Predicting Nanomaterials' Endpoints Using Quasi-SMILES: A Computational Approach

Andrey A. Toropov
IRCCS-Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, Milano, Italy

ABSTRACT

Quantitative structure – property / activity relationships (QSPRs/QSARs) represent efficient and in most cases suitable and accurate computational tools to estimate endpoints of substances with geometric characteristics described adequately by both similarity and variability of molecular structure. Unfortunately, in many cases the QSPR/QSAR analysis is not possible for various nanomaterials. A successful technique to build up a predictive model for an endpoint related to nanomaterials involves holistic elucidation of the endpoint as a mathematical function of all available eclectic data, such as physicochemical and biochemical conditions and circumstances. This chapter offers an introduction to the subject and provides examples of models based on eclectic data represented by so-called quasi-SMILES, analogs of the traditional SMILES utilized in the "classic" QSPR/QSAR analyses. In contrast to traditional SMILES, quasi-SMILES are representation of all available eclectic data (not only information about the molecular structure).

Keywords: nanomaterial; Monte Carlo method; quasi-SMILES; quasi-QSPR/QSAR; nano-QSPR/QSAR; CORAL software

1. Introduction

The influence of various nanomaterials on the everyday life gradually increase owing to their potential be useful for different applications in medicine (De Jong and Borm, 2008; Webster et al., 2013; Toropova et al., 2016).

As a rule, generally, an experimental measurement of an endpoint is not cheap. In addition, performing of the experiment demands considerable time. This promotes developments of alternative techniques, able to provide investigated data faster and more efficient. Such techniques are available in a large pool of computational chemistry methods. Specifically, the techniques of calculations of endpoints, that are able to include experimental data related to untested but similar substances, become attractive alternative for the experiment. Quantitative structure – property / activity relationships (QSPRs/QSARs) represent the practical application of the above-mentioned alternative. The theory and praxis of the QSPR/QSAR have impressive record of successful utilizations for prediction of endpoints related to organic (Toropova et al., 2011a), inorganic (Toropova et al., 2011b), organometallic (Toropova et al., 2011c), and polymeric (Duchowicz et al., 2015) species.

The evolution of the QSPR/QSAR theory/praxis involves a few components. It benefits on improvements of algorithms of the analysis of available data that allow predicting physicochemical and/or biochemical behavior of substances which were not examined in the experiment. The second, less discussed component is of an equal importance. It involves establishment of the definition of the task (target). Historically, the development of correlations "descriptor - endpoint" for a sole endpoint was the main aim of QSPR/QSAR modeling in the beginning of applications of this approach (Wiener, 1947a,b;

1948; Gutman et al., 2005, 2009; Hosoya, 1972; Bonchev et al., 1980). Later the QSPR/QSAR analysis aimed to more challenged tasks – prediction of not a single property but a group of important and sometimes interdependent endpoints (Speck-Planche et al., 2011, 2012a,b, 2013).

The classes of substances considered for the QSPR/QSAR analysis has broaden over the years. An important impetus for development of novel approaches arrived after applications of classical QSAR methodology to nanomaterials – unique class of chemical species - failed. The growing importance of these species is illustrates by fast growing number of publications dedicated to nanomaterials. In 2000 there was about hundred papers related to keyword “nanomaterial”. This number expands to eleven thousands in 2015 (Figure 1).

Obviously, there have been numerous attempts to utilize the QSPR/QSAR approach for nanomaterials with the application of various "nano-descriptors" (Oksel et al., 2015). However, approaches focused on building up “nano-QSAR” were based on hardly accessible physicochemical characteristics of nanomaterials (Sayes and Ivanov, 2010; Glotzer and Solomon, 2007). Interestingly, also the traditional descriptors appropriate for substances, which are not nanomaterials were also examined as a tool to build up "nano-QSAR" (Fourches et al., 2011). However, in this work, all nanoparticles have the same “nano” metal core, and the difference between nanoparticles is defined solely by small organic molecules ((Fourches et al., 2011). Naturally, for such species traditional descriptors can be quite appropriate ones.

So-called optimal descriptors provide possibility to build up predictive model for various nanomaterials using ***ALL*** available eclectic data represented by quasi-SMILES. This holistic approach was successfully applied for the development of model for membrane damage by ZnO and TiO_2 nanoparticles (Toropova and Toropov, 2013; Toropova et al., 2014); mutagenicity of fullerene (Toropov and Toropova, 2014); mutagenic potential of multi-walled carbon-nanotubes (Toropov and Toropova, 2015); and cytotoxicity of metal oxide nanoparticles to bacteria *Escherichia coli* (Toropova et al., 2012).

We believe that the increasing interest in application of efficient methods that reliably predict characteristics of nanomaterials justify review of such approaches and the results obtained using those techniques. The aims of this chapter are: (i) description of the method of building up quasi-SMILES; and (ii) introduction of principles of development of predictive models based on quasi-SMILES. The second aim could be efficiently accomplished using the CORAL software available on the Internet (CORAL, http://www.insilico.eu/coral). The readers are welcome to carry out their own research projects using the CORAL program.

It is to be noted, today the quasi-SMILES likely have no alternative in the case of a situation where one should construct a model based on eclectic data, such as physicochemical and biochemical conditions of a phenomena, presence of large number of factors which can impact the phenomena, together with uncertainty in correctness of classification of all factors into(i) factors with significant impact; and (ii) factors with neglectable influence. Unfortunately, the above-mentioned indeterminacy often takes place in various stages of the drug discovery.

2. Method

2.1. SMILES and quasi-SMILES

Simplified molecular input-line entry system (SMILES) has been introduced by Weiniger and collaborators (Weininger, 1988, 1990; Weininger et al., 1989). This approach allows for simple representation of the molecular structures.

There are defined equivalences between the representation of the molecular structure by graphs and using SMILES approach. However, one needs to be also aware about their significant distinctions. Those are reviewed in the recent publication (Toropov et al., 2011).

Optimal descriptors have been improved along with advances of QSAR approaches. During the initial steps of evolution of the optimal descriptors the molecular graph was the basis for building up a QSAR model. Very similar (if not identical) approach has been developed for SMILES and SMILES attributes. It can be summarized as follows:

(i) each SMILES of the training set provide a list of attributes, x_{kj} (Toropov et al., 2011):

$$SMILES_k \rightarrow \{x_{k1}, x_{k2}, ..., x_{km}\} \quad (1)$$

(ii) The Monte Carlo method provides correlation weights for total list of attributes. They are extracted from all SMILES notations of the training set which give maximal correlation coefficient between examined endpoint and sums of correlation weights for SMILES of the training set:

$$Monte\_Carlo\_method \rightarrow \{CW(x_{k1}), CW(x_{k2}), ..., CW(x_{km})\} \quad (2)$$

(iii) The predictive model is represented by one-variable linear equation:

$$Least\_squares\_method \rightarrow$$
$$EP_k = C_0 + C_1 \times \sum_{x_{kj} \in SMILES} CW(x_{kj}) = C_0 + C_1 \times DCW(T^*, N^*) \quad (3)$$

In the vector and matrix representation this approach can be expressed as the following:

$$\begin{pmatrix} MS_1 \\ MS_2 \\ \dots \\ MS_n \end{pmatrix} \rightarrow \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix} \leftrightarrow \begin{pmatrix} E_1 \\ E_2 \\ \dots \\ E_n \end{pmatrix} \quad (4)$$

where MS_k are molecular structures (represented by graph or SMILES); x_{kj} represent molecular features extracted from molecular graph or molecular features extracted from SMILES. However, an endpoint can be interdependent with some additional impacts related to physicochemical and/or biochemical conditions. In this case, instead of traditional SMILES, one should utilize an extension of the classical parameters referred to as quasi-SMILES. The basis of building up quasi-SMILES can be extracted from a graph, SMILES, and additional eclectic data.

In traditional approach one assumes that an endpoint is interdependent from the molecular structure. However, there are cases in which this approach has to be revised. Obviously, there are also situations where one can expect that the endpoint is interdependent with other conditions (temperature, concentration, dose, etc.) and/or circumstances (the presence/absence of illumination, magnetic field, different times of exposure, etc.). In this case, instead of the paradigm:

"Endpoint = F (Molecular Structure)"

one should apply other paradigm:

"Endpoint = F (Eclectic Data)".

The quasi-SMILES is a representation of the eclectic data. The above-mentioned scheme (1)-(2)-(3) is represented by the quasi-SMILES (Eclectic Data), ED_k, correlation weights of

symbols from quasi-SMILES, $CW(x_{kj})$, and experimental data obtained for the studied endpoint, E_k:

$$\begin{pmatrix} ED_1 \\ ED_2 \\ \dots \\ ED_n \end{pmatrix} \rightarrow \begin{bmatrix} CW(x_{11}) & CW(x_{12}) & \dots & CW(x_{1m}) \\ CW(x_{21}) & CW(x_{22}) & \dots & CW(x_{2m}) \\ \dots & \dots & \dots & \dots \\ CW(x_{n1}) & CW(x_{n2}) & \dots & CW(x_{nm}) \end{bmatrix} \leftrightarrow \begin{pmatrix} E_1 \\ E_2 \\ \dots \\ E_n \end{pmatrix} \quad (5)$$

Thus, the vector of eclectic data represents quasi-SMILES. The quasi-SMILES is a string of symbols similar to traditional SMILES, but the meaning of each symbol in quasi-SMILES is not necessary the representation of molecular features. Figure 2 shows the general scheme of utilization of quasi-SMILES.

2.2. Monte Carlo method

In the case represented by Eq. 5, the Monte Carlo method is used to optimize the correlation weights $CW(x_{kj})$. Here the target function represents the correlation coefficient between the endpoints values E_k and sum of correlation weights of symbols from corresponding quasi-SMILES extracted from the training set.

The sequence of modification of correlation weights is random for each epoch of the Monte Carlo optimization. The epoch represents step-by-step modification of all correlation weights involved into building up a QSAR model. It is to be noted that one should define threshold in order to classify symbols of quasi-SMILES into two classes: 1) rare; and 2) not rare. The correlation weights of rare symbols are fixed to zero, and consequently, they are not involved in building up a model.

In the case of unlimited number of epochs of the Monte Carlo optimization the probability of the overtraining is very high. Under such circumstances, better approach is instead of unlimited number to use the number of epochs which gives preferable statistics for a calibration set. In principle, the measures of the statistical quality of the calibration set can be: 1) correlation coefficient between experimental and calculated values of an endpoint; and 2) root-mean squared error (RMSE). The graphical illustration (Figure 3) shows that these two approaches can give different values of the preferable number of epochs. The computational experiments indicate that the correlation coefficient provides more reliable criterion, because this criterion often gives preferable predictive potential for an external invisible validation set. However, in addition to the preferable number of epochs, one should also select preferable threshold (T*). Thus, the goal is the selection of satisfactory pair of values: T=T* and N=N*, which gives preferable statistical quality for the calibration set (Figure 4).

2.3. Utilization of the model

The result of the Monte Carlo optimization provides the list of correlation weights for symbols involved in the model. Each symbol is representation of defined circumstance. For instance, (i) temperature range, i.e. 100-110'C can be denoted as a code 'a'; and 110-120'C denoted as 'b', etc. (ii) dose ranges, i.e. 20-25 mg/kg denoted as 'c', 25-30 mg/kg denoted as 'd', etc. (iii) time of exposure 1 hour denoted as 'e'; 2 hours denoted as 'f', and so on, according to corresponding conditions and circumstances.

Having the data on the correlation weights, one can extract list of symbols from corresponding quasi-SMILES and calculate:

1) $$DCW(T^*,N^*) = \sum_{x_{kj} \in quasi-SMILES} CW(x_{kj})$$ (6)

2) $$EP_k = C_0 + C_1 \times DCW(T^*,N^*)$$ (7)

2.4. Domain of applicability

The experimental data is used for model development and for evaluation of the model quality. The split of data into the "visible" training set (for the described approach the "visible" training set contains also the calibration set) and "invisible" validation set has apparent influence upon the predictability of a model. A possible measure of the quality of the split can be estimated from prevalence of each feature in the training and calibration sets:

$$defect(x_{kj}) = \sum_{active} \left| P(x_{kj}) - P'(x_{kj}) \right|$$ (8)

where, the probability of feature x_{kj} in the training set $P(x_{kj})$ and the probability of x_{kj} in the calibration set $P(x_{kj})$ are calculated by:

$$P(x_{kj}) = \frac{N_{set}(x_{kj})}{N_{set}}$$ (9)

where $N_{set}(x_{kj})$ is the number of quasi-SMILES which contains x_{kj} and N_{set} represents the total number of quasi-SMILES in the set. The quality of split is evaluated based on the value of a defect. The defect is calculated with active (not blocked) x_{kj} only. If the defect = 0, the split should be considered as an "ideal" one. But in fact, this situation is not possible. However, the value of the defect calculated with Eq. 8 gives possibility to compare quality of various splits.

Sum of *defects* (x_{kj}) of all active attributes of quasi-SMILES can be a measure of a defect of each quasi-SMILES:

$$defect(quasi\_SMILES_k) = \sum_{x_{kj} \in quasi\_SMILES_k} defect(x_{kj})$$ (10)

Summation of all *defects (quasi_SMILES)* can be considered as a measure of quality of split of data into the visible training, calibration, and invisible validation sets:

$$defect(split) = \sum_{quasi\_SMILES_k \in Training} defect(quasi\_SMILES_k)$$ (11)

The probabilistic domain of applicability can be defined via inequality:

$$defect(quasi\_SMILES) < 2 \times \overline{defect(quasi\_SMILES)}$$ (12)

In other words, if quasi-SMILES characterized by the *defect (quasi-SMILES)* which is lower than the doubled average value of this characteristics over compounds included in the training set, then this quasi-SMILES falls into the domain of applicability. Otherwise, this quasi-SMILES is outside of the domain of applicability. In addition, one can compare two splits using the defect (split) calculated with Eq. 11. Split characterized by lower defect is better.

2.5. Mechanistic interpretation

The described approach allows defining the mechanical interpretation of model based on the correlation weights of active features extracted from quasi-SMILES. 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:
1) Features which have positive values of the correlation weight in all runs. These are promoters of endpoint increase;
2) Features which have negative values of the correlation weight in all runs. These are promoters of endpoint decrease;
3) Features which have both negative and positive values 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).

3. Examples of applications of quasi-SMILES for nanomaterials

The principles of the model development and selection of descriptors discussed in the previous sections have been tested on various cases. Examples of such studies are provided in the next few sections.

3.1. Format of representation of a model

The format of representation of a predictive QSAR model represents an extremely important feature for a potential user of the model. There are well-known OECD principles widely used in the QSPR/QSAR analyses. However, in the case of the model based on the quasi-SMILES, the scheme of building up of quasi-SMILES involves additional information. Thus, the format of representation of a model used in this work is the following:

- The description of endpoint;
- The description of quasi-SMILES;
- The statistical characteristics of model;
- Domain of applicability;
- Mechanistic interpretation.

The general scheme of the algorithm of building up a model is described in section "Method".

3.1. Cytotoxicity for metal oxide nanoparticles under different conditions

3.1.1. The description of endpoint

The numerical data on cytotoxicity of metal oxide nanoparticles to bacteria *E. coli* (the concentration of the nanoparticles that proved to be fatal to 50% of the bacteria *E. coli* LC50, in mol/L) have been taken from the literature (Pathakoti et al., 2014). The negative decimal logarithm of the LC50 (pLC50) has been considered as the endpoint. The dark cytotoxicity and photo-induced cytotoxicity were examined as united endpoint, owing to application of the model which is a mathematical function of atomic composition and conditions (the presence/absence of photo-inducing).

3.1.2. The description of quasi-SMILES

In the case of cytotoxicity in darkness, traditional SMILES was used to represent metal oxide nanoparticles. In the case of photo-induced cytotoxicity, the symbol '^' was

added at the end of traditional SMILES. Thus, absence of '^' means the acting of nanoparticle in darkness, presence of '^' means the acting of nanoparticle under illumination (Table 1) (Toropova et al., 2015).

3.1.3. The statistical characteristics of model

The best model for cytoctoxicity of metal oxide nanoparticles based on quasi-SMILES (Toropova et al., 2015) is the following:

$$pLC50 = 1.5185 (\pm 0.0334) + 0.8370 (\pm 0.0110) \times DCW(1,9) \quad (13)$$

n=22, r^2=0.9081, s=0.354, F=198 (training set)
n=6, r^2=0.9943, s=0.454 (calibration set)
n=6, r^2=0.9835, s=0.418 (validation set)

Table 2 contains the correlation weights for calculations with Eq. 13. An example of the calculation with Eq. 13 for quasi-SMILES is provided in the Table 3..

3.1.4. Domain of applicability

A value that characterizes half (50%) of any measured property is widely prevalent measure for rationalization of the research work. Examples include the definition of bit (elementary quantity of information), lethal dose for half of organisms (LD50), the square of correlation coefficient that should be larger than 0.5 (i.e again 50%), and so on. Inequality 12, gives possibility to define SMILES which fall into domain of applicability of prevalence of different molecular features (extracted from SMILES). In addition, the percentage of SMILES, which fall into domain of applicability is a measure of quality for split into the training and validation sets. One assumes that the split is satisfactory if more than 50% of compounds are in the domain of applicability.

The percentages of the domain of applicability, according to inequality 12 are 76%, 76%, 76%, 71%, 71%, and 71% , for splits 1, 2, 3, 4, 5, and 6, respectively. As it was noted above, one can define 50% as a threshold to confirm acceptability of a split. Thus, a split that is characterized by domain of applicability of more than 50% can be considered as satisfactory: all six examined splits are satisfactory ones.

3.1.5. Mechanistic interpretation

The obtained QSAR model allows evaluating a role of various eclectic features on the studied endpoint. Based on developed model one concludes that the double bonds (‘=’) are stable promoter for decrease of cytotoxicity. The illumination represents a promoter of increase of the cytotoxicity for considered metal nano oxides (Table 1).

3.2. Membrane damage by means of TiO_2 nanoparticles under different conditions

3.2.1. The description of endpoint

Recent experimental study on membrane damage by metal oxide nanoparticles provides interesting results that were used to develop another QSAR model. Experimental data on the physicochemical features of TiO_2 nanoparticles and their influence on the membrane damage are taken from the literature (Sayes and Ivanov, 2010). These are 1) engineered size (nm); 2) size in water suspension (nm); 3) size in phosphate buffered saline (BPS, nm); 4) concentration (mg/L); and 5) zeta potential (mV). Table 4 contains these parameters. The above-mentioned physicochemical features of TiO_2 nanoparticles were involved in building up quasi-SMILES and QSAR models for membrane damage values related to various TiO_2 nanoparticles (characterized by different physicochemical features) (Toropova and Toropov, 2013). The physicochemical data were normalized using following equation:

$$Norm(X_k) = \frac{\min X_k + X_k}{\min X_k + \max X_k} \quad (14)$$

Table 5 contains normalized data used to build up the quasi-SMILES. Table 6 contains quasi-SMILES defined according to scale represented in Figure 5. Three various splits of experimental data into the training and test sets were examined (Toropova and Toropov, 2013). These splits obey the following principles: 1) they are random; and 2) the ranges of the endpoint for the training and test sets are similar.

3.2.2. The description of quasi-SMILES

Experimental data were used to develop quasi-SMILES for the investigated phenomena. Table 7 contains the correlation weights of various contributions used in the predictive model.

3.2.3. The statistical characteristics of model

The best predictive model for membrane damage suggested in the recent work (Toropova and Toropov, 2013) is the following:

MD = 0.8054 (± 0.0044) + 0.1273 (± 0.0014)×DCW(2,20) (15)
n=10, r^2=0.9893, q^2=0.9845, s=0.025, F=741 (training set)
n=5, r^2=0.9647, s=0.066, (calibration set)
n=9, r^2=0.8679, s=0.115 (validation set)

3.2.3. Domain of applicability

The quality of the developed model was tested by investigation of the domain of applicability. All quasi-SMILES of the validation set fall into the domain of applicability according to inequality 12.

3.2.4. Mechanistic interpretation

Based on the correlation weights obtained in three runs of the Monte Carlo optimization one can conclude that A4 and A9 are promoters of increase of membrane damage caused by TiO_2 nanoparticles. On the other hand, B2 is the promoter of decrease for the endpoint. These findings help to shed some light on investigated phenomena. Again, the developed model allows to shed a light on the nature of the studied phenomena.

3.3. Mutagenicity of fullerene under different conditions

3.3.1. The description of endpoints

Another study targeted prediction of mutagenicity of the most classical nanoparticle – fullerene. The experimental study provided two endpoints. Both were examined in the recent computational work (Toropova et al., 2016):

1. The bacterial reverse mutation test conducted using *Salmonella typhimurium* strains TA100 (in the presence and absence of metabolic activation under dark conditions and irradiation were taken from the work (Shinohara et al., 2009)), and
2. The bacterial reverse mutation test conducted using *Escherichia coli* strain WP2 uvrA/pKM101 (in the presence and absence of metabolic activation under dark condition and irradiation were taken from the literature (Shinohara et al., 2009)).

The experimental data allows considering a number of features that could be used to develop QSAR model. Twenty quasi-SMILES were defined for the above data. These twenty quasi-SMILES were further randomly distributed into the training, calibration, and validation sets.

3.3.2. The description of quasi-SMILES

The details of the computational work are shown in the Tables 9-11. Table 9 contains the scheme of building up quasi-SMILES. This provide basis for the next steps of the study. Quasi-SMILES and experimental data on mutagenicity TA100 of fullerene under different conditions presented in Table 9 are displayed in the Table 10. Table 11 contains the correlation weights used as the basis of the models for three different splits into the training, calibration, and validation sets. The quasi-SMILES and experimental data on mutagenicity WP2uvrA/pKM101 of fullerene under different conditions (Table 9) are displayed in Table 12. They were used for a model development and Table 13 contains the correlation weights for the models considered in this study.

3.3.3. The statistical characteristics of model

The utilization of the optimal descriptors calculated according to scheme suggested in the literature (Toropova et al., 2016) resulted in the following best models for the above-mentioned two endpoints:

$$TA100 = 117.813 + 12.3159 \times DCW(2,3) \quad (16)$$

n= 10, r^2=0.6810, s=9.78, F=17 (sub-training set)
n=5, r^2=0.9396, s=7.91 (Calibration set)
n=5, r^2=0.7884, s=7.79 (Validation set)

$$WP2uvrA/pKM101 = 84.9481 + 16.1111 \times DCW(3,6) \quad (17)$$

n=10, r^2=0.6805, s=12.1, F=17 (sub-training set)
n=5, r^2=0.7480, s=16.5 (calibration set)
n=5, r^2=0.8367, s=25.7 (validation set)

3.3.4. Domain of applicability

The domains of applicability for quasi-SMILES involved in building up models are presented in Table 10 (TA100) and Table 12 (WP2uvrA/pKM101).

3.3.5. Mechanistic interpretation

Interestingly, almost all correlation weights are positive for the mutagenicity models of fullerene TA100 and WP2uvrA/pKM101. However, their values are different. One can extract features of quasi-SMILES with relative large values. These features represent leading contributions to the investigated phenomena. The two largest contributions include darkness (0) and absence of Mix S9 (-). One needs to note that the obtained results are based on small pool of experimental data. Apparently, it is possible that this interpretation can be adjusted after similar analysis is performed on larger experimental data for the studied endpoints.

4. Conclusions

The chapter reviews a concept of development of quasi-SMILES application that utilize all existing experimental data available for the studied species. This is a major

difference between traditional SMILES and quasi-SMILES approaches. The proposed concept has been used to predict outcomes of various processes involving nanomaterials. We do believe that the suggested hypothesis of building up predictive models is universal and can be relatively simply utilized to solve various non-standard tasks. This allows extending applications of QSAR/QSPR techniques to the cases not cover by the traditional methods.

Unique abilities of nanomaterials are well-known. The probability of these substances be effective pharmaceutical agents is high. However, traditional QSPR/QSAR analyses of these abilities (or these endpoints) often are not convenient for practice, whereas, described quasi-SMILES give possibility to solve tasks unsoluble by traditional paradigm of the QSPR/QSAR.

ACKNOWLEDGMENTS

This work was financially supported by National Science Foundation: NSF-CREST Grant #HRD-1547754 and EPSCoR Grant #362492-190200-01\NSFEPS-0903787. A.A.T. and A.P.T. thank the EC project PeptiCAPS (Project reference: 686141).

References

Bonchev, D., Balaban, A. T. and Mekenyan, O. (1980). Generalization of the graph center concept, and derived topological centric indexes. *Journal of Chemical Information and Computer Sciences* 20, 106-113.

CORAL software, http://www.insilico.eu/coral, accessed April 15, 2016.

De Jong, W.H., and Borm P.J.A. (2008). Drug delivery and nanoparticles: Applications and hazards. *International Journal of Nanomedicine* 3(2), 133–149.

Duchowicz, P.R., Fioressi, S.E., Bacelo, D.E., Saavedra, L.M., Toropova, A.P. and Toropov, A.A. (2015). QSPR studies on refractive indices of structurally heterogeneous polymers. *Chemometrics and Intelligent Laboratory Systems* 140, 86–91.

Fourches, D., Pu, D. and Tropsha, A. (2011). Exploring quantitativenanostructure–activity relationships (QNAR) modeling as a tool for pre-dicting biological effects of manufactured nanoparticles. *Combinatorial Chemistry & High Throughput Screening* 14, 217–225.

Glotzer, S. C. and Solomon, M. J. (2007). Anisotropy of building blocks and their assem-bly into complex structures. *Nature Materials* 6, 557–562.

Gutman, I., Toropov, A.A. and Toropova, A.P. (2005). The graph of atomic orbitals and it's basic properties. 1. Wiener index. *MATCH Communications in Mathematical and in Computer Chemistry* 53, 215-224.

Gutman, I., Furtula, B. and Petrović, M. (2009). Terminal Wiener index. *Journal of Mathematical Chemistry* 46 (2), 522-531.

Hosoya, H. (1972). Topological index as a sorting device for coding chemical structures. *Journal of Chemical Documentation* 12, 181-183.

Oksel,C., Ma, C.Y., Liu, J.J., Wilkins, T. and Wang, X.Z. (2015). (Q)SAR modelling of nanomaterial toxicity: A critical review. *Particuology* 21, 1-19.

Pathakoti, K, Huang, M.-J., Watts, J.D., He, X., Huey-Min Hwang, H.-M. (2014). Using experimental data of Escherichia coli to develop a QSAR model for predicting the photo-induced cytotoxicity of metal oxide nanoparticles. *Journal of Photochemistry and Photobiology A: Chemistry* 130, 234–240.

Sayes, C. and Ivanov, I. (2010). Comparative study of predictive computational modelsfor nanoparticle induced cytotoxicity. *Risk Analysis* 30, 1723–1734.

Shinohara, N., Matsumoto, K., Endoh, S., Maru, J. and Nakanishi, J. (2009). In vitro and in vivo genotoxicity tests on fullerene C60 nanoparticles. *Toxicology Letters* 191, 289-296.

Speck-Planche,A., Kleandrova,V.V., Luan, F. and Cordeiro, M.N.D.S. (2011). Multi-target drug discovery in anti-cancer therapy: Fragment-based approach toward the design of potent and versatile anti-prostate cancer agents. *Bioorganic & Medicinal Chemistry* 19 (21), 6239-6244.

Speck-Planche, A., Kleandrova,V.V., Luan, F. and Cordeiro, M.N.D.S. (2012a). Chemoinformatics in anti-cancer chemotherapy: Multi-target QSAR model for the in silico discovery of anti-breast cancer agents. *European Journal of Pharmaceutical Sciences* 47 (1), 273-279.

Speck-Planche,A., Kleandrova,V.V., Luan, F. and Cordeiro, M.N.D.S. (2012b). Predicting multiple ecotoxicological profiles in agrochemical fungicides: A multi-species chemoinformatic approach. *Ecotoxicology and Environmental Safety* 80, 308-313,

Speck-Planche,A., Kleandrova,V.V. and Cordeiro, M.N.D.S. (2013). New insights toward the discovery of antibacterial agents: Multi-tasking QSBER model for the simultaneous prediction of anti-tuberculosis activity and toxicological profiles of drugs. *European Journal of Pharmaceutical Sciences* 48 (4–5), 812-818.

Toropov, A.A., Toropova, A.P., Martyanov, S.E., Benfenati, E., Gini, G., Leszczynska, D. and Leszczynski, J. (2011). Comparison of SMILES and molecular graphs as the representation of the molecular structure for QSAR analysis for mutagenic potential of polyaromatic amines. *Chemometrics and Intelligent Laboratory Systems* 109, 94-100.

Toropov, A.A., Toropova, A.P. (2014). Optimal descriptor as a translator of eclectic data into endpoint prediction: Mutagenicity of fullerene as a mathematical function of conditions. *Chemosphere* 104, 262–264.

Toropov, A.A., Toropova, A.P. (2015). Quasi-QSAR for mutagenic potential of multi-walled carbon-nanotubes. *Chemosphere* 124, 40–46.

Toropova, A.P., Toropov, A.A., Benfenati, E., Gini, G., Leszczynska, D. and Leszczynski, J. (2011a). CORAL: Quantitative Structure–Activity Relationship models for estimating toxicity of organic compounds in rats. *Journal of Computational Chemistry* 32, 2727-2733.

Toropova, A.P., Toropova, A.A., Benfenati, E. and Gini, G. (2011b). QSAR modelling toxicity toward rats of inorganic substances by means of CORAL. *Central European Journal of Chemistry* 9(1), 75-85.

Toropova, A.P., Toropov, A.A., Benfenati, E. and Gini, G. (2011c). Co-evolutions of correlations for QSAR of toxicity of organometallic and inorganic substances: an unexpected good prediction based on a model that seems untrustworthy. *Chemometrics and Intelligent Laboratory Systems* 105, 215-219.

Toropova, A.P., Toropov, A.A., Rallo, R., Leszczynska, D. and Leszczynski, J. (2012). Novel application of the CORAL software to model cytotoxicity of metal oxide nanoparticles to bacteria *Escherichia coli*. *Chemosphere* 89, 1098–1102.

Toropova, A.P. and Toropov, A.A. (2013). Optimal descriptor as a translator of eclectic information into the prediction of membrane damage by means of various TiO2 nanoparticles. *Chemosphere* 93, 2650–2655.

Toropova, A.P., Toropov, A.A., Benfenati, E., Puzyn, T., Leszczynska, D. and Leszczynski, J. (2014). Optimal descriptor as a translator of eclectic information into the prediction of membrane damage: The case of a group of ZnO and TiO2 nanoparticles. *Ecotoxicology and Environmental Safety* 108, 203–209.

Toropova, A.P., Toropov, A.A., Rallo, R., Leszczynska,D. and Leszczynski, J. (2015). Optimal descriptor as a translator of eclectic data into prediction of cytotoxicity for metal oxide nanoparticles under different conditions. *Ecotoxicology and Environmental Safety* 112, 39-45.

Toropova, A.P., Toropov, A.A., Veselinović, A.M., Veselinović, J.B., Benfenati, E., Leszczynska, D. and Leszczynski, J. (2016). Nano-QSAR:Model of mutagenicity of fullerene as a mathematical function of different conditions. *Ecotoxicology and Environmental Safety* 124, 32–36.

Toropova, A.P., Achary, P.G.R. and Toropov, A.A. (2016). Quasi-SMILES for Nano-QSAR prediction of toxic effect of Al_2O_3 nanoparticles. *Journal of Nanotoxicology and Nanomedicine* 1(1), 17-28.

Webster, D.M., Sundaram, P. and Byrne, M.E. (2013). Injectable nanomaterials for drug delivery: Carriers, targeting moieties, and therapeutics. *European Journal of Pharmaceutics and Biopharmaceutics* 84, 1–20

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* 28(1), 31-36.

Weininger, D., Weininger, A. and Weininger, J.L. (1989). SMILES. 2. Algorithm for generation of unique SMILES notation. *Journal of Chemical Information and Computer Sciences* 29(2), 97-101.

Weininger, D. (1990). Smiles. 3. Depict. Graphical depiction of chemical structures. *Journal of Chemical Information and Computer Sciences*. 30 (3), 237-243.

Wiener, H. (1947a). Correlation of heats of isomerization, and differences in heats of vaporization of isomers, among the paraffin hydrocarbons. *Journal of the American Chemical Society* 69(11), 2636-2638.

Wiener, H. (1947b). Structural determination of paraffin boiling points. *Journal of the American Chemical Society* 69(1), 17-20.

Wiener, H. (1948). Relation of the physical properties of the isomeric alkanes to molecular structure. surface tension, specific dispersion, and critical solution temperature in aniline. *Journal of Physical Chemistry* 52(6), 1082-1089.

Table 1

Quasi-SMILES used to build up model for cytotoxicity of metal oxide nanoparticles.

Splits*						**Quasi-SMILES for metal oxide nanoparticles**	***pLC*$_{50}$** in mol/L [30]
1	**2**	**3**	**4**	**5**	**6**		
v	t	c	t	t	c	O=[Zn]	5.80
t	c	v	t	t	t	[Cu]=O	4.24
t	t	c	t	c	t	O=[V]O[V]=O	3.48
c	t	t	c	t	t	O=[Y]O[Y]=O	5.79
t	c	c	t	t	t	O=[Bi]O[Bi]=O	3.55
t	t	t	t	v	c	O=[In]O[In]=O	2.83
t	c	t	t	c	t	O=[Sb]O[Sb]=O	3.12
t	v	c	v	v	v	O=[Al]O[Al]=O	2.42
t	t	t	c	t	v	O=[Fe]O[Fe]=O	2.40
c	t	v	t	t	t	O=[Si]=O	2.54
v	c	v	v	t	c	O=[Zr]=O	2.58
t	t	t	v	v	c	O=[Sn]=O	2.53
t	t	t	t	t	t	O=[Ti]=O	2.14
t	t	t	c	t	t	[Co]=O	3.13
t	v	t	t	c	t	[Ni]=O	3.79
v	c	c	t	c	c	O=[Cr]O[Cr]=O	2.06
t	t	t	v	t	v	O=[La]O[La]=O	4.96
t	t	t	c	c	t	O=[Zn]^	6.23
t	t	t	t	t	t	[Cu]=O^	5.71
c	c	t	t	t	t	O=[V]O[V]=O^	3.78
t	v	c	t	v	t	O=[Y]O[Y]=O^	5.84
c	t	t	t	t	c	O=[Bi]O[Bi]=O^	4.02
t	t	v	v	t	c	O=[In]O[In]=O^	3.48
v	t	t	c	t	t	O=[Sb]O[Sb]=O^	3.66
t	v	t	v	t	t	O=[Al]O[Al]=O^	2.75
t	t	v	t	t	v	O=[Fe]O[Fe]=O^	2.54
c	v	t	c	t	v	O=[Si]=O^	2.92
v	c	c	t	v	v	O=[Zr]=O^	3.04
t	t	t	v	t	v	O=[Sn]=O^	3.24
t	t	t	t	t	t	O=[Ti]=O^	4.68
t	t	t	t	v	t	[Co]=O^	3.33
v	v	v	c	t	t	[Ni]=O^	3.87

t	t	t	t	t	t	O=[Cr]O[Cr]=O^	2.06
t	t	t	t	c	t	O=[La]O[La]=O^	5.56

*) t=training set; c=calibration set; v=validation set

Table 2
Correlation weights for calculations with Eq. 13

x_{kj}	CW(x_{kj})	Frequency in training set	Frequency in calibration set
=	0.10158	22	6
Al	0.19940	1	0
Bi	1.00478	2	0
Co	1.35252	1	0
Cr	-0.24908	1	1
Cu	3.44583	2	0
Fe	0.20463	2	0
O	-0.27911	22	6
In	0.62103	1	0
La	1.95157	1	1
Ni	2.29898	1	1
V	0.79887	1	1
Sb	0.72663	1	1
Si	0.84577	2	0
Y	2.40333	1	0
Sn	1.05005	1	0
Ti	1.64851	2	0
[	0.20343	22	6
^	1.00105	12	2
Zn	4.60092	1	1
Zr	1.09949	1	0

Table 3

Example of calculation of DCW(1,7) for Eq. 13.
The representation of metal oxide is “[Cu]=O”

DCW(1,9) = $\sum$CW(x_{kj}) = 3.67516;

pLC50 = 1.5185 + 0.8370 * DCW(1,9) = 4.5946

x_{kj}	CW(x_{kj})
[	0.20343
Cu	3.44583
[	0.20343
=	0.10158

O	-0.27911
$\sum_{x_{kj} \in quasi-SMILES} CW(x_{kj})$	3.67516

Table 4

Experimental data on features (impacts) of TiO_2 nanoparticles, and their denotations.

	A	**B**	**C**	**D**	**E**
ID	Enginered size, nm	Size in water, nm	Size in PBS, nm	Concentration, mg/L	Zeta potential, mV
1	30	125	1250	25	10
2	30	102	987	25	12
3	30	281	1543	50	15
4	30	101	1045	50	9
5	30	299	1754	100	11
6	30	134	961	100	11
7	30	600	1876	200	12
8	30	298	1165	200	12
9	45	129	2567	25	9
10	45	129	2309	25	10
11	45	201	2431	50	9
12	45	201	2987	50	11
13	45	451	2941	100	11
14	45	451	1934	100	9
15	45	876	1965	200	11
16	45	876	2109	200	10
17	125	136	3215	25	11
18	125	136	2667	25	10
19	125	149	3782	50	10
20	125	149	2144	50	15
21	125	343	3871	100	12
22	125	343	2890	100	9
23	125	967	3813	200	9
24	125	967	2671	200	8

Table 5
Normalized (Eq. 14) representation of physicochemical features of TiO_2 nanoparticles

	A	B	C	D	E
ID	**Enginered size, normalized**	**Size in water, normalized**	**Size in PBS, normalized**	**Concentration, normalized**	**-Zeta potential, normalized**
1	0.39	0.21	0.46	0.22	0.78
2	0.39	0.19	0.40	0.22	0.87
3	0.39	0.36	0.52	0.33	1.00
4	0.39	0.19	0.42	0.33	0.74
5	0.39	0.37	0.56	0.56	0.83
6	0.39	0.22	0.40	0.56	0.83
7	0.39	0.66	0.59	1.00	0.87
8	0.39	0.37	0.44	1.00	0.87
9	0.48	0.22	0.73	0.22	0.74
10	0.48	0.22	0.68	0.22	0.78
11	0.48	0.28	0.70	0.33	0.74
12	0.48	0.28	0.82	0.33	0.83
13	0.48	0.52	0.81	0.56	0.83
14	0.48	0.52	0.60	0.56	0.74
15	0.48	0.91	0.61	1.00	0.83
16	0.48	0.91	0.64	1.00	0.78
17	1.00	0.22	0.86	0.22	0.83
18	1.00	0.22	0.75	0.22	0.78
19	1.00	0.23	0.98	0.33	0.78
20	1.00	0.23	0.64	0.33	1.00
21	1.00	0.42	1.00	0.56	0.87
22	1.00	0.42	0.80	0.56	0.74
23	1.00	1.00	0.99	1.00	0.74
24	1.00	1.00	0.75	1.00	0.70

Table 6

Building up quasi-SMILES for model of membrane damage values by TiO_2 nanoparticles (MD, units/L)

	A	B	C	D	E	
ID	**Code for Enginered size**	**Code for Size in water**	**Code for Size in PBS**	**Code for Concentration**	**Code for Zeta potential**	**MD, units/L**
1	A3	B2	C4	D2	E7	0.90
2	A3	B1	C4	D2	E8	1.00
3	A3	B3	C5	D3	E9	0.75
4	A3	B1	C4	D3	E7	0.70
5	A3	B3	C5	D5	E8	1.04
6	A3	B2	C3	D5	E8	1.09
7	A3	B6	C5	D9	E8	1.15
8	A3	B3	C4	D9	E8	1.20
9	A4	B2	C7	D2	E7	0.90
10	A4	B2	C6	D2	E7	0.85
11	A4	B2	C7	D3	E7	0.75
12	A4	B2	C8	D3	E8	0.78
13	A4	B5	C8	D5	E8	1.40
14	A4	B5	C5	D5	E7	1.50
15	A4	B9	C6	D9	E8	1.35
16	A4	B9	C6	D9	E7	1.40
17	A9	B2	C8	D2	E8	1.25
18	A9	B2	C7	D2	E7	1.17
19	A9	B2	C9	D3	E7	1.00
20	A9	B2	C6	D3	E9	1.10
21	A9	B4	C9	D5	E8	1.50
22	A9	B4	C7	D5	E7	1.42
23	A9	B9	C9	D9	E7	1.60
24	A9	B9	C7	D9	E6	1.65

Table 7

Correlation weights for calculation of *DCW(T*,N*)*

Split 1		**Split 2**		**Split 3**	
x_{kj}	$CW(x_{kj})$	x_{kj}	$CW(x_{kj})$	x_{kj}	$CW(x_{kj})$
A3	-0.11150	A3	0.71150	A3	0.16450
A4	1.30300	A4	1.19800	A4	0.82400
A9	2.83850	A9	3.25100	A9	2.55400
B1	0.0	B1	0.0	B1	0.40000
B2	-0.69250	B2	-0.85300	B2	-0.04800
B3	0.15000	B3	-0.02800	B3	-0.05200
B4	0.0	B4	0.0	B4	0.0
B9	0.0	B5	1.88450	B5	2.42275
C3	0.0	B6	0.0	B9	2.41350
C4	0.0	B9	0.0	C3	0.0
C5	0.0	C3	0.0	C4	0.36575
C6	0.0	C4	0.0	C5	0.64075
C7	-0.74600	C5	-0.48000	C6	0.0
C8	0.0	C6	0.0	C7	0.58550
C9	0.0	C7	0.0	C8	-0.05400
D2	1.03450	C8	-0.04900	C9	0.63875
D3	-0.56350	C9	-0.20300	D2	1.08450
D5	2.67400	D2	1.03950	D3	-0.03000
D9	3.01650	D3	-0.21050	D5	2.63950
E7	0.14800	D5	2.34800	D9	2.48750
E8	0.27600	D9	3.10850	E6	0.0
E9	0.0	E7	0.56450	E7	0.10200
		E8	0.29250	E8	0.68875
		E9	0.0	E9	0.0

Table 8

An example of model for TiO_2 nanoparticles' membrane damage

Set	Quasi-SMILES	*DCW(2,20)*	Expr	Calc	Expr-Calc	ID
Training	A3B3C5D3E9	-0.52500	0.750	0.739	0.011	3
Training	A3B2C3D5E8	2.14600	1.090	1.079	0.011	6
Training	A3B3C4D9E8	3.33100	1.200	1.229	0.029	8
Training	A4B2C7D2E7	1.04700	0.900	0.939	0.039	9
Training	A4B2C7D3E7	-0.55100	0.750	0.735	0.015	11
Training	A4B9C6D9E7	4.46750	1.400	1.374	0.026	16
Training	A9B2C8D2E8	3.45650	1.250	1.245	0.005	17
Training	A9B2C7D2E7	2.58250	1.170	1.134	0.036	18
Training	A9B2C9D3E7	1.73050	1.000	1.026	0.026	19
Training	A9B4C7D5E7	4.91450	1.420	1.431	0.011	22
Calibration	A3B1C4D2E8	1.19900	1.000	0.958	0.042	2
Calibration	A4B2C8D3E8	0.32300	0.780	0.847	0.067	12
Calibration	A9B2C6D3E9	1.58250	1.100	1.007	0.093	20
Calibration	A9B4C9D5E8	5.78850	1.500	1.542	0.042	21
Calibration	A9B9C9D9E7	6.00300	1.600	1.570	0.030	23
Validation	A3B2C4D2E7	0.37850	0.900	0.854	0.046	1
Validation	A3B1C4D3E7	-0.52700	0.700	0.738	-0.038	4
Validation	A3B3C5D5E8	2.98850	1.040	1.186	-0.146	5
Validation	A3B6C5D9E8	3.18100	1.150	1.210	-0.060	7
Validation	A4B2C6D2E7	1.79300	0.850	1.034	-0.184	10
Validation	A4B5C8D5E8	4.25300	1.400	1.347	0.053	13
Validation	A4B5C5D5E7	4.12500	1.500	1.331	0.169	14
Validation	A4B9C6D9E8	4.59550	1.350	1.390	-0.040	15
Validation	A9B9C7D9E6	5.10900	1.650	1.456	0.194	24

Table 9
The list of conditions, having impact upon mutagenicity of fullerene C_{60} nanoparticles which were utilized to build up quasi-SMILES and models.

Conditions	**Symbols for quasi-SMILES**
Dark or Irradiation	"0" = Darkness "1" = Irradiation
Mix S9	"+" = with Mix S9 "- "= without Mix S9
Dose (g/plate)	"A" = 50 "B" = 100 "C" = 200 "D" = 400 "E" = 1000

Table 10

Experimental and calculated values of the TA100 for fullerene nanoparticles impact under different conditions

ID	**1***	**2**	**3**	**Quasi-SMILES**	**Expriment**	**Split 1**	**Split 2**	**Split 3**	**1***	**2**	**3**
1	V	V	V	0+A	146	132.8046	115.3775	143.2652	Y	Y	Y
2	T	C	V	0+B	141	145.8861	121.3559	144.0029	N	Y	Y
3	T	C	C	0+C	159	157.1709	136.5559	157.8729	N	Y	Y
4	V	C	V	0+D	160	162.0685	142.5034	161.6878	Y	Y	Y
5	T	V	T	0+E	177	165.3027	143.7145	165.5465	Y	N	Y
6	C	C	C	0-A	143	130.9643	134.7925	147.8862	Y	Y	Y
7	T	T	C	0-B	139	144.0458	140.7708	148.6238	N	Y	N
8	V	T	T	0-C	169	155.3307	155.9708	162.4939	N	Y	Y
9	T	V	C	0-D	168	160.2283	161.9183	166.3087	Y	Y	Y
10	T	T	T	0-E	152	163.4625	163.1294	170.1675	Y	N	N
11	C	V	T	1+A	129	116.4477	113.3044	130.1540	Y	Y	Y
12	T	C	T	1+B	131	129.5292	119.2827	130.8917	N	Y	Y
13	V	T	T	1+C	138	140.8141	134.4827	144.7618	N	Y	Y
14	T	T	T	1+D	137	145.7117	140.4303	148.5766	Y	Y	Y
15	C	V	T	1+E	160	148.9459	141.6413	152.4354	Y	N	N
16	V	T	T	1-A	136	114.6075	132.7193	134.7750	Y	Y	Y
17	T	T	V	1-B	136	127.6890	138.6977	135.5127	N	Y	Y
18	T	T	C	1-C	138	138.9739	153.8977	149.3827	N	Y	Y
19	C	T	T	1-D	164	143.8715	159.8452	153.1976	Y	Y	Y
20	C	T	V	1-E	172	147.1057	161.0562	157.0563	Y	N	N

*) Split 1, 2, and 3; T=training set; C=calibration set; and V=validation set; Y=quasi-SMILES falls into Domain of applicability (otherwise "N").

Table 11

Correlation weights for symbols which represent conditions of fullerene impact on TA100 mutagenicity

Symbols of quasi-SMILES, x_{kj}	***CW*(x_{kj}) *run 1***	***CW*(x_{kj}) *run 2***	***CW*(x_{kj}) *run 3***
+	1.06698	0.47336	0.19086
-	0.99597	1.77789	0.56606
0	1.37861	0.96051	1.93566
1	0.74747	0.82121	0.87109
A	0.0	0.0	-0.05990
B	0.50476	0.40170	0.0
C	0.94020	1.42303	1.12619
D	1.12918	1.82266	1.43594
E	1.25398	1.90403	1.74925

Table 12

Experimental and calculated values of the WP2uvrA/pKM101 for fullerene nanoparticles impact under different conditions

ID	**1***	**2**	**3**	**Quasi-SMILES**	**Expriment**	**Split 1**	**Split 2**	**Split 3**	**1***	**2**	**3**
1	T	C	T	0+A	113	118.9590	95.2449	133.2322	Y	Y	Y
2	T	V	C	0+B	106	118.9590	127.8341	126.7746	Y	N	Y
3	V	T	V	0+C	112	118.9590	127.9124	126.7746	Y	Y	Y
4	T	T	C	0+D	115	118.9590	95.2449	126.7746	Y	Y	Y
5	T	C	T	0+E	145	152.9472	169.4336	144.8871	Y	Y	Y
6	C	T	T	0-A	160	159.2997	136.8401	162.2816	Y	Y	Y
7	V	T	C	0-B	162	159.2997	169.4294	155.8240	Y	N	Y
8	C	V	C	0-C	174	159.2997	169.5076	155.8240	Y	Y	Y
9	V	V	T	0-D	179	159.2997	136.8401	155.8240	Y	Y	Y
10	T	T	V	0-E	220	193.2879	211.0289	173.9365	Y	Y	Y
11	V	C	T	1+A	114	84.2194	53.6913	104.3638	Y	Y	Y
12	C	V	V	1+B	105	84.2194	86.2806	97.9062	Y	N	Y
13	V	C	V	1+C	113	84.2194	86.3588	97.9062	Y	Y	Y
14	T	V	C	1+D	110	84.2194	53.6913	97.9062	Y	Y	Y
15	C	T	T	1+E	123	118.2076	127.8801	116.0187	Y	Y	Y
16	T	V	T	1-A	127	124.5601	95.2866	133.4132	Y	Y	Y
17	C	T	T	1-B	133	124.5601	127.8759	126.9556	Y	N	Y
18	T	T	V	1-C	121	124.5601	127.9541	126.9556	Y	Y	Y
19	C	C	T	1-D	117	124.5601	95.2866	126.9556	Y	Y	Y
20	T	C	T	1-E	138	158.5483	169.4754	145.0681	Y	Y	Y

*) Split 1, 2, and 3; T=training set; C=calibration set; and V=validation set; Y=quasi-SMILES falls into Domain of applicability (otherwise "N").

Table 13

Correlation weights for symbols which represent conditions of impact of fullerene C_{60} nanoparticles on mutagenicity for WP2uvrA/pKM101

Symbols of quasi-SMILES, S_k	***CW*(S_k) *run 1***	***CW*(S_k) *run2***	***CW*(S_k) *run 3***
+	0.50000	0.69675	0.39982
-	0.94039	1.60115	2.20288
0	1.12778	1.60305	2.19630
1	0.74854	0.69727	0.40447
A	0.0	0.0	0.40082
B	0.0	0.99558	0.0
C	0.0	0.70300	0.0
D	0.0	0.0	0.0
E	0.37104	1.59979	1.12422

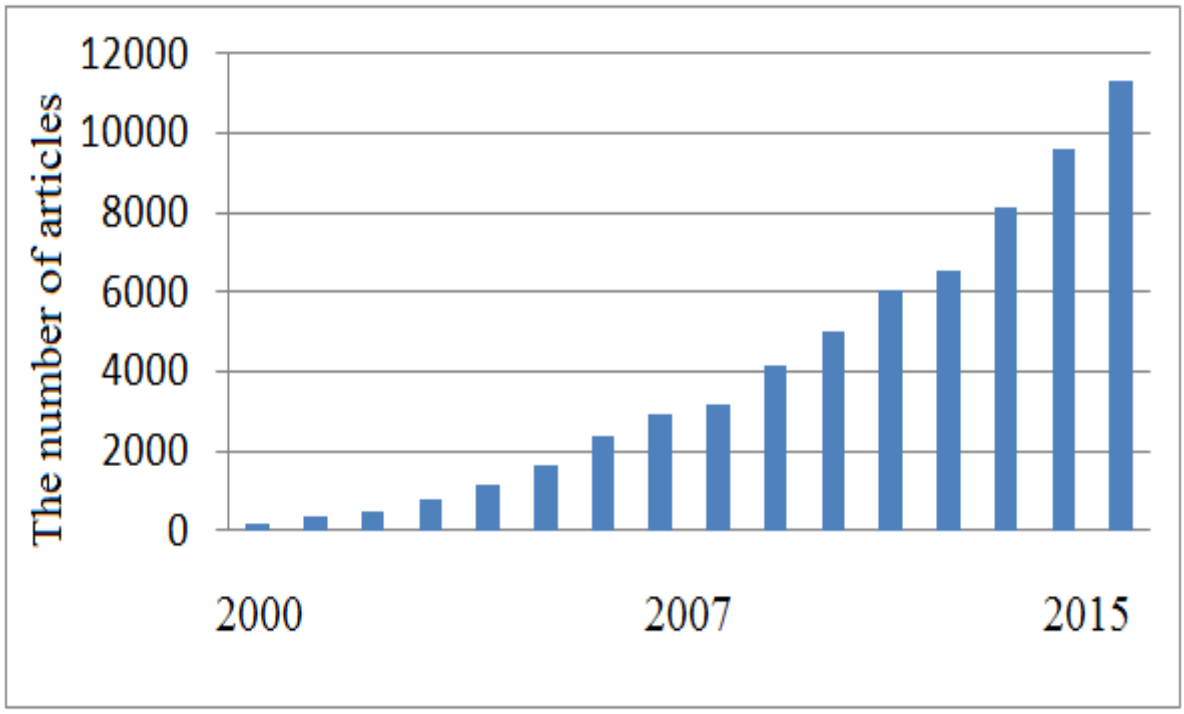

Figure 1
The increase of numbers of articles related to keyword “nanomaterial” (2000 to 2015), according to: www.sciencedirect.com.

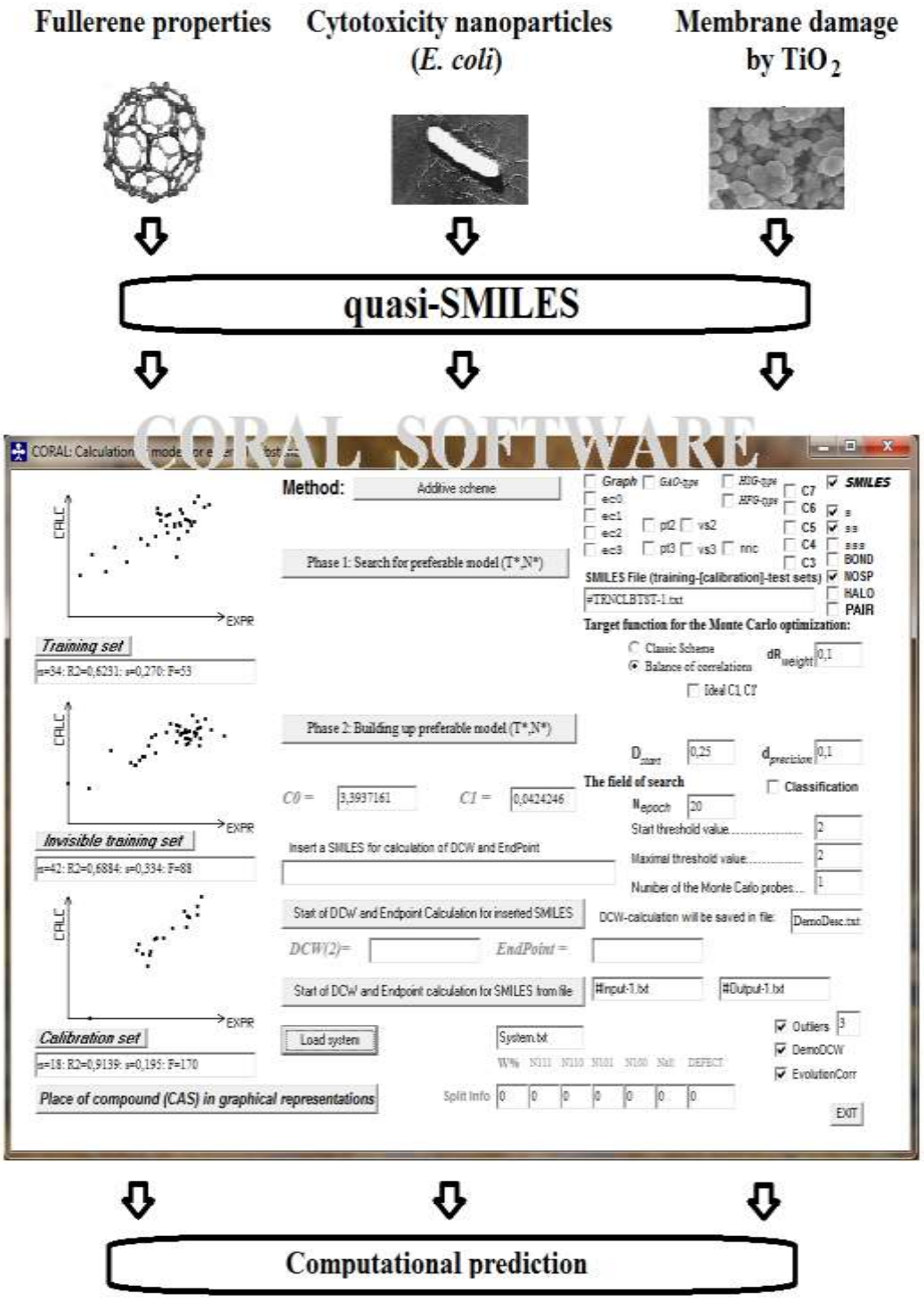

Figure 2
The general scheme of utilization of quaisi-SMILES

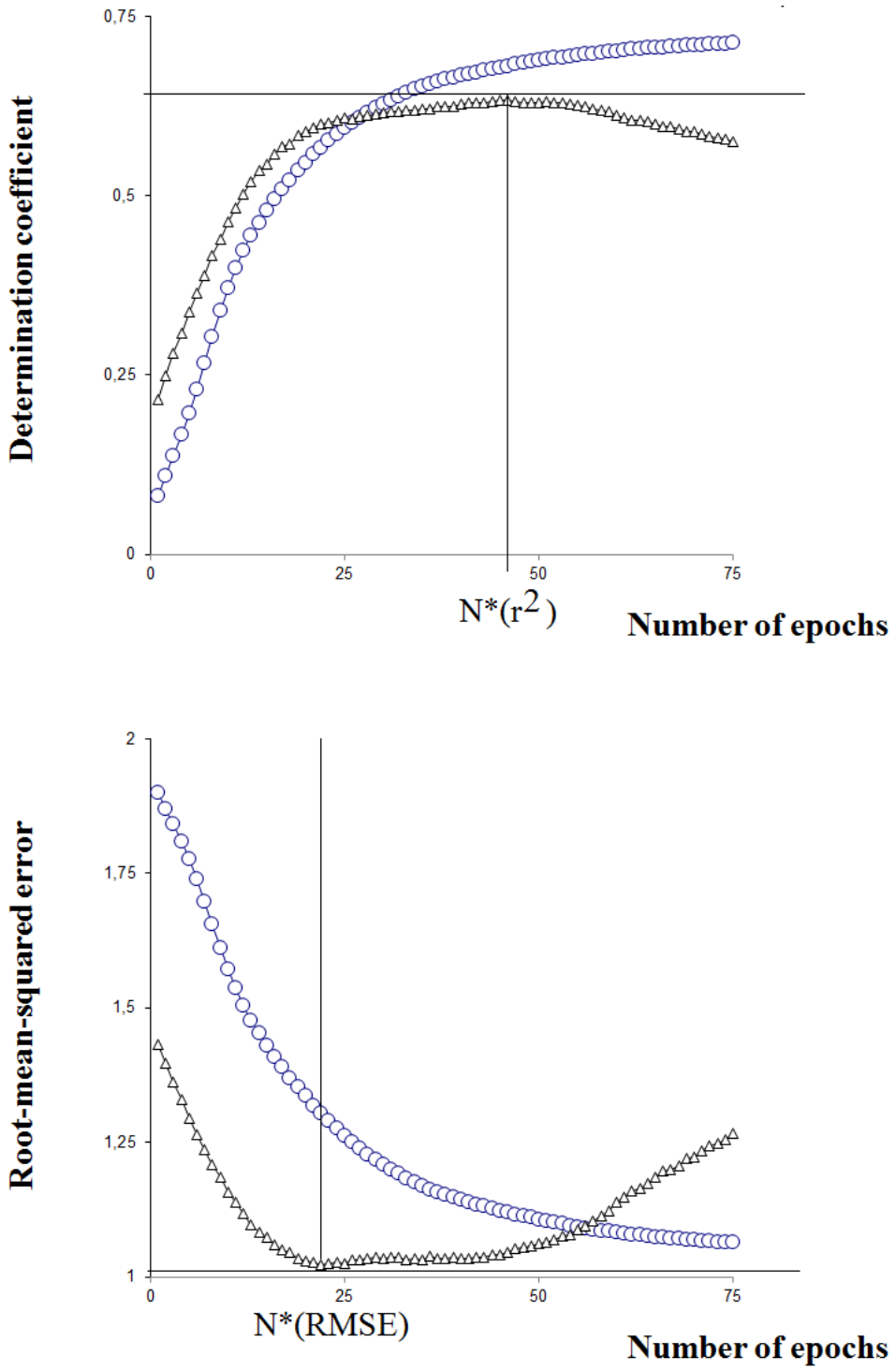

Figure 3
The selection of the number of epochs of the Monte Carlo optimization using: (i) the correlation coefficient between experimental and predicted values of an endpoint; and (ii) root-mean squared error.

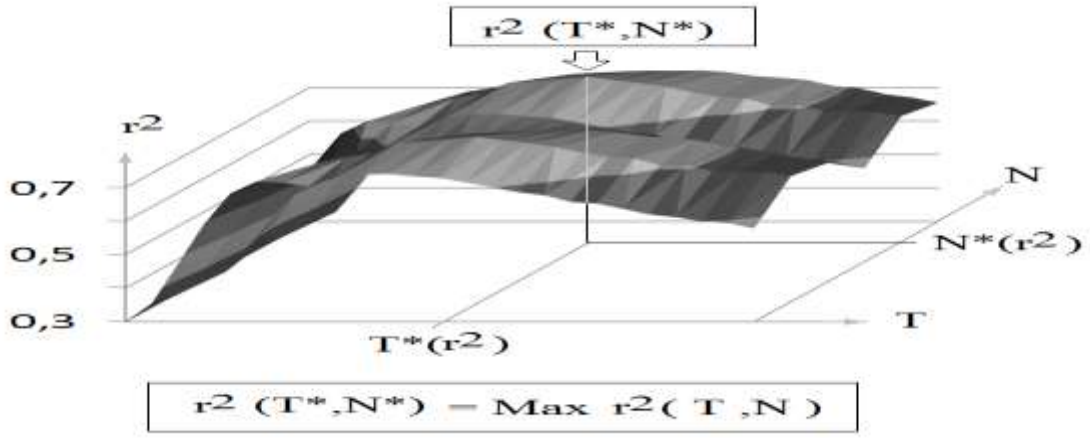

Figure 4
The scheme of the definition of the T* and N* values which give preferable statistics for the calibration set.

X=A,B,C,D,E

9, Norm(X)>0.9	**X9**
8, 0.8<Norm(X)<0.9	**X8**
7, 0.7<Norm(X)<0.8	**X7**
6, 0.6<Norm(X)<0.7	**X6**
5, 0.5<Norm(X)<0.6	**X5**
4, 0.4<Norm(X)<0.5	**X4**
3, 0.3<Norm(X)<0.4	**X3**
2, 0.2<Norm(X)<0.3	**X2**
1, 0.1<Norm(X)<0.2	**X1**
0, Norm(X)<0.1	**X0**

Figure 5
Partition of normalized physicochemical features into categories 1, 2, …, 9 according to its value.