [1-6].

There are a number of tasks related to medicine, in particular drug discovery. Metabolism of pharmaceutical agents is very complex phenomenon. Pharmaceutical agents as rule are toxic substances. Under such circumstances, the necessity of the systematization of pharmaceutical agents according to their toxicity becomes task no less important than knowledge about their positive abilities. Quantitative structure – property / activity relationships (QSPRs/QSARs) are a more or less convenient tool to solve these tasks.

There are a large number of programs available on the Internet, which are related to the QSPR/QSAR analysis. However, unfortunately, the majority of these programs is very specialized and sometimes not convenient for the potential user.

Often a program which generated molecular descriptor (e.g. DRAGON, <https://chm.kode-solutions.net/products_dragon.php>; ODDesripotrs, <http://www.softpedia.com/get/Science-CAD/ODDescriptors.shtml>;  PaDEL, <http://padel.nus.edu.sg/software/padeldescriptor>)

is not able to build up a model, whereas other programs

(e.g.  BuildQSAR, <http://www.profanderson.net/files/buildqsar.php>; VEGA,  [http://www.vegazz.net//](http://www.vegazz.net/);   E-DRAGON, <http://146.107.217.178/lab/edragon/>;  Genetic Algorithm(GA), <http://teqip.jdvu.ac.in/QSAR_Tools/>) give possibility to build up a model, but with descriptors which should be imported from other source.

The CORAL software which is discussed in this review has rare enough quality: this free software is generating special optimal descriptors and building up models [1-6].

There is wide variety of approaches aimed to predict metabolism of different pharmaceutical agents using the QSAR [7-13].  In this review, optimal descriptors, which are calculated with simplified molecular input-line entry systems (SMILES) [14-16]   and some alternative approaches are compared as a tool to predict various endpoints.

The drug discovery establishment has been probably one of the original industries to appreciate the QSPR/QSAR technology and remains the most important user. In fact, the drug discovery protocol needs to define two groups of endpoints related to new molecular entities which relate to both the therapeutic and toxic effects [3-6].

***History of optimal descriptors***

The first version of the optimal descriptor has been suggested by Randic and Basak [17]. The modification of diagonal elements of adjacency matrix with further calculation of topological indices by usual manner is the main idea of the descriptor suggested by Randic and Basak [17]. This approach gave promising results; in particular, the descriptor in the case of normal boiling point of organic compounds provides better accuracy in comparison with multiple linear regression analysis [18].

Instead of the above-mentioned modification of the adjacency matrix [17] one can define an optimal descriptor which is a mathematical function of so-called correlation weights of different topological features of the molecular graph [19]. Finally, the optimal descriptor can be defined as a mathematical function of the correlation weights of molecular features extracted from SMILES [20].

The intensive applications of optimal descriptors to solution of physicochemical and biochemical tasks [1-5, 21-49] are beginning approximately ten years ago owing to their availability on the Internet via the CORAL software.

***Definition of optimal descriptors***

The first definition of the optimal descriptor according to works [17, 18] can be formulated as the following: “Optimal descriptor is a topological index calculated with taken into account the presence of heteroatoms (oxygen or nitrogen) via replacing of zero values of the diagonal elements of the adjacency matrix by special coefficients which give improvement of the correlation of endpoint with this index for a set of compounds" [17, 18].

The next version of the optimal descriptor [19] can be defined as the following: “Optimal descriptor is a mathematical function of all molecular features extracted from molecular graph”.

***Calculation of optimal descriptors***

In praxis, the optimal descriptor is a coefficient, which is calculated with so-called correlation weights of structural attributes *SAk* extracted from SMILES or from molecular graph:

 (1)

The features (*SAk*) of the molecular structure extracted from graph are extended connectivity of various orders, numbers of path length two or three, valence shells of second or third orders, special descriptors related to rings (size, presence or absence of heteroatoms, aromaticity) [20].

The features extracted SMILES are SMILES atoms, i.e. one character or two characters which cannot be examined separately (e.g. ‘Cl’, ‘Br’, etc.), global SMILES attributes calculated according to presence or absence of nitrogen, oxygen, Sulphur, phosphorus, halogens, double, triple bonds [20].

The numerical values of correlation weights are calculated by the Monte Carlo optimization procedure which gives values of correlation weights CW(*Xk*) maximized correlation coefficient between the *DCW(T,N)* and endpoint (*E*) for compounds of the training set:

 (2)

Having the numerical data on the CW(*SAk*) one can calculate the DCW(T,N) for compounds of the training set and build up model:

 (3)

***Applicability domain***

The applicability domain (AD) is a characteristic of developed QSAR model that can be applied for further validation. The AD is defined as biological, structural or physico-chemical space, knowledge or information on which the model of the training set is developed, and for which it is applicable to make predictions for new compounds. The QSPR models are more reliable if predicted compounds are within the applicability. However, when a compound is much dissimilar to all compounds of the modeling set, a reliable prediction of its property is uncertain. For reasons stated above defining AD is one of the main aims of all developed QSAR models.

The distribution into the “visible” training set (for the described approach the “visible” training set contains also invisible training, calibration, and validation sets) and “invisible” validation set has apparent influence upon the predictability of a model. A possible measure of the quality of the split can be as follows:

 (4)

where the probability of attribute SA in the sub-training set P(SA) and the probability of SA in the test set or in the calibration set P’(SA) are calculated by

 (5)

where *Nset(SA)* is the number of SMILES which contains SA and *Nset* is the total number of SMILES in the set. The defect is calculated with active (not blocked) SA only. If the defect = 0, the split should be estimated as “ideal” one. However, in fact, this situation is not possible. However, the value of the defect calculated with Eq. 4 gives possibility to compare various splits.

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

 (6)

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

 (7)

The probabilistic domain of applicability can be defined via inequality

 (8)

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. 7.

***Mechanistic interpretation***

The developed QSAR models allow for mechanical interpretation of the studied phenomena. 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).

However, it should be noted that prevalence of attributes extracted from SMILES (or from graph) in the training and calibration sets plays an important role. It is clear that, SMILES attributes (and corresponding molecular features) which are absent in the training set cannot be examined as basis of the mechanistic interpretation even if their correlation weights are stably positive (or stably negative). All molecular features, which are examined in this work as promoters of increase or decrease of endpoints have significant prevalence in the training and calibration sets (Tables 4-6).

***Representation of features of the molecular structure***

Optimal descriptors are calculated with different features (*Xk*) of the molecular structure. The CORAL software gives possibility to use as the representation of the molecular structure (i) SMILES able to reflect directly presence of various chemical elements, various covalent chemical bonds, presence of cycles, and others.; (ii) hydrogen-suppressed graph (HSG) represents chemical elements which are not hydrogen atoms; hydrogen-filled graph (HFG) represents chemical elements including hydrogen atoms; and (iii) graph of atomic orbitals reflects presence of different atomic orbitals such as 1s2, 2p5, 4d10, and others.

The combination of molecular features is defined by special options. The combination can contain attributes extracted solely from SMILES, or solely from HSG, or solely from HFG, or solely from GAO. In addition, there are possibilities to utilize hybrid optimal descriptors (i) SMILES and HSG; (ii) SMILES and HFG; and (iii) SMILES and GAO. This work suggested so-called super-attribute as alternative for preliminary utilized ATOMPAIR [20]. The definition of the attribute based on preliminary suggested BOND, NOSP, and HALO [20], which are combinations of zero (if absent) and unit (if takes place) in accordance with presence or absence of the following molecular features (i) double bonds “=”; (ii) triple bonds “#”; (iii) stereo chemical bonds “@”; (iv) nitrogen “N”; (v) oxygen “O”; (vi) Sulphur “S”; (vii) phosphorus “P”; (viii) fluorine “F”; (ix) chlorine “Cl”; (x) bromine “Br”; and (xi) iodine “I”. Table 1 contains example of the twelve symbols, which are representation of the super-attribute.

***Applications of optimal descriptors***

In fact, the majority of therapeutic agents is aimed to impact human metabolism [50]. Unfortunately, as rule, the pharmaceutical effect for defined target in organism is accompanied by specific toxic phenomena. The positive pharmaceutical effect as well as toxicological effect is characterized by various endpoints. Reliable QSAR models for these endpoints are valuable data for medicinal chemistry. The optimal descriptors are involved in the solution of this task.

***Medicinal tasks***

The review of the recent literature gives possibility to extract works, which are carried out with the optimal descriptors and aimed to solve various tasks of the medicinal chemistry.

The number of targets of pharmaceutical agents (diseases) increases every year. However, there are “classic” diseases, such as malaria [21-23], tuberculosis [24], hepatitis [25], cardiac diseases [26], Alzheimer disease [27], Anti-HIV-1 [28-34], sarcoma [35], influenza virus [36, 37], and many others [38-40]. The thyroid hormones are tyrosine-based hormones produced by the thyroid gland that are primarily responsible for regulation of metabolism [41].

**SEARCH FOR POTENTIAL PHARMACEUTIC AGENTS**

***Anticancer activity***

Mullen et al. [42] have developed a model for anticancer activity of triphenylmethyl pharmacophore. The linear additive model of carcinogenicity has been used as a basis of some heuristic principles of the CORAL software in terms of the probability theory [43]. Furthermore, the balance of correlations has been used as tool to build up model for carcinogenicity [44], and the balance of correlations with ideal slopes [45] gave some improvement of the accuracy and reliability of the model. The conception of the CORAL models gives possibility to benchmarking of the SMILES attributes as participants of the QSAR model for carcinogenicity and anticancer activity [46]. In addition, the QSAR model carcinogenicity has shown that sometimes the outliers are able to be useful for development of a model, since they also contain valuable hints on the endpoint [47]. The attempt of standardization of the Monte Carlo algorithms is carried out in work [48]. The possibility to compare correlations of molecular features with mutagenicity and carcinogenicity is demonstrated in work [49]. There is some homology of results obtained in work [49] with results described in the literature [51]. The QSAR for antineoplastic agents based on the optimal descriptors is suggested in work [52].

The QSAR model for large-scale of aromatase inhibitors (AIs are a class of drugs used in the treatment of breast cancer in postmenopausal women and gynecomastia in men) was built up with using SMILES-based descriptors in work [53]. There are interest to enzyme AKR1C3 (Aldo-keto reductase family 1 member C3) as an indicator of way to find a potential effective anticancer agent [54-60]. The AKR1C3 is a participant of carcinogenic processes. In particular, the 2,3-diarylpropenoic acids are effective inhibitors of the enzyme AKR1C3. The QSAR model for inhibitory rate of group of 2,3-diarylpropenoic acids based on optimal descriptors is built up and suggested in the literature [61].

In addition to QSAR models of positive effect of therapeutic agents, the potential carcinogenicity of large set of drugs by the Monte Carlo method using the CORAL software is analyzed, in work [62] database available on the Internet [63] has been used for this analysis.

Thus, the Monte Carlo method for the above-listed medicinal tasks (as well as other accompanied tasks [64, 66-68, 70, 71]) gave statistically significant models [64-77].

***Anticonvulsant activity***

Garro Martinez et al. [64], using the optimal descriptors calculated with SMILES have developed QSAR model for therapeutical efficacy of different anticonvulsants.

***Anti-malaria activity***

QSAR models, based on the optimal descriptors for anti-malaria activity, are suggested in a few works [21-23].

***Anti-influenza drugs***

The model for the H1N1 neuraminidase inhibitors from influenza a virus based on the optimal descriptors is described in work [36,37].

***Vitamins***

Octanol/water partition coefficient is a physicochemical parameter that is usually correlated with different kinds of biological activity. Taking into account biochemical role of vitamins the prediction of the parameter given above for vitamins becomes important task [66, 67]. The optimal descriptors for this parameter give quite good results comparable with models obtained by means of other approaches [66, 67].

***Inhibitors of serine proteases***

Selective inhibitors of target serine proteinases have a potential therapeutic role for the treatment of various inflammatory diseases. The existence of a protease–antiprotease imbalance is generally associated to depressed levels of physiological protein inhibitors. It is for this reason that there has been so much interest in developing highly selective and potent irreversible inhibitors of serine proteases. García et al., [79] have built up the SMILES-based QSAR model for inhibitors of serine proteinases.

***Psychotropic drugs***

The influence of psychotropic drugs to metabolism is very complex and important task of the medicinal chemistry [65, 68, 73]. The CORAL free software gives QSAR models for toxicity of psychotropic drugs which are characterized by reasonable well predictive potential and which have mechanistic interpretation [65, 68, 73].

Thus, search for potential pharmaceutic agents using the CORAL software can be considered as successful. Likely, in the future, the list of pharmacological agents studied with the optimal descriptors will be extended [80-85].

***Prediction of endpoints, which are related to peptides***

The valuable advantage of the optimal descriptors is possibility to build up a model of an endpoint based on eclectic data. For instance, the optimal descriptors give possibility to build up model for antimicrobial activity of peptides using as the basis to build up a model so-called quasi-SMILES which in this case is the sequence of one-letter abbreviations of corresponding amino acids [86-88].

***Prediction of endpoints, which are related to nanomaterials***

Finally, it should be noted that optimal descriptors can be used as a tool for the predictive modelling in the case of nanomaterials [28-30, 89-99]. The experimental data on the nanomaterials are eclectic since these experiments are carried out with very exotic equipment and under extraordinary conditions. Under such circumstances, the above-mentioned quasi-SMILES become convenient basis for models of various endpoints related to nanomaterials [28-30, 89-99]. It has to be noted that the influence of nanomaterials to human metabolism takes place via manifold cosmetics, food, and medical treatments.

Thus, the CORAL software can be recognized as a reasonable, useful, and validated tool for the QSPR/QSAR analyses. Whether this software can be improved? In order to answer the question, three endpoints related to drug discovery are examined in this work. The comparison of previously developed CORAL models with fresh improved model can be interesting and useful from heuristic point of view. The comparison of CORAL models with models which are built up using other approaches also can be interesting and useful from practical point of view. Finally, it is important to check up the idea suggested in work [90,93]: whether QSAR is a random event, in the case of building up models with a set of random splits into the training set and the validation set, in aspect of the variations of the predictive potential of these models.

**FRESH IMPROVED MODELS, WHICH ARE CALCULATED WITH THE CORAL**

***Endpoint 1. Therapeutic agents for the treatment of Alzheimer disease:*** The binding affinity data (*IC50* nM converted into negative decimal logarithm *pIC50*) of 233 gamma-secretase inhibitors and their simplified molecular input-line entry system (SMILES) were taken from the literature [27]. Three random splits into the training (≈35%), invisible training (≈35%), calibration (≈15%), and validation (≈15%) sets have been defined.

***Endpoint 2. Cardiac toxicity (hERG blocker compounds):*** The numerical data of cardiac toxicity (*IC50*, half-maximal response dose the endpoint is *pIC50*, i.e. the negative decimal logarithm of the *IC50*) for 400 compounds were taken from the literature [26]. However, six compounds are duplicated in data from [100]. Thus, the total number of cardiac toxic compounds examined in this work is 394. Three random splits into the training (≈35%), invisible training (≈35%), calibration (≈15%), and validation (≈15%) sets have been defined.

***Endpoint 3. The p53-HDM2 inhibitors as antiproliferative agents:*** The set of compounds which have p53-HDM2 inhibitory activity (n=155) taken in the literature [101] were examined in this work. The negative decimal logarithm of inhibitor concentration (*pIC50*) is examined as endpoint. Three random splits into the training (≈30%), invisible training (≈30%), calibration (≈20%), and validation (≈20%) sets have been defined.

***Distribution available data into the training and validation sets:*** In this work, the training set is structured and contains three components: training, invisible training, and calibration sets. The validation set is a group of compounds which are not involved in the process of building up a model.The total sets of the above-mentioned three endpoints were split into the above-mentioned four special sets: (i) the training set (≈35%); (ii) invisible training set (≈35%); (iii) calibration set (≈15%); and (iv) validation set (≈15%). The functions of these sets are as follows. The training set is builder of model. The invisible training set is inspector of the current status of the model (whether the model has evolution to real predictive ability or the model is exclusive property of the training set). The calibration set is indicator of absence of overtraining. Computational experiments have shown that there is maximum of the correlation coefficient for the calibration set: after the maximum as rule the overtraining takes place. The validation set gives possibility of final estimation of the predictive potential for the model.

Computational experiments have shown that there are successful and unsuccessful distributions into the training, invisible training, calibration, and validation sets. The successful distribution is characterized by good statistical quality of a model for the external validation set. Consequently, one should check up an approach with several distributions in order to obtain real estimation of the predictive potential of a model.

Table 2 contains a group of examples of molecular features, which can be involved for building up a CORAL model.

***Examples of building up the coral models***

Different CORAL methods were involved to build up models for three endpoints examined in this work. However, in all cases, the same number of epochs (ten) and the same start values (1.5) of the Monte Carlo optimization procedure are used.

***Endpoint 1***

The hybrid optimal descriptor utilized to build up models is as follows:

 (9)

where





The models for three different split into the training (training-invisible training-calibration) and validation sets are the following:

*pIC50* = 0.0000191 (± 0.0307575) + 0.0766192 (± 0.0003023) \* DCW(1,10) (10)

*pIC50* = 0.0041613 (± 0.0441199) + 0.0758459 (± 0.0004240) \* DCW(1,10) (11)

*pIC50* = -0.0009292 (± 0.0309708) + 0.0786591 (± 0.0003255) \* DCW(1,10) (12)

***Endpoint 2***

The SMILES based optimal descriptor utilized to build up models is as follows:

 (13)

The models for three different splits into the training (training-invisible training-calibration) and validation sets are as follows:

*pIC50* = -1.6674012 (± 0.0025895) + 0.0827729 (± 0.0001780) \* DCW(1,10) (14)

*pIC50* = -1.8795895 (± 0.0033574) + 0.0848448 (± 0.0002020) \* DCW(1,10) (15)

*pIC50* = -1.8294581 (± 0.0032312) + 0.0681604 (± 0.0001136) \* DCW(1,10) (16)

***Endpoint 3***

The hybrid optimal descriptor utilized to build up models is as follows:

 (17)

where





The models for three different splits into the training (training-invisible training-calibration) and validation sets are as follows:

*pIC50* = 0.0004414 (± 0.0392571) + 0.0251339 (± 0.0002250) \* DCW(1,10) (18)

*pIC50* = -0.0000281 (± 0.0242385) + 0.0216757 (± 0.0001221) \* DCW(1,10) (19)

*pIC50* = 0.0010377 (± 0.0335129) + 0.0309062 (± 0.0002228) \* DCW(1,10) (20)

Table 3 contains statistical characteristics of the models calculated with Eqs. 10-12; 14-16; and 18-20 together with the numerical data on Y-randomization test [102].

***Examples of utilization of the CORAL models***

Optimal descriptors are untypical descriptors. In contrast to traditional descriptors, which are calculated with molecular graphs or which are physicochemical endpoints, the optimal descriptors are calculated with directly data on the endpoint of compounds from the training set, in order to predict the endpoint for the external validation set.

Under such circumstances, the impact of the distribution into the training set and the validation set becomes significantly more powerful factor in comparison with QSPR/QSAR models based on traditional descriptors. This is the disadvantage of these descriptors. However, one can reduce influence of this circumstance by means of selection of rational distributions: percentage of available substances in training set should be larger than percentage of substances in validation set. In this case, optimal descriptors can be more effective than traditional descriptors (topological indices, 3D descriptors, physicochemmical parameters).

One can see (Table 3) all suggested models are better in comparison with previous models. Consequently, the answer on the question “whether a CORAL model can be improved?” should be “yes”. The improvement obtained in this work is caused by new "super-attribute" (Table 1) as well as by the involvement of the hybrid optimal descriptors.

As it is noted above, after several runs of the optimization one can obtain the mechanistic interpretation for the CORAL model.

***Endpoint 1.*** In the case of the endpoint 1, presence of oxygen “O...(.......”, presence of chlorine “Cl..........” are promoters of increase for endpoint 1, whereas, presence of nitrogen “N...........”, and presence of aromaticity “1...c...(...” are promoters of decrease for endpoint 1 (Table 4).

***Endpoint 2.*** In the case of the endpoint 2, presence of aromaticity “c...c...c...”, presence of two rings “c...c...2...”, and presence of nitrogen “N...(...C...” are promoters of increase for endpoint 2, whereas, presence of oxygen “O...........”, and presence of double bonds “=...........” are promoters of decrease for endpoint 2 (Table 5).

***Endpoint 3.*** In the case of the endpoint 3, presence of rings “1...........”, “2...........”, “4...........”, presence of vertex degree equal to 6 for vertex 1s2 in GAO “EC0-1s2 6...”, and presence of vertex degree equal to 3 for vertex 2p4 in GAO “EC0-2p4 3...” are promoters of increase for endpoint 3, whereas, branching “(...........”,”C...(.......” and presence of nitrogen “N...........” as well presence of vertex 2s2 with NNC equal to 725 are promoters of decrease for endpoint 3 (Table 6).

In addition, the improvement of the software can also be obtained by means of so-called quasi-SMILES described in the literature [95]. The quasi-SMILES give possibility to modify the paradigm of the QSPR/QSAR models by means of the involvement of all available eclectic parameters and circumstances, which are able to impact endpoint, which should be predicted.

Finally, the CORAL software can be improved via accumulation of new ideas from other approaches described in the literature [103-106].

**POSSIBLE FUTURE OF THE OPTIMAL DESCRIPTORS**

Optimal descriptors as well as the CORAL software have possibilities to adaptation for fresh tasks**.** In particular, the CORAL software is able to build up a model based on quasi-SMILES described in the literature. The quasi-SMILES is a sequence of symbols, which are representation of arbitrary eclectic data, which can have impact upon endpoint that should be predicted [93-99]. The list of attributes, which are extracting from SMILES or from molecular graph, can be modified. This can change predictive potential of the CORAL models. The algorithm of the Monte Carlo optimization can also be improved.

The current version of the CORAL free software remains a convenient tool to build up predictive models. The software has been checked up in a few works related to various endpoints. The CORAL free software gives possibility to build up QSPR/QSAR models according to the OECD principles [107]. The CORAL models can be improved by means of the utilization of the hybrid optimal descriptors. The Monte Carlo technique involved for the software confirms the principle “QSAR (as well QSPR) is a random event” is important paradigm for the QSPR/QSAR analyses. According to this principle, the real validation of a model is the checking up of the model for group of different splits into the training set and validation set. The evolution of Optimal Descriptors is not completed [108]. Hopefully it will be continued.

**CONFLICT OF INTEREST**

The authors confirm that this article content has no conflict of interest.

**ACKNOWLEDGEMENTS**

Authors thank the EC project PeptiCAPS (Project reference: 686141).

**REFERENCES**

[1] Toropov, A.A.; Benfenati, E. SMILES in QSPR/QSAR modeling: Results and perspectives. *Curr. Drug Discov. Technol*., **2007**, *4* (2), 77-116.

[2] Toropov, A.A.; Toropova, A.P.; Benfenati E.; Gini G. OCWLGI Descriptors: Theory and Praxis. *Curr. Comput. Aided Drug Des*., **2013**, *9*, 226-232

[3] Toropov, A.A.; Toropova, A.P.; Raska, Jr I.; Leszczynska D., Leszczynksy J. Comprehension of drug toxicity: Software and databases. *Comput. Biol. Med.*, **2014**, *45*, 20-25.

[4] Toropov, A.A.; Toropova, A.P.; Nesmerak, K.; Veselinović, A.M.; Veselinović, J.B.; Leszczynska, D.; Leszczynski, J. Development of the latest tools for building up “nano-QSAR”: Quantitative features—property/activity relationships (QFPRs/QFARs). Chapter 12, In Book: Practical Aspects of Computational Chemistry IV. Leszczynski, J.; Shukla, M.K. (Eds.). Springer **2016**. ISBN 78-1-4899-7699-4, p. 353-396. DOI: 10.1007/978-1-4899-7699-4\_12

[5] Toropov, A.A.; Toropova, A.P.; Benfenati, E.; Nicolotti, O.; Carotti, A.; Nesmerak, K.; Veselinović, A.M.; Veselinović, J.B.; Duchowicz, P. R.; Bacelo, D.; Castro, E. A.; Rasulev, B. F.; Leszczynska, D.; Leszczynski, J. QSPR/QSAR analyses by means of the CORAL software: results, challenges, perspectives. Chapter 15, in Book: Roy, K. (Eds.) Quantitative Structure-Activity Relationships in Drug Design, Predictive Toxicology, and Risk Assessment. Hershey, **2015**. PA: IGI Global, p. 1-531. doi:10.4018/978-1-4666-8136-1

[6] Melagraki, G.; Afantitis, A.; Toropov, A.A.; Sarimveis,H.; Igglessi – Markopoulou . O. QSPR-QSAR Studies on Desired Properties for Drug Design. Chapter 4, in Book: Castro, E.A. (Eds.) QSAR models constructed by optimal descriptors and by multiple regression analysis for the prediction of carbonic anhydrase II inhibitory activity of substituted aromatic sulphonamides. **2010**: 95-116 ISBN: 978-81-308-0404-0

[7] González-Díaz, H.; Arrasate, S.; Juan, A.G.-S.; Sotomayor, N.; Lete, E.; Speck-Planche, A.; Ruso, J.M.; Luan, F.; Cordeiro, M.N.D.S. Matrix trace operators: From spectral moments of molecular graphs and complex networks to perturbations in synthetic reactions, micelle nanoparticles, and drug ADME processes. *Curr. Drug Metab.,* **2014**, *15* (4), 470-488.

[8] Vergara-Galicia, J.; Prado-Prado, F.J.; González-Díaz, H. Galvez-markov network transferability indices: Review of classic theory and new model for perturbations in metabolic reactions. *Curr. Drug Metab.*, **2014**, *15* (5), 557-564.

[9] De Julián-Ortiz, J.V.; Zanni, R.; Gálvez-Llompart, M.; García-Domenech, R. The prediction of human intestinal absorption based on the molecular structure. *Curr. Drug Metab.,* **2014**, *15* (4), 380-388.

[10]Chen, A.; Yarmush, M.L.; Maguire, T. Physiologically based pharmacokinetic models: Integration of in silico approaches with micro cell culture analogues. *Curr. Drug Metab.*, **2012**, *13* (6), 863-880.

[11] Khan, M.T.H. Predictions of the ADMET properties of candidate drug molecules utilizing different QSAR/QSPR modelling approaches. *Curr. Drug Metab.*, **2010**, *11* (4), 285-295.

[12] Ekins, S.; Iyer, M.; Krasowski, M.D.; Kharasch, E.D. Molecular characterization of CYP2B6 substrates. *Curr. Drug Metab.*, **2008**, *9* (5), 363-373.

[13] Mankowski, D.C.; Ekins, S. Prediction of human drug metabolizing enzyme induction

*Curr. Drug Metab.*, **2003**, *4* (5), 381-391.

[14] Weininger, D. Smiles. 3. Depict. Graphical depiction of chemical structures. *J. Chem. Inf. Comput. Sci.,* **1990**, *30* (3), 237-243.

[15] Weininger, D.; Weininger, A.; Weininger, J.L. SMILES: 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.*, **1989**, 29 (2), 97-101.

[16] Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.,* **1988**, *28*(1), 31-36.

[17] . Randic, M.; Basak, S. C. Optimal Molecular Descriptors based on weighted path numbers. *J. Chem. Inf. Comput. Sci.,* **1999**, *39*, 261-266.

[18] Randic, M.; Pompe, M. The variable connectivity Index 1Xf versus the traditional molecular descriptors: A comparative study of 1Xf Against descriptors of CODESSA  *J. Chem. Inf. Comput. Sci.,* **2001**, *41*, 631-638.

[19] Toropov, A.A.; Toropova, A.P. QSAR modeling of toxicity on optimization of correlation weights of Morgan extended connectivity. *J. Mol. Struct: THEOCHEM,*  **2002**, *578*(1-3), 129-134.

[20] Toropova, A.P.; Toropov, A.A. CORAL software: Prediction of carcinogenicity of drugs by means of the Monte Carlo method. *Eur. J. Pharm. Sci.* **2014**, *52*, 21-25.

[21] Toropov, A.A.; Toropova , A P.; Benfenati, E.; Gini, G.; Fanelli, R. The definition of the molecular structure for potential anti-malaria agents by the Monte Carlo method. *Struct. Chem*. **2013**, *24*,1369–1381.

[22] Masand,V.H.; Toropov, A.A.; Toropova, A P.; Mahajan, D.T. QSAR models for anti-malarial activity of 4-aminoquinolines. *Curr. Comp. Aided Drug Des*., **2014**, *10*, 75-82.

[23] Ibezim, E.; Duchowicz, P.R.; Ortiz, E.V.; Castro, E.A. QSAR on aryl-piperazine derivatives with activity on malaria. *Chemometr. Intell. Lab. Syst*., **2012**, *110* (1), 81-88.

[24] Toropov, A.A.; Toropova , A P.; Benfenati, E.; Fanelli, R. QSAR as a random event: selecting of the molecular structure for potential anti-tuberculosis agents. *Anti Infect.Agents*, **2016**, *14*(1), 3 – 10.

[25] Worachartcheewan, A.; Prachayasittikul, V.; Toropova, A.P.; Toropov, A.A.; Nantasenamat, C. Large-scale structure-activity relationship study of hepatitis C virus NS5B polymerase inhibition using SMILES-based descriptors. *Molec. Divers*., **2015**, *19*, 955–964.

[26] Gobbi, M.; Beeg, M.; Toropova, M.A.; Toropov, A.A.; Salmona, M. Monte Carlo method for predicting of cardiac toxicity: hERG blocker compounds. *Toxicol. Lett*., **2016**, *250-251*, 42-46.

[27] Toropova, M.A.; Toropov, A.A.; Raška, I. Jr; Rašková, M. Searching therapeutic agents for treatment of Alzheimer disease using the Monte Carlo method. *Comput. Biol. Med*., **2015**, *64*(1), 148–154.

[28] Toropova, A.P.; Toropov, A.A.; Benfenati, E.; Leszczynska, D.; Leszczynski, J. QSAR modeling of measured binding affinity for fullerene-based HIV-1 PR inhibitors by CORAL. *J.Math. Chem*., **2010**, *47*, 959-987.

[29] Toropov, A.A.; Toropova, A.P.; Benfenati , E.; Leszczynska, D.; Leszczynski, J. SMILES-based optimal descriptors: QSAR analysis of fullerene-based HIV-1 PR inhibitors by means of balance of correlations. *J. Comput. Chem*., **2010**, *31*, 381-392.

[30] Toropov, A.A.; Toropova, A.P.; Benfenati , E.; Leszczynska, D.; Leszczynski, J. InChI-based optimal descriptors: QSAR analysis of fullerene[C60]-based HIV-1 PR inhibitors by correlation balance. *Eur. J. Med. Chem*., **2010**, *45*, 1387–1394.

[31] Toropova, A.P.; Toropov, A.A.; Benfenati , E.; Gini, G. Simplified molecular input-line entry system and international chemical identifier in the QSAR analysis of styrylquinoline derivatives as HIV-1 integrase inhibitors. *Chem. Biol. Drug Des*., **2011**, *77*, 343-360.

[32] Toropov, A.A.; Toropova, A.P.; Raska, I. Jr ; Benfenati , E.; Gini, G. Development of QSAR models for predicting anti-HIV-1 activity using the Monte Carlo method . *Cent.Eur. J. Chem*., **2013**, *11*, 371-380.

[33] Veselinović, J.B.; Veselinović, A.M.; Toropov, A.A.; Toropova,A.P.; Damnjanović, I.; Nikolić, G. Monte Carlo method based QSAR modeling of coumarin derivates as potent HIV‐1 integrase inhibitors and molecular docking studies of selected 4‐phenyl hydroxycoumarins. *Sci. J. Faculty Med. Niš*, **2014**, *31*(2), 95-103.

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

[35] Toropov, A.A.; Toropova, A.P.; Benfenati, E.; Gini, G.; Leszczynska, D.; Leszczynksy, J. CORAL: Classification model for predictions of anti-sarcoma activity. *Curr. Top. Med. Chem*., **2012**, *12*, 2741-2744.

[36] Karthick, V.; Toropova, A.P.; Toropov, A.A.; Ramanathan, K. Discovery of potential, non-toxic influenza virus inhibitor by computational techniques. *Mol. Inform*., **2014**, *33* (8), 559-565.

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

[38] Toropov, A.A.; Toropova, A.P.; Rasulev, B.F.; Benfenati, E.; Gini, G.; Leszczynska, D.; Leszczynksy, J. CORAL: binary classifications (Active/Inactive) for Liver-Related adverse effects of drugs. *Curr. Drug Saf*., **2012**, *7*, 257-261.

[39] Toropov, A.A.; Toropova, A.P.; Benfenati, E.; Gini, G.; Leszczynska, D.; Leszczynksy, J.; De Nucci, G. QSAR models for inhibitors of physiological impact of *Escherichia coli* that leads to diarrhea. *Biochem. Biophys. Res. Commun*., **2013**, *432*, 214-225.

[40] Sokolović, D.; Stanković, V.; Dragan Toskić, D.; Lilić, L.; Ranković, G.; Ranković, J.; Nedin-Ranković, G.; Veselinović, A.M. Monte Carlo-based QSAR modeling of dimeric pyridinium compounds and drug design of new potent acetylcholine esterase inhibitors for potential therapy of *Myasthenia Gravis*. *Struct.Chem*., **2016**, Online: 07 June 2016. DOI: 10.1007/s11224-016-0776-z

[41] Toropova, A.P.; Toropov, A.A.; Benfenati, E. CORAL: Prediction of binding affinity and efficacy of thyroid hormone receptor ligands. *J. Med. Chem*., **2015**, *101*, 452-461.

[42] Mullen, L.M.A.; Duchowicz, P.R.; Castro, E.A. QSAR treatment on a new class of triphenylmethyl-containing compounds as potent anticancer agents. *Chemometr. Intell. Lab. Syst*., **2011**, *107* (2), 269-275.

[43] Toropov, A. A.; Toropova, A. P.; Benfenati, E. Additive SMILES-Based carcinogenicity models: probabilistic principles in the search for robust predictions. *Int. J. Mol. Sci*., **2009**, *10*, 3106-3127.

[44] Toropov, A. A.; Toropova, A. P.; Benfenati, E.; Manganaro, A. QSAR modelling of carcinogenicity by balance of correlations. *Molec. Divers*., **2009**, *13*, 367-373.

[45] Toropov, A. A.; Toropova, A. P.; Benfenati, E. SMILES-based optimal descriptors: QSAR modelling of carcinogenicity by balance of correlations with ideal slopes. *Eur. J. Med. Chem.,* **2010**, *45*, 3581-3587.

[46] Toropov, A. A.; Toropova, A. P.; Benfenati, E.; Gini, G.; Leszczynska, D.; Leszczynksy, J. SMILES-based QSAR approaches for carcinogenicity and anticancer activity: Comparison of correlation weights for identical SMILES attributes. *Anticancer. Agents Med. Chem*., **2011**, *11*, 974-982.

[47] Toropova, A.P.; Toropov, A.A.; Gonella Diaza, R.; Benfenati, E.; Gini, G. Analysis of the co-evolutions of correlations as a tool for QSAR-modeling carcinogenicity: an unexpected good prediction based on a model that seems untrustworthy. *Cent. Eur. J. Chem*., **2011**, *9*, 165-174.

[48] Benfenati, E.; Toropov, A.A.; Toropova, A.P.; Manganaro, A.; Gonella Diaza, R. CORAL software: QSAR for anticancer agents. *Chem. Biol. Drug Des*., **2011**, *77*, 471-476.

[49] Toropov, A.A.; Toropova, A.P.; Benfenati, E.; Gini, G.; Leszczynska, D.; Leszczynski, J. Calculation of molecular features with apparent impact on both activity of mutagens and activity of anticancer agents. *Anticancer. Agents Med. Chem*., **2012**, *12*, 807-817.

[50] Shaked, I.; Oberhardt, M.A.; Atias, N.; Sharan, R.; Ruppin, E. Metabolic network prediction of drug side effects. *Cell Syst.,* **2016**, *2*(3), 209–213.

[51] Simon, L.; Imane, A.; Srinivasan, K. K.; Pathak, L.; Daoud, I. *In silico* drug-designing studies on flavanoids as anticolon cancer agents: pharmacophore mapping, molecular docking, and Monte Carlo method-based QSAR modeling. *Interdiscip. Sci. Comput. Life Sci.* Published online: April 8, 2016. doi:10.1007/s12539-016-0169-4

[52] Toropov, A.A.; Toropova, A.P.; Raska, Jr I.; Benfenati, E. QSPR modeling of octanol/water partition coefficient of antineoplastic agents by balance of correlations.  *Eur. J. Med. Chem.,* **2010**, *45*, 1639–1647.

[53] Worachartcheewan, A.; Mandi, P.; Prachayasittikul, V.; Toropova, A.P.; Toropov, A.A.; Nantasenamat, C. Large-scale QSAR study of aromatase inhibitors using SMILES-based descriptors. *Chemometr. Intell. Lab. Syst*., **2014**, *138*, 120-126.

[54] Brožič, P.; Golob, B.; Gomboc, N.; Rižner, T.; Gobec, S. Cinnamic acids as new inhibitors of 17-hydroxysteroid dehydrogenase type 5 (AKR1C3). *Mol. Cell. Endocrinol*., **2006**, *248*, 233–235.

[55] Gobec, S.; Brožič, P.; Rižner, T. Nonsteroidal anti-inflammatory drugs and their analogues as inhibitors of aldo-keto reductase AKR1C3: new lead compounds for the development of anticancer agents. *Bioorg. Med. Chem. Lett*., **2005**, *15*, 5170–5175.

[56] Štefane, B.; Brožič, P.; Vehovc, M.; Rižner, T.; Gobec, S. New cyclopentane derivatives as inhibitors of steroidmetabolizing enzymes AKR1C1 and AKR1C3. *Eur. J.Med. Chem*., **2009**, *44*, 2563–2571.

[57] Beranič, N.; Gobec, S.; Rižner, T. Progestins as inhibitors of the human 20-ketosteroid reductases, AKR1C1 and AKR1C3. *Chem. Biol. Interact*., **2011**, *191*, 227–233.

[58] Adeniji, A.; Twenter, B.; Byrns, M.; Jin, Y.; Winkler, J.; Penning, T. Discovery of substituted 3-(phenylamino) benzoic acids as potent and selective inhibitors of type 5 17bhydroxysteroid dehydrogenase (AKR1C3). *Bioorg. Med. Chem. Lett*., **2011**, *21*, 1464–1468.

[59] Chen, M.; Adeniji, A.; Twenter, B.; Winkler, J.; Christianson, D.; Penning, T. Crystal structures of AKR1C3 containing an N-(aryl) amino-benzoate inhibitor and a bifunctional AKR1C3 inhibitor and androgen receptor antagonist. Therapeutic leads for castrate resistant prostate cancer. *Bioorg. Med. Chem. Lett*., **2012**, *22*, 3492–3497.

[60] Sinreih, M.; Sosič, I.; Beranič, N.; Turk, S.; Adeniji, A.; Penning, T.; Rižner, T.; Gobec, S. NBenzoyl anthranilic acid derivatives as selective inhibitors of aldo-keto reductase AKR1C3. *Bioorg. Med. Chem. Lett*., **2012**, *22*, 5948–5951.

[61] Li, Q.; Ding, X.; Si, H.; Gao, H. QSAR model based on SMILES of inhibitory rate of 2, 3-diarylpropenoic acids on AKR1C3. *Chemometr. Intell. Lab. Syst*., **2014**, *139*, 132-138.

[62] Toropova, A.P., Toropov, A.A. CORAL software: Prediction of carcinogenicity of drugs by means of the Monte Carlo method. *Eur. J. Pharm. Sci*., **2014**, *52* (1), 21-25.

[63] Galvez, J. **2000.** Personal webpage. <http://www.uv.es/~galvez/tablevi.pdf> (accessed May 30, 2016).

[64] Garro Martinez, J.C ; Duchowicz, P.R.; Estrada, M.R.; Zamarbide, G.N.; Castro, E.A. QSAR study and molecular design of open-chain enaminones as anticonvulsant agents. *Int. J. Mol. Sci*., **2011**, *12* (12), 9354-9368.

[65] Roy, K.; Toropov, A.A.; Raska, Jr I. QSAR modeling of peripheral versus central benzodiazepine receptor binding affinity of 2-phenylimidazo[1,2-a]pyridineacetamides using optimal descriptors calculated with SMILES. *QSAR Comb. Sci*., **2007**, *26* (4), 460-468.

[66] Toropov, A.A.; Toropova, A.P.; Raska, Jr I. QSPR modeling of octanol/water partition coefficient for vitamins by optimal descriptors calculated with SMILES. *Eur. J. Med. Chem.*, **2008**, *43* (4), 714-740.

[67] Raska, Jr I.; Toropov, A.A. Comparison of QSPR models of octanol/water partition coefficient for vitamins and non vitamins. *Eur. J. Med. Chem.*, **2006**, *41* (11), 1271-1278.

[68] Toropova, A.P.; Toropov, A.A.; Benfenati, E.; Gini, G.; Leszczynska, D.; Leszczynski, J. QSAR modeling of anxiolytic activity taking into account the presence of keto- and enol-tautomers by balance of correlations with ideal slopes. *Cent. Eur. J. Chem*., **2011**, *9*(5), 846-854.

[69] Toropov, A.A.; Toropova, A.P.; Gonella Diaza, R.; Benfenati, E.; Gini, G. SMILES-based optimal descriptors: QSAR modeling of estrogen receptor binding affinity by correlation balance. *Struct. Chem*., **2012**, *23*, 529-544.

[70] Nesměrák, K.; Toropov, A.A.; Toropova, A.P.; Kohoutova, P.; Waisser, K. SMILES-based quantitative structure-property relationships for half-wave potential of N-benzylsalicylthioamides. *Eur. J. Med. Chem.*, **2013**, *67*, 111-114.

[71] Veselinović, A.M.; Milosavljević, J.B.; Toropov, A.A.; Nikolić, G.M. SMILES-based QSAR model for arylpiperazines as high-affinity 5-HT1A receptor ligands using CORAL. *Eur. J. Pharm. Sci*., **2013**; *48*, 532-541

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

[73] Gissi, A.; Toropov, A.A.; Toropova, A.P.; Nicolotti, O.; Carotti, A.; Benfenati, E. Building up QSAR model for toxicity of psychotropic drugs by the Monte Carlo method. *Struct. Chem*., **2014**, *25* (4), 1067-1073.

[74] Nesměrák, K.; Toropov, A.A.; Toropova, A.P. SMILES-based quantitative structure-retention relationships for RP-HPLC of 1-phenyl-5-benzylsulfanyltetrazoles. *Struct.Chem*., **2014**, *25*, 311–317.

[75] Toropova, A.P.; Toropov, A.A.; Veselinović, J.B.; Miljković, F. N.; Veselinović, A.M. QSAR models for HEPT derivates as NNRTI inhibitors based on Monte Carlo method. *Eur. J. Med. Chem.*, **2014**, *77*, 298–305.

[76] Veselinović, J.B.; Toropov, A.A.; Toropova, A.P.; Nikolic´, G. M.; Veselinović, A.M. Monte Carlo Method-Based QSAR modeling of penicillins binding to human serum proteins. *Arch. Pharm.,* **2015**, *348* (1), 62–67.

[77] Toropov, A.A.; Veselinović, J.B.; Veselinović, A.M.; Miljkovic´, F.N.; Toropova, A.P. QSAR models for 1,2,4-benzotriazines as Src inhibitors based on Monte Carlo method. *Med. Chem. Res*. **2015**, *24* (1), 283-290.

[78] Veselinović, A.M.; Veselinović, J.B.; Toropov, A.A.; Toropova, A.P.; Nikolić, G. M. QSAR models for the reactivation of Sarin inhibited AChE by quaternary pyridinium oximes based on Monte Carlo method. *Curr. Comput. Aided Drug Des*., **2015** , *10* (3), 266-273.

[79] García, J.; Duchowicz, P.R ; Rozas, M.F.; Caram, J.A.; Mirífico, M.V.; Fernández, F.M.; Castro, E.A. A comparative QSAR on 1,2,5-thiadiazolidin-3-one 1,1-dioxide compounds as selective inhibitors of human serine proteinases. *J. Mol. Graph. Model*., **2011**, *31*, 10-19.

[80] Živković, J.V.; Trutić, N.V.; Veselinović, J.B.; Nikolić, G.M.; Veselinović, A.M. Monte Carlo method based QSAR modeling of maleimide derivatives as glycogen synthase kinase-3β inhibitors. *Comput. Biol. Med.,* **2015**, *64*(1), 276–282.

[81] Yilmaz, H.; Novoselska, N.; Rasulev, B.F.; Toropov, A.A. ; Guzel, Y.; Kuz'min, V.; Leszczynska, D.; Leszczynski, J. Amino substituted nitrogen heterocycle ureas as kinase insert domain containing receptor inhibitors: Performance of structure - activity relationship approaches. *J. Food Drug Anal*. **2015**, *23*, 168 - 175.

[82] Comelli, N.C.; Ortiz, E.V.; Kolacz, M.; Toropova, A.P.; Toropov, A.A.; Duchowicz,P.R.; Castro, E.A. Conformation-Independent QSAR on c-Src Tyrosine Kinase Inhibitors. *Chemometr. Intell. Lab. Syst*., **2014**, *134*, 47–52.

[83] Veselinović, A.M.; Veselinović, J.B.; Živković, J.V.; Nikolić, G.M. Application of smiles notation based optimal descriptors in drug discovery and design. *Curr. Top. Med.Chem*., **2015**, *15* (18), 1768-1779.

[84] Ghaedi, A. Predicting the cytotoxicity of ionic liquids using QSAR model based on SMILES optimal descriptors. *J. Mol. Liq*., **2015**, *208,* 269–279.

[85] Deng, F.-F.; Xie, M.-H.; Li, P.-Z.; Tian, Y.-L.; Zhang, X.-Y.; Zhai, H.-L. Study on the antagonists for the orphan G protein-coupled receptor GPR55 by quantitative structure-activity relationship. *Chemometr. Intell. Lab. Syst*., **2014**, *131*, 51-60.

[86] Toropova, A .P.; Toropov, A. A.; Rasulev, B.F.; Benfenati, E.; Gini, G.; Leszczynska, D.; Leszczynksy, J. QSAR models for ACE-inhibitor activity of tri-peptides based on representation of the molecular structure by a graph of atomic orbitals and SMILES. *Struct. Chem*., **2012**, *23*, 1873-1878.

[87] Toropov, A.A.; Toropova, A .P.; Raska, Jr I.; Benfenati, E.; Gini, G. QSAR modeling of endpoints for peptides which is based on representation of the molecular structure by a sequence of amino acids. *Struct. Chem*., **2012**, 23, 1891-1904.

[88] Toropova, M.A.; Veselinović, A.M.; Veselinović, J.B.; Stojanović, D.B.; Toropov, A.A. QSAR modeling of the antimicrobial activity of peptides as a mathematical function of a sequence of amino acids. *Comp. Biol. Chem.*, **2015**, *59*, 126-130.

[89] Toropov, A.A.; Rasulev, B.F.; Leszczynska, D.; Leszczynski, J. New Approach to QSPR Modeling of Fullerene C60 Solubility in Organic Solvents: An Application of SMILES- Based Optimal Descriptors. Chapter 14, in Book (Edited by Cataldo, F.; Da Ros, T.): Medicinal Chemistry and Pharmacological Potential of Fullerenes and Carbon Nanotubes, **2008**, *1*, 337-350.

[90] Toropov, A.A.; Toropova, A.P.; Puzyn, T.; Benfenati, E.; Gini, G.; Leszczynska, D.; Leszczynski, J. QSAR as a random event: modeling of nanoparticles uptake in PaCa2 cancer cells. *Chemosphere*, **2013**, *92* (1), 31-37.

[91] Toropova, A.P.; Toropov, A.A.; Rallo, R.; Leszczynska, D.; Leszczynski, J. Optimal descriptor as a translator of eclectic data into prediction of cytotoxicity for metal oxide nanoparticles under different conditions. Ecotoxicol. Environ. Saf., **2015**, *112*, 39-45.

[92] Toropova, A.P.; Toropov, A.A.; Benfenati, E.; Korenstein, R.; Leszczynska, D.; Leszczynski, J. Optimal nano-descriptors as translators of eclectic data into prediction of the cell membrane damage by means of nano metal-oxides. *Environ. Sci. Pollut. Res. Int*., **2015**, *22* (1), 745-757.

[93] Toropov, A.A.; Toropova, A.P. Quasi-SMILES and nano-QFAR: United model for mutagenicity of fullerene and MWCNT under different conditions. *Chemosphere*, **2015**, *139*, 18-22.

[94] Toropov, A.A.; Toropova, A.P. Editorial. Special issue: From Chemoinformatics to Nanoinformatics: New tools for drug discovery and nanoparticles design in medicinal chemistry. *Curr. Top. Med.Chem*., **2015**, *15*(18), 1767.

[95] Toropova, A.P.; Toropov, A.A. Mutagenicity: QSAR -quasi-QSAR -nano-QSAR. *Mini. Rev. Med. Chem*., **2015**, *15* (8), 608-621.

[96] Toropova, A.P.; Toropov, A.A.; Veselinović, A.M.; Veselinović, J.B.; Benfenati, E.; Leszczynska, D.; Leszczynski, J. Nano-QSAR: Model of mutagenicity of fullerene as a mathematical function of different conditions. *Ecotoxicol. Environ. Saf*., **2016**, *124*, 32-36.

[97] Toropova,A.P.; Achary, P.G.R.; Toropov, A.A. Quasi-SMILES for Nano-QSAR prediction of toxic effect of Al2O3 nanoparticles. *J. Nanotoxicology & Nanomedicine*, **2016**, *1*(1), 17-28.

[98] Toropova, A.P.; Toropov, A.A.; Manganelli, S.; Leone,C.; Baderna, D.; Benfenati,E.; Fanelli, R. Quasi-SMILES as a tool to utilize eclectic data for predicting the behavior of nanomaterials. *Nanoimpact*, **2016**, *1*, 60–64.

[99] Toropova, A.P.; Toropov, A.A.; Rallo, R.; Leszczynska, D.; Leszczynski, J. Nano-QSAR: Genotoxicity of multi-walled carbon nanotubes. *Int. J. Environ. Res*., **2016**, *10* (1), 59-64.

[100] Obiol-Pardo, C.; Gomis-Tena, J.; Sanz, F.; Saiz, J.; Pastor, M. A multiscale simulation system for the prediction of drug-induced cardiotoxicity. *J. Chem. Inf. Model*. **2011**, *51* (2), 483–492.

[101] Mondal, C.; Halder, A.K.; Adhikari, N.; Saha, A.; Saha, K.D.; Gayen, S.; Jha, T. Comparative validated molecular modeling of p53-HDM2 inhibitors as antiproliferative agents. *Eur. J. Med. Chem.,* **2015**, *90*, 860-875.

[102] P.K. Ojha, K. Roy, Comparative QSARs for antimalarial endochins: Importance of descriptor-thinning and noise reduction prior to feature selection, *Chemometr. Intell. Lab. Syst.* **2011**, *109*, 146-161.

[103] Scotti, L.; Mendonça, F.J.B.; Moreira, D.R.M.; da Silva, M.S.; Pitta, I.R.; Scotti, M.T. SAR, QSAR and Docking of anticancer flavonoids and variants: a review. *Curr. Top. Med. Chem*., **2012** , *12*, 2785-2809.

[104] Kleandrova, V.V.; Luan, F.; González-Díaz, H.; Ruso, J.M.; Melo, A.; Speck-Planche, A.; Cordeiro, M.N.D.S. Computational ecotoxicology: Simultaneous prediction of ecotoxic effects of anoparticles under different experimental conditions. *Environ. Int*., **2014**, *73*, 288-294.

[105] Vrontaki, E.; Melagraki, G.; Mavromoustakos, T.; Afantitis, A. Exploiting ChEMBL database to identify indole analogs as HCV replication inhibitors. *Methods*, **2015**, *71*, 4-13.

[106] Carbó-Dorca, R. Least squares estimation of unknown molecular properties and quantum QSPR fundamental equation. *J. Math. Chem.,* **2015**, *53*,1651–1656.

[107] OECD, **2007**. (Organization for Economic Co-operation and Development). Guidance Document on The Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models No. 69.

[108] Toropova, A.P.; Toropov, A.A. Evolution of Optimal Descriptors: Solved, Unsolved, and Unsoluble Tasks. *Int. J. Quant. Struct.-Prop. Relat.*, **2016**, *1* (2), 52-71.

Table 1

Examples of the “super-attribute” that is suggested instead of ATOMPAIR

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **BOND** | | | **NOSP** | | | | **HALO** | | | |
| **SMILES** |  | **=** | **#** | **@** | **N** | **O** | **S** | **P** | **F** | **Cl** | **Br** | **I** |
| c1ccccc1/C(c2ccccn2)=N/Nc(n3)sc(c34)cccc4 | $ | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| COCCCC\C(c1ccc(cc1)C(F)(F)F)=N/OCCN | $ | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| C[C@H](CN(C)C)CN1c2ccccc2Sc2ccc(cc12)C#N | $ | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |

Table 2

Examples of molecular features used to build up a CORAL model

|  |  |  |
| --- | --- | --- |
| **Source** | **Molecular feature, *Xk*** | **Comment** |
| **SMILES** | 1........... | Presence of ring |
|  | Cl.......... | Presence of chlorine |
|  | N...#...C... | Presence of fragment “N≡C─“ |
|  | BOND01100000 | Presence of triple bonds and stereo chemical bonds together with absence of double bond |
|  | NOSP01000000 | Presence of oxygen and absence of nitrogen, Sulphur, and phosphorus |
|  | HALO01000000 | Presence of chlorine atom and absence of fluorine, bromine and iodine |
|  | $10111000000 | Presence of double bonds, stereo chemical bonds, nitrogen and oxygen, together with absence of triple bonds, Sulphur, phosphorus, fluorine, chlorine, bromine and iodine |
| **HSG** | C5...A..1... | Presence of one five members ring with aromaticity |
|  | C6...AH.2... | Presence of two six members ring with aromaticity and heteroatoms |
| **GAO** | EC0-1s1 3... | Presence of 1s1 vertex with vertex degree 3 |
|  | EC0-2s2 17.. | Presence of 2s2 vertex with vertex degree 17 |
|  | NNC-2p3 927. | Presence of vertex with the nearest neighbor code equal to 927\* |

\*) The nearest neighbor code for k-th vertex in GAO is calculated as the following



Table 3

The statistical characteristics of QSAR models for three endpoints examined in this work

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Endpoint** | **Split** | **Set** | **n** | **r2** | **RMSE** | **CR2\*\*** | **r2 / RMSE**  **in the literature** |
| 1 | 1 | Training | 82 | 0.8283 | 0.570 | 0.825 |  |
|  |  | Invisible training | 84 | 0.7921 | 0.576 | 0.781 |  |
|  |  | Calibration | 34 | 0.8378 | 0.425 | 0.824 |  |
|  |  | validation | 33(6\*) | 0.9318 | 0.453 |  | 0.842 / 0.564 [27] |
|  | 2 | Training | 80 | 0.7980 | 0.612 | 0.791 |  |
|  |  | Invisible training | 76 | 0.8490 | 0.501 | 0.842 |  |
|  |  | Calibration | 39 | 0.8536 | 0.521 | 0.831 |  |
|  |  | validation | 38(2) | 0.8535 | 0.617 |  | 0.690 / 0.702 [27] |
|  | 3 | Training | 78 | 0.8471 | 0.521 | 0.843 |  |
|  |  | Invisible training | 85 | 0.8423 | 0.672 | 0.835 |  |
|  |  | Calibration | 35 | 0.8939 | 0.678 | 0.882 |  |
|  |  | validation | 35(4) | 0.8058 | 0.617 |  | 0.689 / 0.816 [27] |
| 2 | 1 | Training | 146 | 0.9216 | 0.307 | 0.917 |  |
|  |  | Invisible training | 138 | 0.9079 | 0.386 | 0.905 |  |
|  |  | Calibration | 55 | 0.8225 | 0.435 | 0.817 |  |
|  |  | validation | 55(9) | 0.8875 | 0.388 |  | 0.8618 / 0.396 [26] |
|  | 2 | Training | 136 | 0.9031 | 0.327 | 0.900 |  |
|  |  | Invisible training | 146 | 0.9025 | 0.394 | 0.899 |  |
|  |  | Calibration | 56 | 0.8684 | 0.432 | 0.861 |  |
|  |  | validation | 56(4) | 0.9111 | 0.340 |  | 0.9035 / 0.343 [26] |
|  | 3 | Training | 144 | 0.9121 | 0.342 | 0.910 |  |
|  |  | Invisible training | 133 | 0.9362 | 0.354 | 0.932 |  |
|  |  | Calibration | 60 | 0.8481 | 0.474 | 0.833 |  |
|  |  | validation | 57(9) | 0.9326 | 0.303 |  | 0.9359 / 0.298 [26] |
| 3 | 1 | Training | 40 | 0.9000 | 0.275 | 0.891 |  |
|  |  | Invisible training | 40 | 0.8998 | 0.322 | 0.893 |  |
|  |  | Calibration | 38 | 0.7230 | 0.576 | 0.703 |  |
|  |  | validation | 37(6) | 0.7553 | 0.497 |  | 0.743 / 0.581 [101] |
|  | 2 | Training | 41 | 0.9220 | 0.296 | 0.910 |  |
|  |  | Invisible training | 39 | 0.9222 | 0.455 | 0.915 |  |
|  |  | Calibration | 38 | 0.7711 | 0.670 | 0.743 |  |
|  |  | validation | 37(4) | 0.7714 | 0.541 |  | 0.743 / 0.581 [101] |
|  | 3 | Training | 46 | 0.8926 | 0.269 | 0.869 |  |
|  |  | Invisible training | 45 | 0.8929 | 0.377 | 0.879 |  |
|  |  | Calibration | 32 | 0.8714 | 0.356 | 0.856 |  |
|  |  | validation | 32(4) | 0.9120 | 0.476 |  | 0.743 / 0.581 [101] |

\*) In brackets, the number of compounds which are not satisfactory according to inequality 8.

\*\*) Y-randomization test, i.e.; the *R* is the correlation coefficient between experimental (x) and predicted (y) values of an endpoint, theis the random correlation coefficient between x and y’, where y’ is vector y after N random permutations [102].

Table 4

Molecular features which are promoters of increase or decrease for endpoint 1

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| SAk | CWs in run 1 | CWs in run 1 | CWs in run 1 | NT\* | NiT | NC | SAdefect |
| O...(....... | 0.81520 | 2.87589 | 1.56638 | 82 | 84 | 34 | 0.0000 |
| =...O...(... | 0.99845 | 3.43804 | 2.30860 | 81 | 84 | 34 | 0.0001 |
| C...1....... | 3.25408 | 1.75387 | 2.50301 | 78 | 71 | 32 | 0.0001 |
| C...(....... | 0.43773 | 0.81491 | 0.24626 | 74 | 75 | 31 | 0.0001 |
| Cl.......... | 0.06270 | 1.56733 | 1.18755 | 51 | 49 | 18 | 0.0013 |
| O...(...(... | 2.12427 | 0.99705 | 3.62613 | 48 | 46 | 21 | 0.0005 |
| C........... | -0.87775 | -0.68950 | -0.50429 | 82 | 84 | 34 | 0.0000 |
| N........... | -0.87706 | -2.37439 | -1.24976 | 70 | 76 | 28 | 0.0003 |
| N...C....... | -0.49847 | -0.68803 | -0.12058 | 53 | 47 | 21 | 0.0004 |
| 1...c...(... | -0.12708 | -1.80886 | -1.62854 | 49 | 50 | 19 | 0.0006 |
| N...(...C... | -1.62135 | -1.25234 | -1.06077 | 41 | 38 | 13 | 0.0022 |

\*) NT, NiT, and NC are the number of SMILES, which contain SAk in the training, invisible training, and calibration sets, respectively.

Table 5

Molecular features which are promoters of increase or decrease for endpoint 2

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| SAk | CWs in run 1 | CWs in run 1 | CWs in run 1 | NT\* | NiT | NC | SAdefect |
| c...c...c... | 3.25483 | 1.37947 | 2.87991 | 129 | 125 | 54 | 0.0005 |
| c...1...c... | 2.12748 | 2.12990 | 1.75197 | 119 | 120 | 48 | 0.0003 |
| N...(...C... | 4.00340 | 0.99691 | 2.49549 | 87 | 76 | 36 | 0.0005 |
| c...c...2... | 1.37921 | 3.24870 | 1.56535 | 82 | 88 | 25 | 0.0010 |
| C...N...(... | 0.81208 | 1.75272 | 1.93310 | 81 | 73 | 33 | 0.0004 |
| O...C....... | 0.62349 | 1.37452 | 1.19245 | 75 | 61 | 31 | 0.0005 |
| c...(...C... | 0.24641 | 1.56446 | 0.06168 | 57 | 61 | 17 | 0.0011 |
| 2...c...(... | 2.50471 | 1.37535 | 1.75368 | 51 | 56 | 13 | 0.0018 |
| 1........... | -0.12521 | -2.00292 | -0.50003 | 143 | 136 | 55 | 0.0001 |
| C...(....... | -0.49517 | -0.12900 | -0.49989 | 139 | 126 | 54 | 0.0002 |
| c...c....... | -0.12757 | -0.87051 | -2.94158 | 139 | 132 | 54 | 0.0002 |
| c...(....... | -0.50490 | -0.49843 | -0.69067 | 129 | 128 | 51 | 0.0002 |
| O........... | -1.06648 | -1.62539 | -1.62659 | 122 | 120 | 45 | 0.0001 |
| O...(....... | -0.49805 | -0.68386 | -0.12220 | 100 | 96 | 34 | 0.0005 |
| =........... | -1.80909 | -0.87334 | -1.62973 | 94 | 93 | 29 | 0.0009 |
| [........... | -0.68736 | -0.87037 | -1.06255 | 92 | 76 | 43 | 0.0011 |

\*) NT, NiT, and NC are the number of SMILES, which contain SAk in the training, invisible training, and calibration sets, respectively.

Table 6

Molecular features which are promoters of increase or decrease for endpoint 3

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| SAk | CWs in run 1 | CWs in run 1 | CWs in run 1 | NT\* | NiT | NC | SAdefect |
| 1........... | 5.49596 | 10.00498 | 11.87834 | 40 | 40 | 38 | 0.0000 |
| 2........... | 8.50229 | 6.99835 | 9.43689 | 40 | 40 | 38 | 0.0000 |
| O...(....... | 0.06256 | 0.62707 | 1.74608 | 39 | 40 | 37 | 0.0000 |
| EC0-1s2 6... | 2.87812 | 2.12133 | 2.12593 | 38 | 39 | 37 | 0.0003 |
| EC0-2p4 3... | 0.25353 | 0.25417 | 2.50062 | 38 | 40 | 36 | 0.0000 |
| 4........... | 4.00405 | 2.49602 | 3.05800 | 35 | 36 | 34 | 0.0003 |
| c...1....... | 6.81297 | 5.50399 | 5.87561 | 34 | 37 | 34 | 0.0007 |
| (........... | -1.25059 | -1.44119 | -1.81384 | 40 | 40 | 38 | 0.0000 |
| C...(....... | -0.49545 | -0.68526 | -0.12092 | 40 | 40 | 38 | 0.0000 |
| EC0-1s2 9... | -0.68738 | -0.12180 | -0.50083 | 40 | 40 | 38 | 0.0000 |
| N........... | -1.06519 | -1.06319 | -0.49859 | 40 | 37 | 37 | 0.0003 |
| NNC-2s2 725. | -0.49938 | -0.50430 | -0.50245 | 40 | 40 | 37 | 0.0003 |
| NNC-1s2 918. | -2.37336 | -3.31085 | -4.62002 | 37 | 40 | 36 | 0.0003 |
| NNC-2p2 918. | -2.18985 | -2.94182 | -3.12990 | 37 | 40 | 36 | 0.0003 |
| EC0-1s2 8... | -0.69205 | -0.50372 | -0.87240 | 34 | 35 | 27 | 0.0023 |

\*) NT, NiT, and NC are the number of SMILES, which contain SAk in the training, invisible training, and calibration sets, respectively.