Quasi-SMILES as a tool to utilize eclectic data for predicting the behavior of nanomaterials

Alla P. Toropova*, Andrey A. Toropov, Serena Manganelli, Caterina Leone, Diego Baderna, Emilio Benfenati, Roberto Fanelli

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

Published version of this paper could be find here https://doi.org/10.1016/j.impact.2016.04.003

NanoImpact Volume 1, January 2016, Pages 60-64

Abstract

Nowadays, nanomaterials are often considered a scientific hit. However, despite the immense advantages of nanomaterials, there are studies, which have shown that these materials can also harmfully impact both human health and the environment. A preliminary evaluation of the hazards related to nanomaterials can be performed using predictive models. The aim of the present study is building up a single QSAR model for predicting cytotoxicity of metal oxide nanoparticles on (i) *Escherichia coli* (*E. coli*) and (ii) human keratinocyte cell line (HaCaT) based on the representation of the available eclectic data, encoded into quasi-SMILES. Quasi-SMILES is an analogue and an attractive alternative of traditional simplified molecular input-line entry systems (SMILES). In contrast to traditional SMILES quasi-SMILES are a tool to represent not only molecular structures, but also different conditions, such as physicochemical properties and experimental conditions. The statistical quality of the models are average correlation coefficient (r^2) and root mean squared error (RMSE) for the training set 0.79 and 0.216; the average r^2 and RMSE for validation set are 0.90 and 0.247, respectively.

Keywords: Nano-QSAR; Nanoparticles; Cytotoxicity; HaCaT; *Escherichia coli*; quasi-SMILES

*) Corresponding author
Alla P. Toropova, PhD
Laboratory of Environmental Chemistry and Toxicology,
IRCCS - Istituto di Ricerche Farmacologiche Mario Negri,
Via La Masa 19, 20156 Milano, Italy
Tel: +39 02 3901 4595
Fax: +3902 3901 4735
Email: alla.toropova@marionegri.it

1. Introduction

Human exposure to NPs has been in existence for many years. It involves public and occupational health exposure to ultrafine particulate air pollution. A broader source of exposure is related to nanoparticles which are abundant in nature, as they are produced in many natural processes, including photochemical reactions, volcanic eruptions, forest fires, and simple erosion, and by plants and animals [1].

In more recent years, due to the rapid expansion of nanotechnology, environmental and human exposure to engineered nanoparticles has also become unavoidable [2].

For this reason, the need to gain knowledge about safety and potential hazards of nanoparticles is dramatically increasing. Within this context, nanotoxicology has become an emerging discipline. However, while the number of nanoparticle types and their applications continues to increase, studies to characterize their effects after exposure and to address their potential toxicity are few in comparison. In the medical field in particular, nanoparticles are being utilized in diagnostic and therapeutic tools to better understand, detect, and treat human diseases. Exposure to nanoparticles for medical purposes involves intentional contact and control; therefore, understanding the properties of nanoparticles and their effect on the body is crucial before clinical use can occur. The first step towards understanding how an agent will react in the body often involves cell-culture studies. Compared to animal studies, cellular testing is less ethically ambiguous, is easier to control and reproduce, and is less expensive [3].

Building up predictive models for endpoints related to nanomaterials is an important task of modern natural sciences [4]. Likely, the traditional quantitative structure – property / activity relationships (QSPRs/QSARs) [5-13] based on the molecular structure are not able to solve this task.

The problem with nanomaterials is that a chemical structure is not sufficient to describe them so that a range of other unique properties needs to be considered, including particle size, shape and surface [14].

A model for endpoints related to nanomaterials can be organized in the following form: the measured calculated endpoint is a mathematical function of all available eclectic information, which may be (i) chemical structure, (ii) atom compositions, (iii) conditions of synthesis/preparation of the nanomaterial, (iv) the features of nanomaterials related to their manufacture. This list can be easily extended (size, porosity, symmetry, electromechanical properties, etc.). To define a predictive model for an endpoint related to nanomaterials the traditional paradigm for QSAR modeling, 'Endpoint = F (molecular structure)', can be replaced by 'Endpoint = F (eclectic information)' [15-19].

The aim of the present work is an attempt to build up united predictive model for two endpoints : (i) cytotoxicity to *Escherichia coli* and (ii) human keratinocyte cell line (HaCaT) for metal nanoparticles using optimal descriptors based on quasi-SMILES. Quasi-SMILES is a modification of the traditional simplified molecular input-line entry systems (SMILES) [20-22] representing eclectic data using a string of characters, encoding particular conditions, not of the molecular structure. In fact, the aim of the present work can be also defined as an attempt to answer question: “How one should organize databases related to nanomaterials in order to extract from these databases satisfactory prediction of the behavior for nanomaterials, which were not examined in experiment?”

2. Method

2.1. Data

The endpoint considered for the QSAR analysis was cytotoxicity of metal oxide nanoparticle on *Escherichia Coli* (*E. coli*) [23] and human keratinocyte cell line (HaCaT) [24], expressed as the negative logarithm of half maximal effective concentration (pEC_{50}). pEC_{50} data (mol/L) were taken from the literature (see Table 1). Figure 1 shows the toxicity data for nano-sized metal oxides against E.coli and HaCaT cells: pEC_{50} values on HaCaT are higher in comparison to those obtained from *E. coli*. This trend of toxicity is reversed only for In_2O_3, SnO_2, and TiO_2, which are more toxic to HaCaT than to *E. Coli* [25].

[Table 1 around here]

The total set of available data has been split (three times) into the training (n=22), calibration (n=5), and validation (n=5) sets. These splits are built up according to principles: (i) these splits are random; (ii) the ranges of endpoints are similar for each sub-set (i.e. for the training, calibration, and validation set); and (iii) these splits are different. It is possible to notice that there is a good balance of cytotoxicity data between the two sets of values. Furthermore, the cytotoxicity ranges are also quite similar going from 1.76 to 3.32 in the case of line cell line and in the case of E.coli from 1.74 to 3.45. These values are given as pEC50 where EC50 is the cytotoxicity effect observed the dose which produces effect on 50% of the cells.

In fact these endpoints are a mathematical function of the same conditions (same structures of nano oxides) and two additional codes (%11 and %12) give possibility to attempt to build up united model for these endpoints. The similar approach was used in work [26] for united model of mutagenicity for fullerene and multi walled carbon nanotubes (MWCNTs) under different conditions.

2.2. Optimal descriptor

Optimal descriptors also called ‘quasi-SMILES’, of nanoQSAR analysis were calculated with CORAL software [27]. These were built and optimized starting by the coding of an experimental condition (*in vitro* test): HaCaT and E. Coli were encoded as "%11" and "%12" respectively. These codes were combined with the traditional SMILES of nano-oxides (see Table 1). The 32 resulting combined systems (traditional SMILES- *in vitro* test) were randomly split into training, calibration and validation sets, with similar distribution of endpoint values.

Optimal descriptors were calculated as follows:

$$DCW(T,N) = \sum CW(S_k) \quad (1)$$

where $CW(S_k)$ are the correlation weights for each fragment S_k contained in the quasi-SMILES (Table 2).

[Table 2 around here]

The correlation weights are calculated using the Monte Carlo optimization method [12-19]. The optimization process make use of two parameters: (i) the threshold (T), which is a tool for classifying codes as either rare (and thus likely less reliable features, probably introducing noise into the model) or not rare features, which are used by the model and labeled as active; and (ii) the number of epochs (N), which is the number of cycles (sequence of modifications of correlation weights for all codes involved in model development) for the optimization [15-18]. The target function of the optimization procedure is the correlation coefficient between cytotoxicity and descriptors calculated with Eq. 1 for the training set. However, the process should be stopped when the correlation coefficient for the calibration set reach maximum. If the process will be continued after this maximum, the model most probably will give the overtraining (i.e. excellent statistical quality for the training set, but poor quality for the calibration and for the validation set).

Thus, the model should be optimized using codition the T=T* and N=N* which give the maximum of the correlation coefficient for the calibration set. These T* and N* should be defined from computational calculations with T from range $\{T_1, T_2, \ldots, T_n\}$ and N from range $\{1, 2, \ldots, N\}$. Having the correlation weights obtained by described manner, one can calculate with using the Eq. 1 the optimal descriptor for any system of eclectic conditions and by utilizing the systems of the training set build up a model:

$$pEC50 = C_0 + C_1 * DCW(T^*,N^*) \quad (2)$$

The model should be checked up with the calibration set and if the statistical quality is satisfactory, then the obtained model should has a predictive potential. The validation set in the described scheme of building up models plays role of the final estimator of the predictive potential for Eq. 2.

Thus, as it was noted above, instead of the traditional QSAR paradigm “Endpoint = F (Molecular structure)” the new paradigm “Endpoint = F(Eclectic data)” is suggested.

3. Results and Discussion

Comparison of suggested approach with models suggested in work [25] has apparent limitations. First of all the aim of the above work is to develop nano quantitative toxicity– toxicity relationship (nano-QTTR) with involving some descriptors of quantum mechanics whereas this work is aimed to develop integrated model based on elementary data on molecular structure of metal nano oxides together with taking into account objects for their impacts (E. coli and HaCaT). Thus one can note (i) the models calculated with Eq. 3, Eq. 4, and Eq. 5 are identical for all thirty two situations of acting of metal nano oxides represented by the quasi-SMILES; (ii) the models suggested here do not involve additional information (descriptors of quantum mechanics).

3.1. How one can utilize these models?

How, one should define "input" data and how one should define expected results?

One should define request as Eclectic data which contain two components: (i) traditional SMILES for metal nano-oxide (Table 1); and (ii) code %11 in order to obtain prediction of pEC50 for cytotoxicity human keratinocyte cell line (HaCaT) or code %12 in order to obtain pEC50 for cytotoxicity to *Escherichia coli.*

3.2. Predictive models

The described approach gives the following models:

Split 1

$$pEC50 = 1.6840375 (\pm 0.0214373) + 0.2883483 (\pm 0.0063152) * DCW(1,15) \quad (3)$$

Split 2

$$pEC50 = 1.3816828 (\pm 0.0300053) + 0.3657955 (\pm 0.0089238) * DCW(1,15) \quad (4)$$

Split 3

$$pEC50 = -0.0009168 (\pm 0.0455860) + 0.4622782 (\pm 0.0074398) * DCW(1,30) \quad (5)$$

Table 2 contains the correlation weights $CW(S_k)$ for calculation $DCW(T^*,N^*)$ with Eq. 1 Table 3 contains the statistical characteristics of models for three random splits. One can see that statistical characteristics of models for each split are different, but quite good. Table 4 contains an example of the DCW(T*,N*) calculation. Table 5 contains the splits into the training, calibration, and validation sets together with the numerical data on the experimental and predicted pEC50. Table 6 contains the comparison of the statistical quality of models suggested in work [25] and models calculated with quasi-SMILES.

[Table 3 around here]

[Table 4 around here]

[Table 5 around here]

[Table 6 around here]

3.3. OECD principles

The described approach build up predictive models according to OECD principles (Table 7) [29].

[Table 7 around here]

4. Conclusions

The suggested approach gives quite satisfactory models for the eclectic data related to cytotoxicity towards *Escherichia coli* and human keratinocyte cell line (HaCaT) for metal nanoparticles. The possibility to build up predictive databases using eclectic data is demonstrated. The quasi-SMILES are analogy of the traditional SMILES, but have additional possibility to involve in building up a model different conditions. Described actions can be repeated and improved by means of utilization available on the Internet the CORAL software [27].

Acknowledgments

Authors thank EC project PeptiCAPS (Project reference: 686141) for financial support.

References

[1] C. Buzea, I. Pacheco, K. Robbie, Nanomaterials and nanoparticles: Sources and toxicity, Biointerphases 2 (2007) MR17-MR71.

[2] P. C. Ray, H. Yu, P. P. Fu, Toxicity and Environmental Risks of Nanomaterials: Challenges and Future Needs, J. Environ. Sci. Health., Part C Environ. Carcinog. Ecotoxicol. Rev. 27 (1) (2009) 1-35.

[3] N. Lewinski, V. Colvin, R. Drezek, Cytotoxicity of nanopartides, Small 4 (1) (2008) 26-49.

[4] K.P. Singh, S. Gupta, Nano-QSAR modeling for predicting biological activity of diverse nanomaterials RSC Adv., 4 (2014) 13215-13230.

[5] G. Melagraki, A. Afantitis, Enalos KNIME nodes: Exploring corrosion inhibition of steel in acidic medium, Chemom . Intell. Lab. Syst. 123 (2013) 9-14.

[6] L. Scotti, E.O. Lima, M.S. da Silva, H. Ishiki, I.O. Lima, F.O. Pereira, F.J.B. Mendonça Jr, M.T. Scotti, Docking and PLS Studies on a Set of Thiophenes RNA Polymerase Inhibitors Against Staphylococcus aureus, Curr. Top. Med. Chem. 14 (1) (2014) 64-80.

[7] A.A. Toropov, A.P. Toropova,I. Raska Jr, D. Leszczynska, J. Leszczynski,Comprehension of drug toxicity: Software and databases, Comput. Biol. Med. 45 (2014) 20-25.

[8] V.V. Kleandrova, F. Luan, A. Speck-Planche, M.N.D.S. Cordeiro, In silico assessment of the acute toxicity of chemicals: Recent advances and new model for multitasking prediction of toxic effect, Mini Rev. Med. Chem. 15 (8) (2015) 677-686.

[9] A. Speck-Planche, M.N.D.S. Cordeiro, Multi-target QSAR approaches for modeling protein inhibitors. Simultaneous prediction of activities against biomacromolecules present in Gram-negative bacteria, Curr. Top. Med. Chem. 15 (18) (2015) 1801-1813.

[10] P.R. Duchowicz, S.E. Fioressi, D.E. Bacelo, L.M. Saavedra, A.P. Toropova, A.A. Toropov, QSPR studies on refractive indices of structurally heterogeneous polymers, Chemom. Intell. Lab. Syst. 140 (2015) 86-91.

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

[12] A.M. Veselinović, J.B. Veselinović, A.A. Toropov, A.P. Toropova, G.M. Nikolić, In silico prediction of the β-cyclodextrin complexation based on Monte Carlo method, Int. J. Pharm. 495 (1) (2015) 404-409.

[13] J.B. Veselinović, A.A. Toropov, A.P. Toropova, G.M. Nikolić, A.M. Veselinović, Monte Carlo method-based QSAR modeling of penicillins binding to human serum proteins, Arch. Pharm. 348 (1) (2015) 62-67.

[14] A.A. Toropov, A.P. Toropova, E. Benfenati, G. Gini, T. Puzyn, D. Leszczynska, J. Leszczynski, Novel application of the CORAL software to model cytotoxicity of metal oxide nanoparticles to bacteria *Escherichia coli*, Chemosphere 89 (9) (2012) 1098-1102.

[15] A.A. Toropov, A.P. Toropova, Quasi-QSAR for mutagenic potential of multi-walled carbon-nanotubes, Chemosphere 124 (1) (2015) 40-46.

[16] S. Manganelli, C. Leone, A.A. Toropov, A.P. Toropova, E. Benfenati, QSAR model for predicting cell viability of human embryonic kidney cells exposed to SiO2 nanoparticles. Chemosphere 144 (2016) 995-1001.

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

[18] A.A. Toropov, R. Rallo, A.P. Toropova, Use of Quasi-SMILES and monte carlo optimization to develop quantitative feature property/activity relationships (QFPR/QFAR) for nanomaterials, Curr. Top. Med. Chem. 15 (18) (2015) 1837-1844.

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

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

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

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

[23] T. Puzyn, B. Rasulev, A. Gajewicz, X. Hu, T.P. Dasari, A. Michalkova, H.M. Hwang, A. Toropov, D. Leszczynska, J. Leszczynski, Using nano-QSAR to predict the cytotoxicity of metaloxide nanoparticles. Nat. Nanotechnol.6 (2011) 175–178.

[24] A. Gajewicz, N. Schaeublin, B. Rasulev, S. Hussain, D. Leszczynska, T. Puzyn, J. Leszczynski, Towards understanding mechanisms governing cytotoxicity of metaloxides nanoparticles: hints from nano-qsar studies. Nanotoxicology 9 (2015) 313–325.

[25] S. Kar, A. Gajewicz, K. Roy, J. Leszczynski, T. Puzyn, Extrapolating between toxicity endpoints of metal oxide nanoparticles: Predicting toxicity to Escherichia coli and human keratinocyte cell line (HaCaT) with Nano-QTTR, Ecotoxicol. Environ. Saf. 126 (2016) 238-244.

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

[27] CORAL, http://www.insilico.eu/coral (Accessed Feb 5, 2016)

[28] P.K. Ojha, K. Roy, Comparative QSARs for antimalarial endochins: Importance of descriptor-thinning and noise reduction prior to feature selection, Chemom. Intell. Lab. Syst. 109 (2) (2011) 146-161.
[29] Organisation For Economic Co-Operation And Development (OECD). 2007; http://www.oecd.org/dataoecd/55/35/38130292.pdf (Accessed Feb 5, 2016)

Table 1
Numerical data on the toxicity to *Escherichia coli* and human keratinocyte cell line (HaCaT)

No.	Nano-oxide	Traditional SMILES	Additional codes: HaCaT=%11 *E. coli*=%12	pEC50 in molar scale
1.	Al2O3	O=[Al]O[Al]=O	%11	1.85
2.	Bi2O3	O=[Bi]O[Bi]=O	%11	2.5
3.	CoO	[Co]=O	%11	2.83
4.	Cr2O3	O=[Cr]O[Cr]=O	%11	2.3
5.	Fe2O3	O=[Fe]O[Fe]=O	%11	2.05
6.	In2O3	O=[In]O[In]=O	%11	2.92
7.	La2O3	O=[La]O[La]=O	%11	2.87
8.	NiO	[Ni]=O	%11	2.49
9.	Sb2O3	O=[Sb]O[Sb]=O	%11	2.31
10.	SiO2	O=[Si]=O	%11	2.12
11.	SnO2	O=[Sn]=O	%11	2.67
12.	TiO2	O=[Ti]=O	%11	1.76
13.	V2O3	O=[V]O[V]=O	%11	2.24
14.	Y2O3	O=[Y]O[Y]=O	%11	2.21
15.	ZnO	O=[Zn]	%11	3.32
16.	ZrO2	O=[Zr]=O	%11	2.02
17.	Al2O3	O=[Al]O[Al]=O	%12	2.49
18.	Bi2O3	O=[Bi]O[Bi]=O	%12	2.82
19.	CoO	[Co]=O	%12	3.51
20.	Cr2O3	O=[Cr]O[Cr]=O	%12	2.51
21.	Fe2O3	O=[Fe]O[Fe]=O	%12	2.29
22.	In2O3	O=[In]O[In]=O	%12	2.81
23.	La2O3	O=[La]O[La]=O	%12	2.87
24.	NiO	[Ni]=O	%12	3.45
25.	Sb2O3	O=[Sb]O[Sb]=O	%12	2.64
26.	SiO2	O=[Si]=O	%12	2.2
27.	SnO2	O=[Sn]=O	%12	2.01
28.	TiO2	O=[Ti]=O	%12	1.74
29.	V2O3	O=[V]O[V]=O	%12	3.14
30.	Y2O3	O=[Y]O[Y]=O	%12	2.87
31.	ZnO	O=[Zn]	%12	3.45
32.	ZrO2	O=[Zr]=O	%12	2.15

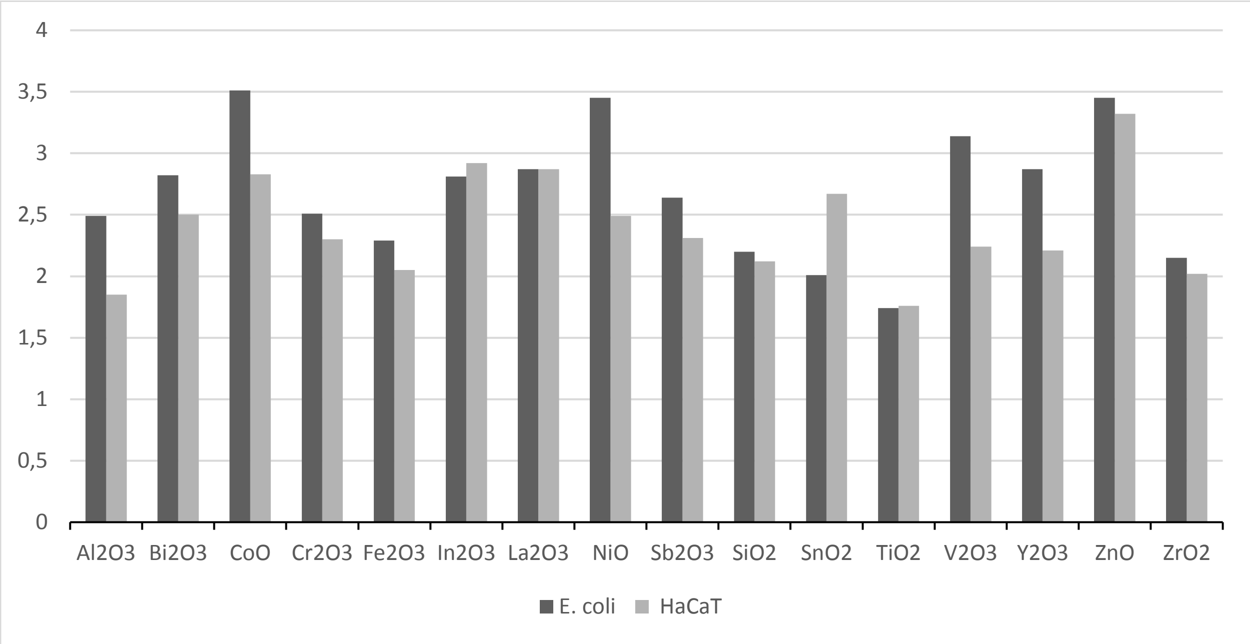

Figure 1

Nanoparticle toxicity data, expressed as pEC_{50}, for E.coli and HaCaT

Table 2
Correlation weights for calculation with Eq. 1 for three random splits

Split 1		**Split 2**		**Split 3**	
S_k	***$CW(S_k)$***	***S_k***	***$CW(S_k)$***	***S_k***	***$CW(S_k)$***
%11	1.17055	%11	0.92518	%11	1.58993
%12	2.17144	%12	1.61214	%12	2.14798
=	-0.55123	=	-0.04516	=	1.07580
Al	-0.26129	Al	0.56062	Al	0.59662
Bi	0.95121	Bi	0.97207	Bi	0.92776
Co	3.16018	Co	2.80291	Co	4.23560
Cr	0.45955	Cr	0.59633	Cr	0.67787
Fe	0.0	Fe	0.0	Fe	0.41164
O	-0.34249	O	-0.30102	O	-0.42509
In	1.37575	In	1.24544	In	1.34598
La	1.63282	La	1.25245	La	1.29975
Ni	3.69871	Ni	2.79766	Ni	3.35802
V	1.05102	V	1.00488	V	1.31730
Sb	0.58840	Sb	0.78321	Sb	0.76343
Si	-0.06530	Si	0.87119	Si	1.06687
Y	0.77211	Y	0.80024	Y	0.79239
Sn	0.95754	Sn	1.44623	Sn	1.32284
Ti	-0.70465	Ti	-0.49905	Ti	0.0
[	0.25238	[	0.32036	[	0.28833
Zn	4.21583	Zn	4.00446	Zn	4.19637
Zr	0.00869	Zr	0.69706	Zr	0.77528

Table 3
The statistical characteristics of the models of pEC50 for three splits into the training, calibration, and validation sets

		Split 1	Split 2	Split 3
Training set (n=22)	r^2	0.79	0.74	0.85
	q^2	0.76	0.69	0.83
	RMSE	0.230	0.227	0.191
Calibration set (n=5)	r^2	0.84	0.90	0.90
	RMSE	0.248	0.237	0.441
	${}^{c}R_p^2$ *(should be>0.5)	0.76	0.77	0.70
	$\overline{r_m^2}$ (should be >0.5)	0.78	0.79	0.68
	Δr_m^2 (should be<0.2)	0.062	0.103	0.137
Validation set (n=5)	r^2	0.96	0.88	0.87
	RMSE	0.242	0.257	0.244

*) Description of ${}^{c}R_p^2$, $\overline{r_m^2}$, and Δr_m^2 is available in work [28].

Table 4
An example of the ***DCW(T*,N*)*** calculation for Eq. 3

Attributes of quasi-SMILES, $\boldsymbol{S_k}$	$\boldsymbol{CW(S_k)}$	Frequency in training set	Frequency in calibration set
O	-0.3496	22	5
=	-0.3017	22	5
[	0.3244	22	5
Al	-0.0963	1	1
[	0.3244	22	5
O	-0.3496	22	5
[	0.3244	22	5
Al	-0.0963	1	1
[	0.3244	22	5
=	-0.3017	22	5
O	-0.3496	22	5
%11	1.1293	13	1
$\boldsymbol{DCW(1,15) = \sum CW(S_k) =}$	**0.58213**		

pEC50 = 1.6840 + 0.28835 * 0.58213 = 1.851857

Table 5
The splits into the training (t), calibration (c), and validation (v) sets. Numerical data on experimental and predicted values of the pEC50

1	2	3	Quasi-SMILES	Expriment [25]	Eq. 3	Eq. 4	Eq. 5	Model from Ref. 25
t	v	v	O=[Al]O[Al]=O%11	1.85	1.8519	2.2356	2.2239	1.98
t	t	t	O=[Bi]O[Bi]=O%11	2.50	2.5146	2.5366	2.5301	2.58
t	t	c	[Co]=O%11	2.83	3.0169	2.8531	3.2595	2.97
c	c	t	O=[Cr]O[Cr]=O%11	2.30	2.1537	2.2618	2.2991	2.28
v	c	t	O=[Fe]O[Fe]=O%11	2.05	1.9075	1.8255	2.0529	2.17
t	t	t	O=[In]O[In]=O%11	2.92	2.7312	2.7366	2.9168	2.92
t	t	t	O=[La]O[La]=O%11	2.87	2.8856	2.7418	2.8740	2.83
c	t	t	[Ni]=O%11	2.49	2.8421	2.8512	2.8538	2.55
t	c	v	O=[Sb]O[Sb]=O%11	2.31	2.3122	2.3985	2.3782	2.33
v	t	t	O=[Si]=O%11	2.12	1.8343	2.0199	2.0955	1.99
t	t	t	O=[Sn]=O%11	2.67	2.1539	2.2302	2.2138	2.24
t	v	c	O=[Ti]=O%11	1.76	1.7621	1.5187	1.6023	1.90
t	t	c	O=[V]O[V]=O%11	2.24	2.5135	2.5606	2.8903	2.17
t	t	t	O=[Y]O[Y]=O%11	2.21	2.3708	2.4109	2.4049	2.15
t	t	t	O=[Zn]%11	3.32	3.1637	3.2927	3.2414	3.26
t	t	t	O=[Zr]=O%11	2.02	1.8993	1.9562	1.9607	2.23
c	t	t	O=[Al]O[Al]=O%12	2.49	2.1970	2.4869	2.4819	2.50
v	c	t	O=[Bi]O[Bi]=O%12	2.82	2.8598	2.7879	2.7881	2.75
t	c	t	[Co]=O%12	3.51	3.3621	3.1044	3.5175	3.49
t	t	v	O=[Cr]O[Cr]=O%12	2.51	2.4988	2.5130	2.5570	2.60
c	v	t	O=[Fe]O[Fe]=O%12	2.29	2.2526	2.0768	2.3109	2.43
t	t	c	O=[In]O[In]=O%12	2.81	3.0763	2.9879	3.1747	2.81
v	t	v	O=[La]O[La]=O%12	2.87	3.2307	2.9930	3.1320	2.95
t	t	t	[Ni]=O%12	3.45	3.1873	3.1025	3.1118	3.38
v	t	t	O=[Sb]O[Sb]=O%12	2.64	2.6573	2.6498	2.6361	2.63
t	t	v	O=[Si]=O%12	2.20	2.1795	2.2712	2.3534	1.98
t	t	t	O=[Sn]=O%12	2.01	2.4991	2.4815	2.4718	2.15
c	t	c	O=[Ti]=O%12	1.74	2.1072	1.7700	1.8602	1.92
t	t	t	O=[V]O[V]=O%12	3.14	2.8587	2.8119	3.1482	2.68
t	t	t	O=[Y]O[Y]=O%12	2.87	2.7160	2.6622	2.6629	2.83
c	v	t	O=[Zn]%12	3.45	3.5088	3.5440	3.4993	3.63
t	v	t	O=[Zr]=O%12	2.15	2.2445	2.2075	2.2186	2.14

Table 6
Comparison of statistical characteristics of models from work [25] and models calculated with quasi-SMILES (i.e. Eqs. 3, 4, and 5)

Endpoint	**n**	$\mathbf{r^2}$	**RMSE**	$\overline{r_m^2}$	Δr_m^2
pEC_{50} HaCaT [25]	16	0.88	0.22	0.74	0.04
pEC_{50} E. coli [25]	16	0.91	0.19	0.82	0.09
pEC_{50}(HaCaT, E.coli), split 1	32	0.80	0.23	0.71	0.04
pEC_{50}(HaCaT, E.coli), split 2	32	0.80	0.23	0.71	0.11
pEC_{50}(HaCaT, E.coli), split 3	32	0.80	0.24	0.71	0.08

Table 7

The compliance to the OECD principles

No.	Definition	How a principle is taken into account in this work?
1	a defined endpoint	Two endpoints are united into one
2	an unambiguous algorithm	Monte Carlo optimization with available software [27]
3	a defined domain of applicability	Probabilistic criteria to define domain of applicability according to distribution of available data into the training and calibration set [15-19, 27]
4	appropriate measures of goodness-of-fit, robustness and predictivity	The traditional criteria which are utilized for the QSPR/QSAR [15-19, 27]
5	a mechanistic interpretation, if possible	Available after several runs of the Monte Carlo optimization [15-19, 27]