Nano-QSAR in cell biology: Model of cell viability as a mathematical function of available eclectic data

Alla P. Toropova*, Andrey A. Toropov

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

Published version of this paper could be find here

https://doi.org/10.1016/j.jtbi.2017.01.012

Journal of Theoretical Biology, Volume 416, 7 March 2017, Pages 113-118

*) 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

ABSTRACT

The prediction of biochemical endpoints is an important task of the modern medicinal chemistry, cell biology, and nanotechnology. Simplified molecular input-line entry system (SMILES) is a tool for representation of the molecular structure. In particular, SMILES can be used to build up the quantitative structure – property / activity relationships (QSPRs/QSARs). The QSPR/QSAR is a tool to predict an endpoint for a new substance, which has not been examined in experiment. Quasi-SMILES are representation of eclectic data related to an endpoint. In contrast to traditional SMILES, which are representation of the molecular structure, the quasi-SMILES are representation of conditions (in principle, the molecular structure also can be taken into account in quasi-SMILES). In this work, the quasi-SMILES were used to build up model for cell viability under impact of the metal-oxides nanoparticles by means of the CORAL software (http://www.insilico.eu/coral). The eclectic data for the quasi-SMILES are (i) molecular structure of metals-oxides; (ii) concentration of the nanoparticles; and (iii) the size of nanoparticles. The significance of different eclectic facts has been estimated. Mechanistic interpretation and the domain of applicability for the model are suggested. The statistical quality of the models is satisfactory for three different random distribution of available data into the training (sub-training and calibration) and the validation sets.

Keywords: quasi-SMILES; Nano-QSAR; Nano-QFAR; Monte Carlo method; CORAL software

1. Introduction

The satisfactory prediction of various endpoints related to nanomaterials is valuable knowledge for the development and modification of nanomaterials and nanotechnologies (Melagraki and Afantitis, 2014, 2015). Apparently, however, that the influence of nanomaterials upon live systems (cells, organs, organisms) can be dangerous, hence the risk assessment of the nanomaterials is one of important tasks of the medicinal chemistry (Esposito et al., 2015). The traditional toolbox for prediction of endpoints related to the organic substances (Duchowicz et al., 2012), polymers (Mercader and Duchowicz, 2016), organometallic compounds (Toropova et al., 2013) contains approaches based on the descriptors calculated with the molecular graph (Furtula and Gutman,

2011) and/or data on physicochemical endpoints (Sayes and Ivanov, 2010). In the case of the nanomaterials, application of the molecular graph is limited, because (i) as rule the molecular structure of nanomaterials is very large, and (ii) the percentage of dissimilarity of molecules of nanomaterials, which are members of congeneric group (e.g., fullerene derivatives, single walled carbon nanotubes, multi walled carbon nanotubes, etc.) is very small. Thus, the using of the traditional quantitative structure – property / activity relationships (QSPRs/QSARs), which are based on manipulation with similarity and dissimilarity of the molecular structures becomes very problematic. The second above-mentioned conception of the predictive models, which are based on physicochemical data also meet difficulties: in particular, large databases on physicochemical parameters of nanomaterials remain unavailable. Under such circumstances, the definition of fresh conceptions of predictive models for nanomaterials becomes important task (Kahru and Ivask, 2013; Kleandrova et al., 2014a,b; Luan et al., 2014; Speck-Planche et al., 2015).

Simplified molecular input-line entry system (SMILES) (Weininger, 1988; Weininger et al., 1989; Toropova et al., 2014) is the representation of the molecular structure for the traditional QSPR/QSAR analyses. Modification of SMILES by means of extension of available meanings of the sequence of symbols can be a way to define a new approach of building up predictive models for endpoints related to situations where molecular structure has not key role and an endpoint value is defined by conditions of acting of a nanomaterial. Essence of this paradigm is "*endpoint is a mathematical function of eclectic data*". The eclectic data can involve the following: size of nanoparticle; concentration; various technological conditions; condition of synthesis of nanoparticles; dark or irradiation; presence of different chemical elements; exposure time; and others. The above mentioned eclectic data can be united in quasi-SMILES (Toropova and Toropov, 2015; Toropov and Toropova, 2015a,b; Toropova et al., 2015a; Toropov et al., 2015; Toropova et al., 2016). Building up of a predictive model for cell viability under impact of different metal-oxide nanoparticles characterized by various metals, size, and various concentration is the aim of this work.

2. Method

2.1. Data

The data on the viability of BEAS-2B cells taken in the literature (Huang et al., 2010). The nanoparticles of the following metal-oxides are examined: Fe_2O_3, Cr_2O_3, TiO_2, Mn_2O_3, NiO, CoO, ZnO, and CuO. In order to extract numerical data graphics (Huang et al., 2010), the DataThief software (http://www.datathief.org/) has been used (Table 1). Fifty quasi-SMILES with various value of the cell viability (CV%) are extracted. The total set has been distributed into the training (structured as sub-training and calibration sets) and validation set. Three distributions were examined in order to check up the approach. The distribution were build up according to the following rules (i) these distributions are random; (ii) these distributions are different; and (iii) the range of the cell viability for the training and validation sets is similar.

2.2. Building up quasi-SMILES

Quasi-SMILES contain three components: (i) data on the metal-oxide nanoparticle in the form of traditional SMILES generated with the ACD/ChemSketch software (http://www.acdlabs.com); (ii) code of nanoparticle size (Table 2); and (iii) code of nanoparticle concentration (Table 3). The traditional SMILES separated by dot (Table 4).

2.3. Bulding up nano-QFAR

Since the model for cell viability is a mathematical function of structure of metal-oxides, size, and concentration of nanoparticles, words “structure – activity” should be replaced by “feature-activity”. Consequently, instead of traditional abbreviation "QSAR", one should use more adequate “QFAR”. The QFAR built up in this work are based on the optimal descriptors calculated with the quasi-SMILES:

$$DCW(T^*, N^*) = \Sigma CW(F_k) \quad (1)$$

where F_k is *k-th* feature of a nanoparticle (Table 2 and Table 3);

The $CW(F_k)$ is the correlation weight for the *k-th* feature. The correlation weights for various features involved in building up model are special coefficients calculated by the Monte Carlo method optimization.

The correlation coefficient between cell viability and $DCW(T^*,N^*)$ is a mathematical function of the correlation weights { $CW(F_k)$ } and of two parameters of the optimization. The T (threshold) is a coefficient to classify all features into two classes (i) rare or noise, if the number of a $CW(F_k)$ in the training set is less than T; and (ii) active if the number of a $CW(F_k)$ in the training set is larger than T (or equal to T). The rare features are blocked: their correlation weights are defined equal to zero. The T can be 1, 2, …, m. The N is the number of epochs of the Monte Carlo optimization. The N can be 10, 15, 70, 100, etc.

The $T=T^*$ and $N=N^*$ are such values of the above-mentioned parameters of the optimization which gives maximum for correlation coefficient between *CV%* and $DCW(T^*,N^*)$ for calibration set.

$$R_{CALIBRATION}[CV\%, DCW(T^*, N^*)] \rightarrow \max \quad (2)$$

In other words, the model is based on the hypothesis that good statistical quality of model for the calibration set (data on this set serve only to check up the model, whereas data on the training set are used to build up the model) should be accompanied by satisfactory statistical quality for the external validation set.

Having numerical data on the correlation weights of all features involved in building up the model together with values of the T^* and N^*, predictive model can be calculated using the training set:

$$CV\% = C_0 + C_1 \times DCW(T^*, N^*) \quad (3)$$

The predictive potential of the model calculated with Eq. 3 should be checked up with the validation set. The above-mentioned calculations are carried out with the CORAL software (http://www.insilico.eu/coral).

3. Results

3.1. nano-QFAR for three random distributions into training and validation sets

The QFAR calculated by the above-mentioned scheme (Eq.1 – Eq. 3) for prediction of the cell viability (*CV%*) are the following:

$$CV\% = -141.8151(\pm 5.0497) + 24.5908(\pm 0.6115) * DCW(1,20) \quad (4)$$

$$CV\% = -129.0514(\pm 4.8267) + 23.7234(\pm 0.6063) * DCW(1,18) \quad (5)$$

$$CV\% = -168.5388(\pm 5.8212) + 24.0969(\pm 0.5994) * DCW(1,11) \quad (6)$$

Table 4 contains the experimental and calculated with Eqs. 4-6 CV% values. Table 5 contains the statistical characteristics of these models.

3.2. Mechanistic interpretation

Three runs of Monte Carlo optimizations with *DCW(1,20), DCW(1, 18)* and *DCW(1,11)* for distributions 1, 2, and 3, respectively, indicate that there are stable promoters of the CV% increase. Table 6 contains the following promoters of CV% increase: (i) ‘A’, concentration <1 μg/mL; (ii) ‘1’, size 50 nm; (iii) ‘8’ size 20 nm; (iv) ‘Co’, cobalt; (v) ‘Zn’, zink; (vi) ‘H’, 7 μg/mL < concentration < 8 μg/mL; (vii) ‘J’, 9 μg/mL < concentration < 10 μg/mL; (viii) ‘G’, 6 μg/mL < concentration < 7 μg/mL; (ix) ‘Cr’, chrome. Table 6 contains only one promoter of CV% decrease: this is ‘S’, i.e. 90 μg/mL < concentration < 100 μg/mL. Other features are not available for the

analysis, since they absent in the calibration sets for three examined distributions into training, calibration, and validation sets.

3.3. Domain of applicability

Different distributions into the training, calibration, and validation sets are characterized by different prevalence of features in the training and calibration sets. The measure of influence of a feature F_k for possible predictive potential of a model can be estimated via defect of F_k , d(F_k) calculated as the following:

$$d(F_k) = \frac{P_T(F_k) - P_C(F_k)}{N_T(F_k) + N_C(F_k)} \quad (7),$$

where $P_T(F_k)$ and $P_C(F_k)$ are probabilities of attribute F_k in the training and the calibration set, respectively; $N_T(F_k)$ and $N_C(F_k)$ are prevalence of attribute F_k in the training and the calibration set, respectively. The $d(F_k)$ =1, if $N_C(F_k)$ = 0.

The defect of quasi-SMILES $d(qS)$ can be estimated via defects of F_k which presence in the quasi-SMILES:

$$d(qS) = \sum_{F_k \in qS} d(F_k) \rightarrow \min \quad (8)$$

The defect of a distribution (Split) into the training, calibration, and validation sets can be estimated via sum of defects of quasi-SMILES:

$$d(Split) = \sum d(qS) \rightarrow \max$$

Computational experiments have shown, that described models have preferable predictive potential if, (i) each d(qS) is minimal; but (ii) d(Split) is maximal. The statistically robust domain of applicability can be introduced via inequality

$$d(qS) < 2 \times \overline{d(qS)} \quad (9),$$

where $\overline{d(qS)}$ is average defect of quasi-SMILES over the training set. Table 7 contains defects of quasi-SMILES related to three distributions. Table 4 contains domain of applicability (Y/N) for

models calculated with Eqs. 4-6 defined by inequality 9. The numbers of outliers for distributions 1, 2, and 3 are 14, 3, and 16, respectively. It is large percentage of outliers, however, taking into account the small number of available quasi-SMILES and their heterogeneity, this situation should be recognized as quite realistic.

4. Discussion

Described models have very similar statistical characteristics on the training, calibration, and validation sets. The statistical quality of these models can be estimated as at least semi-quantitative one. Of course, the dispersion of the experimental measurement of the CV% characterized by more accuracy (Huang et al., 2010).

Unfortunately, the fifty quasi-SMILES examined in this work (Table 4) have low level of similarities. Owing to this circumstances, the distribution where majority of the features takes place in both the training and calibration sets at least several times becomes almost unavailable (Table 6).

However, in fact, the possibility of more or less satisfactory predictions by means of described scheme is demonstrated. It is to be noted, the scheme of definition of the mechanistic interpretations via the promoters of increase/decrease for various endpoints is checked up in traditional QSPR/QSAR (Toropova et al., 2015a; Toropova et al., 2015b) as well as with some untypical predictive models (Toropova et al., 2015b; Veselinović et al., 2015), and, finally, for nano-QSAR (Toropov and Toropova; 2015a).

The domain of applicability via inequality 9, has been checked up for the traditional QSPR/QSAR (Toropova et al., 2015a; Toropova et al., 2015b). In this work, the above mentioned conception of the domain of applicability lead to large percentage of outliers owing to heterogeneity of available data, but in the case of more congeneric datasets, the percentage of outliers can be smaller even if the total number of quasi-SMILES is not large (Toropov and Toropova; 2015a).

5. Conclusions

The described approach based on the Monte Carlo technique gives semi-quantitative prediction for cell viability under impact different metal-oxide nanoparticles. The limitation in variation of quasi-SMILES lead to necessity of construction of models for small number of external validation sets (Table 7). However, this approach can give prediction with more accuracy for extended datasets, expected in the near future. The suggested models built up according to OECD principles (OECD, 2007).

Acknowledgements

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

References

Duchowicz, P.R., Comelli, N.C., Ortiz, E.V., Castro, E.A., 2012. QSAR study for carcinogenicity in a large set of organic compounds. Curr. Drug Saf. 7 (4), 282-288.

Esposito, E.X., Hopfinger, A.J., Shao, C.-Y., Sue, B.-H., Chen, S.-Z., Tseng, Y.J., 2015. Exploring possible mechanisms of action for the nanotoxicity and protein binding of decorated nanotubes: interpretation of physicochemical properties from optimal QSAR models. Toxicol. Appl. Pharm. 288, 52–62.

Furtula, B., Gutman, I., 2011. Relation between second and third geometric-arithmetic indices of trees. J. Chemometrics 25 (2), 87-91.

Huang, Y.-W., Wu, C.-H., Aronstam, R.S., 2010. Toxicity of transition metal oxide nanoparticles: Recent insights from in vitro studies. Materials 3 (10), 4842-4859.

Kahru, A., Ivask, A., 2013. Mapping the dawn of nanoecotoxicological research. Acc. Chem. Res. 46 (3), 823-833.

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

Kleandrova, V.V., Luan, F., González-Díaz, H., Ruso, J.M., Speck-Planche, A., Cordeiro, M.N.D.S., 2014b. Computational tool for risk assessment of nanomaterials: Novel QSTR-perturbation model for simultaneous prediction of ecotoxicity and cytotoxicity of uncoated and coated nanoparticles under multiple experimental conditions. Environ. Sci. Technol. 48 (24), 14686-14694.

Luan, F., Kleandrova, V.V., González-Díaz, H., Ruso, J.M., Melo, A., Speck-Planche, A., Cordeiro, M.N.D.S., 2014. Computer-aided nanotoxicology: Assessing cytotoxicity of nanoparticles under diverse experimental conditions by using a novel QSTR-perturbation approach. Nanoscale 6 (18), 10623-10630.

Melagraki, G., Afantitis, A., 2014. Enalos InSilicoNano platform: An online decision support tool for the design and virtual screening of nanoparticles. RSC Advances 4 (92), 50713-50725.

Melagraki, G., Afantitis, A., 2015. A risk assessment tool for the virtual screening of metal oxide nanoparticles through enalos insiliconano platform. Curr. Top. Med. Chem. 15 (18), 1827-1836.

Mercader, A.G., Duchowicz, P.R., 2016. Encoding alternatives for the prediction of polyacrylates glass transition temperature by quantitative structure-property relationships. Mater. Chem. Phys. 172, 158-164.

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

Sayes, C., Ivanov, I., 2010. Comparative study of predictive computational models for nanoparticle-induced cytotoxicity. Risk Anal. 30 (11), 1723-1734.

Speck-Planche, A., Kleandrova, V.V., Luan, F., Cordeiro, M.N.D.S., 2015. Computational modeling in nanomedicine: Prediction of multiple antibacterial profiles of nanoparticles using a quantitative structure-activity relationship perturbation model. Nanomedicine 10 (2), 193-204.

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

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

Toropov, A.A., Rallo, R., Toropova, A.P., 2015. 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), 1837-1844.

Toropova, A.P., Toropov, A.A., Benfenati, E., Gini, G., Leszczynska, D., Leszczynski, J., 2013. CORAL: QSPRs of enthalpies of formation of organometallic compounds. J. Math. Chem. 51 (7), 1684-1693.

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

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

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

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

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

Toropova, M.A., Toropov, A.A., Raška, I., Rašková, M., 2015a. Searching therapeutic agents for treatment of Alzheimer disease using the Monte Carlo method. Comput. Biol. Med. 64, 148-154.

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

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

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

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

Table 1

Experimental data on the cell viability (BEAS-2B cells exposed for 24 h) extracted from Figure 2, represented in work (Huang et al., 2010).

Metal-oxide nanoparticle	Size, nm	Concentration, µg/ml	Cell Viability (% of control)
CoO	50	0.00E+00	1.00E+02
		4.60E+00	9.00E+01
		7.05E+00	5.96E+01
		9.80E+00	2.61E+01
		2.51E+01	8.95E+00
		4.99E+01	4.98E+00
		1.00E+02	5.76E-01
Cr2O3	60	0.00E+00	1.01E+02
		6.43E+00	1.01E+02
		9.80E+00	1.01E+02
		2.48E+01	9.80E+01
		4.99E+01	8.96E+01
		1.00E+02	8.03E+01
CuO	30-50	0.00E+00	1.01E+02
		0.00E+00	8.61E+01
		9.19E-01	7.02E+01
		1.23E+00	5.26E+01
		1.84E+00	3.85E+01
		5.21E+00	1.86E+01
		1.01E+01	1.20E+01
		1.99E+01	8.07E+00
Fe2O3	20-60	0.00E+00	1.00E+02
		9.80E+00	8.61E+01
		2.51E+01	8.39E+01
		5.02E+01	8.08E+01
		1.00E+02	7.73E+01
Mn2O3	30-60	0.00E+00	1.01E+02
		9.80E+00	9.22E+01
		2.48E+01	7.06E+01
		4.99E+01	3.94E+01
		1.00E+02	2.35E+01
NiO	10-20	0.00E+00	1.01E+02
		5.51E+00	8.61E+01
		1.01E+01	6.84E+01
		2.48E+01	4.82E+01
		4.99E+01	3.80E+01
		1.00E+02	2.61E+01
TiO2	10-30	0.00E+00	1.01E+02
		7.66E+00	9.53E+01

		2.51E+01	9.36E+01
		4.99E+01	9.75E+01
		1.00E+02	1.06E+02
ZnO	20	0.00E+00	1.01E+02
		5.82E+00	9.75E+01
		6.13E+00	8.30E+01
		6.43E+00	6.14E+01
		7.35E+00	4.51E+01
		7.66E+00	2.57E+01
		7.97E+00	1.56E+01
		1.01E+01	1.03E+01

Table 2

The scheme to build up codes of nanoparticle size

Size, nm	**Code**
50	‘1’
60	‘2’
30-50	‘3’
20-60	‘4’
30-60	‘5’
10-20	‘6’
10-30	‘7’
20	‘8’

Table 3

The scheme to build up codes of nanoparticle concentration

Concentration, C, μg/mL	**Code**
$0< C < 1$	‘A’
$1< C \leq 2$	‘B’
$2< C \leq 3$	‘C’
$3< C \leq 4$	‘D’
$4< C \leq 5$	‘E’
$5< C \leq 6$	‘F’
$6< C \leq 7$	‘G’
$7< C \leq 8$	‘H’
$8< C \leq 9$	‘I’
$9< C \leq 10$	‘J’
$10< C \leq 20$	‘K’
$20< C \leq 30$	‘L’
$30< C \leq 40$	‘M’
$40< C \leq 50$	‘N’
$50< C \leq 60$	‘O’
$60< C \leq 70$	‘P’
$70< C \leq 80$	‘Q’
$80< C \leq 90$	‘R’
$90< C \leq 100$	‘S’

Table 4

Experimental and calculated cell viability (CV, %)

	Distributions			Quasi-SMILES	Cell viability, CV%				DA*		
ID	**1**	**2**	**3**		**CV% experiment**	**Eq.4**	**Eq.5**	**Eq.6**	**1**	**2**	**3**
1	T	T	T	[Co]=O.1A	100.0	81.5766	81.1472	80.1179	Y	Y	Y
2	T	T	T	[Co]=O.1E	90.0	90.2321	90.2944	90.2356	Y	Y	Y
3	T	V	T	[Co]=O.1H	59.6	33.6832	33.0560	34.6469	Y	Y	Y
4	T	T	C	[Co]=O.1J	26.1	49.1138	49.1292	46.8824	Y	Y	Y
5	T	T	T	[Co]=O.1L	8.95	30.8629	30.2375	34.6363	Y	Y	Y
6	V	T	V	[Co]=O.1N	4.98	27.5125	21.3324	23.2744	Y	Y	Y
7	C	C	T	[Co]=O.1S	0.576	16.3959	15.7609	16.6091	Y	Y	Y
8	C	T	V	O=[Cr]O[Cr]=O.2A	101.0	139.3566	139.6393	135.8546	Y	Y	Y
9	T	T	T	O=[Cr]O[Cr]=O.2G	101.0	109.6141	110.4138	109.3573	Y	Y	Y
10	C	C	C	O=[Cr]O[Cr]=O.2J	101.0	106.8939	107.6213	102.6192	Y	Y	Y
11	T	V	T	O=[Cr]O[Cr]=O.2L	98.0	88.6429	88.7296	90.3730	Y	Y	Y
12	V	T	V	O=[Cr]O[Cr]=O.2N	89.6	85.2925	79.8245	79.0112	Y	Y	Y
13	V	V	V	O=[Cr]O[Cr]=O.2S	80.3	74.1759	74.2530	72.3459	Y	Y	Y
14	T	T	T	[Cu]=O.3A	101.0	70.9656	71.3665	71.0413	Y	Y	Y
15	T	T	T	[Cu]=O.3A	86.1	70.9656	71.3665	71.0413	Y	Y	Y
16	T	T	T	[Cu]=O.3A	70.2	70.9656	71.3665	71.0413	Y	Y	Y
17	T	T	V	[Cu]=O.3B	52.6	52.9216	52.9652	52.8416	N	Y	N
18	V	V	T	[Cu]=O.3B	38.5	52.9216	52.9652	52.8416	N	Y	N
19	T	T	T	[Cu]=O.3F	18.6	53.5494	53.7140	53.9941	N	Y	N
20	T	T	V	[Cu]=O.3K	12.0	13.9745	14.6261	15.2059	N	Y	N
21	T	T	T	[Cu]=O.3K	8.07	13.9745	14.6261	15.2059	N	Y	N
22	T	T	T	O=[Fe]O[Fe]=O.4A	100.0	112.5027	112.6714	119.3886	N	Y	N
23	T	T	T	O=[Fe]O[Fe]=O.4J	86.1	80.0399	80.6534	86.1531	N	Y	N
24	V	T	T	O=[Fe]O[Fe]=O.4L	83.9	61.7890	61.7617	73.9070	N	N	N
25	T	T	T	O=[Fe]O[Fe]=O.4O	80.8	75.0173	71.8501	72.7402	N	Y	N
26	V	T	V	O=[Fe]O[Fe]=O.4S	77.3	47.3220	47.2851	55.8798	N	Y	N
27	T	T	T	O=[Mn]O[Mn]=O.5A	101.0	109.0449	106.8798	105.4754	Y	Y	N
28	T	T	T	O=[Mn]O[Mn]=O.5J	92.2	76.5821	74.8618	72.2399	Y	Y	N
29	T	T	T	O=[Mn]O[Mn]=O.5L	70.6	58.3312	55.9701	59.9938	Y	N	N
30	C	T	T	O=[Mn]O[Mn]=O.5N	39.4	54.9808	47.0650	48.6320	Y	N	N
31	T	T	T	O=[Mn]O[Mn]=O.5S	23.5	43.8642	41.4935	41.9667	Y	Y	N
32	T	T	T	[Ni]=O.6A	101.0	101.9638	102.8669	102.2622	Y	Y	Y
33	T	T	T	[Ni]=O.6F	86.1	84.5476	85.2145	85.2150	N	Y	Y
34	T	T	T	[Ni]=O.6K	68.4	44.9726	46.1266	46.4268	N	Y	Y
35	T	T	C	[Ni]=O.6L	48.2	51.2500	51.9572	56.7806	N	Y	Y
36	T	T	T	[Ni]=O.6N	38.0	47.8996	43.0521	45.4188	Y	Y	Y
37	T	V	T	[Ni]=O.6S	26.1	36.7831	37.4806	38.7535	Y	Y	Y
38	T	T	T	O=[Ti]=O.7A	101.0	142.2408	143.7393	140.5071	Y	Y	Y
39	V	T	T	O=[Ti]=O.7H	95.3	94.3474	95.6481	95.0361	Y	Y	Y

40	T	T	T	O=[Ti]=O.7L	93.6	91.5271	92.8296	95.0255	N	Y	Y
41	T	V	T	O=[Ti]=O.7N	97.5	88.1767	83.9244	83.6637	Y	Y	N
42	T	C	T	O=[Ti]=O.7S	106.0	77.0601	78.3529	76.9984	Y	Y	Y
43	C	T	C	[Zn]=O.8A	101.0	81.9821	82.3986	80.0647	Y	Y	Y
44	T	C	T	[Zn]=O.8F	97.5	64.5659	64.7462	63.0174	Y	Y	Y
45	C	T	C	[Zn]=O.8G	83.0	52.2396	53.1732	53.5673	Y	Y	Y
46	T	T	T	[Zn]=O.8G	61.4	52.2396	53.1732	53.5673	Y	Y	Y
47	C	T	C	[Zn]=O.8H	45.1	34.0887	34.3074	34.5936	Y	Y	Y
48	T	T	T	[Zn]=O.8H	25.7	34.0887	34.3074	34.5936	Y	Y	Y
49	T	C	T	[Zn]=O.8H	15.6	34.0887	34.3074	34.5936	Y	Y	Y
50	T	T	T	[Zn]=O.8K	10.3	24.9910	25.6583	24.2293	Y	Y	Y

*) T, C, and V are denominations of the training, calibration, and validation sets, respectively.

**) The DA is the domain of applicability according to inequality 9, for distribution 1, 2, and 3

Table 5

The statistical characteristics of models calculated with Eqs. 4-6.

Distribution	n_{train}	r^2_{train}	q^2_{train}	S_{train}	n_{calib}	r^2_{calib}	S_{calib}	n_{valid}	r^2_{valid}	S_{valid}
1, Eq.4	36	0.7405	0.6953	17.6	7	0.6992	23.9	7	0.7041	18.9
2, Eq.5	37	0.7406	0.6932	17.5	6	0.7149	25.0	7	0.7388	18.3
3, Eq.6	37	0.7192	0.6728	17.8	6	0.8132	18.7	7	0.7732	20.0

Table 6

Promoters of increase/decrease of cell viability

F_k	*CWs Run 1*	*CWs Run 2*	*CWs Run 3*	N_T^*	N_C	$d(F_k)$
	Distribution 1					
A	2.37958	2.24967	2.18587	8	2	0.0063
1	1.54152	1.41253	1.61298	5	1	0.0007
8	1.54204	1.50417	1.67687	5	3	0.0362
Co	1.52466	1.67447	1.47718	5	1	0.0007
Zn	1.54831	1.57997	1.42919	5	3	0.0362
H	0.39014	0.27965	0.39725	3	1	0.0149
J	1.04045	0.91325	0.95866	3	1	0.0149
G	1.17620	1.04800	1.08340	2	1	0.0291
Cr	0.89792	0.87660	0.72016	2	2	0.0575
S	-0.30166	-0.37694	-0.25158	3	1	0.0149
	Distribution 2					
A	2.41110	2.41714	2.40492	8	2	0.0117
1	1.61375	1.56122	1.72912	5	1	0.0053
8	1.69636	1.28871	1.58639	5	3	0.0456
Co	1.44138	1.46524	1.51403	5	1	0.0053
Zn	1.39537	1.77047	1.69193	5	3	0.0456
H	0.31591	0.43931	0.41304	3	1	0.0214
J	1.00769	1.06499	1.11695	3	1	0.0214
G	1.16438	1.19861	1.22387	2	1	0.0375
Cr	0.93809	0.87972	0.94967	2	2	0.0698
S	-0.44907	-0.31108	-0.27845	3	1	0.0214
	Distribution 3					
A	2.47293	2.44183	2.54674	8	1	0.0055
1	1.43477	1.52346	1.54594	5	1	0.0053
8	1.52450	1.77538	1.65361	5	3	0.0456
Co	1.80204	1.62009	1.52433	5	1	0.0053
Zn	1.72790	1.38502	1.42610	5	3	0.0456
H	0.58583	0.65760	0.66175	4	1	0.0117
J	1.15095	1.14805	1.17687	3	1	0.0214
G	1.39911	1.44591	1.47288	2	1	0.0375
Cr	0.88310	0.84926	0.85494	2	1	0.0375
S	-0.10426	-0.04951	-0.04710	3	1	0.0214

*) The N_T is the number of F_k in the training set; the N_C is the number of F_k in the calibration set; the $d(F_k)$ is defect of a feature calculated with Eq. 7.

Table 7

Defects of quasi-SMILES and values of average defects of quasi-SMILES over training set for three distributions into the training, calibration, and validation sets examined in this work.

ID	Quasi SMILES	*D(qS)*	*D(qS)*	*D(qS)*
1	[Co]=O.1A	0.0077	0.0222	0.0160
2	[Co]=O.1E	1.0013	1.0105	1.0105
3	[Co]=O.1H	0.0162	0.0319	0.0222
4	[Co]=O.1J	0.0162	0.0319	0.0319
5	[Co]=O.1L	1.0013	1.0105	0.0158
6	[Co]=O.1N	0.0304	1.0105	1.0105
7	[Co]=O.1S	0.0162	0.0319	0.0319
8	O=[Cr]O[Cr]=O.2A	0.1790	0.2212	0.1181
9	O=[Cr]O[Cr]=O.2G	0.2017	0.2470	0.1502
10	O=[Cr]O[Cr]=O.2J	0.1875	0.2309	0.1340
11	O=[Cr]O[Cr]=O.2L	1.1726	1.2095	0.1179
12	O=[Cr]O[Cr]=O.2N	0.2017	1.2095	1.1126
13	O=[Cr]O[Cr]=O.2S	0.1875	0.2309	0.1340
14	[Cu]=O.3A	2.0063	2.0117	2.0055
15	[Cu]=O.3A	2.0063	2.0117	2.0055
16	[Cu]=O.3A	2.0063	2.0117	2.0055
17	[Cu]=O.3B	3.0000	3.0000	3.0000
18	[Cu]=O.3B	3.0000	3.0000	3.0000
19	[Cu]=O.3F	3.0000	3.0000	3.0000
20	[Cu]=O.3K	3.0000	3.0000	3.0000
21	[Cu]=O.3K	3.0000	3.0000	3.0000
22	O=[Fe]O[Fe]=O.4A	3.0063	3.0117	3.0055
23	O=[Fe]O[Fe]=O.4J	3.0149	3.0214	3.0214
24	O=[Fe]O[Fe]=O.4L	4.0000	4.0000	3.0053
25	O=[Fe]O[Fe]=O.4O	3.0000	3.0000	3.0000
26	O=[Fe]O[Fe]=O.4S	3.0149	3.0214	3.0214
27	O=[Mn]O[Mn]=O.5A	0.0254	3.0117	3.0055
28	O=[Mn]O[Mn]=O.5J	0.0339	3.0214	3.0214
29	O=[Mn]O[Mn]=O.5L	1.0190	4.0000	3.0053
30	O=[Mn]O[Mn]=O.5N	0.0481	4.0000	4.0000
31	O=[Mn]O[Mn]=O.5S	0.0339	3.0214	3.0214
32	[Ni]=O.6A	2.0063	2.0117	0.0160
33	[Ni]=O.6F	3.0000	3.0000	1.0105
34	[Ni]=O.6K	3.0000	3.0000	1.0105
35	[Ni]=O.6L	3.0000	3.0000	0.0158
36	[Ni]=O.6N	2.0291	3.0000	1.0105
37	[Ni]=O.6S	2.0149	2.0214	0.0319
38	O=[Ti]=O.7A	2.0063	2.0117	2.0055
39	O=[Ti]=O.7H	2.0149	2.0214	2.0117
40	O=[Ti]=O.7L	3.0000	3.0000	2.0053
41	O=[Ti]=O.7N	2.0291	3.0000	3.0000

42	O=[Ti]=O.7S	2.0149	2.0214	2.0214
43	[Zn]=O.8A	0.0788	0.1029	0.0967
44	[Zn]=O.8F	1.0724	1.0912	1.0912
45	[Zn]=O.8G	0.1015	0.1288	0.1288
46	[Zn]=O.8G	0.1015	0.1288	0.1288
47	[Zn]=O.8H	0.0873	0.1126	0.1029
48	[Zn]=O.8H	0.0873	0.1126	0.1029
49	[Zn]=O.8H	0.0873	0.1126	0.1029
50	[Zn]=O.8K	1.0724	1.0912	1.0912
$2\times\overline{d(qS)}$		**2.7344**	**3.5868**	**2.8659**
$d(Split)$		**13.31**	**16.35**	**13.29**