CORAL and Nano-OFAR: Quantitative feature – activity relationships (QFAR) for bioavailability of nanoparticles (ZnO, CuO, Co_3O_4, and TiO_2)

Alla P. Toropova[1*], Andrey A. Toropov[1], Danuta Leszczynska[2], Jerzy Leszczynski[3]

[1]*IRCCS-Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20156 Milan, Italy*

[2]*Interdisciplinary Nanotoxicity Center, Department of Civil and Environmental Engineering, Jackson State University, 1325 Lynch Street, Jackson, MS 39217-0510, USA*

[3]*Interdisciplinary Nanotoxicity Center, Department of Chemistry and Biochemistry, Jackson State University, 1400 J. R. Lynch Street, P.O. Box 17910, Jackson, MS 39217, USA*

Published version of this paper could be find here https://doi.org/10.1016/j.ecoenv.2017.01.054

Ecotoxicology and Environmental Safety, Volume 139, May 2017, Pages 404-40

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

Abstract

Quantitative feature – activity relationships (QFAR) approach was applied to prediction of bioavailability of metal oxide nanoparticles. ZnO, CuO, Co_3O_4, and TiO_2 nanoxides were considered. The computational model for bioavailability of investigated species is asserted. The model was calculated using the Monte Carlo method. The CORAL free software (http://www.insilico.eu/coral) was used in this study. The developed model was tested by application of three different splits of data into the training and validation sets. So-called, quasi-SMILES are used to represent the conditions of action of metal oxide nanoparticles. A new paradigm of building up predictive models of endpoints related to nanomaterials is suggested. The paradigm is the following "An endpoint is a mathematical function of available eclectic data (conditions)". Recently, the paradigm has been checked up with endpoints related to metal oxide nanoparticles, fullerenes, and multi-walled carbon-nanotubes.

Keywords: QSAR; nano-QSAR; QFAR; quasi-SMILES; CORAL free software

1. Introduction

There are two types of works dedicated to searching for predictive models for endpoints related to nanomaterials. The first type is reviews of models, approaches, and paradigms suggested in the literature related to nanomaterials and endpoints in most generalized form (Posner, 2009; Puzyn et al., 2009; Gottschalk et al., 2013; Vanli et al., 2014; Ying et al., 2015; Winkler, 2016). Works of the second type are detailed description of fresh predictive models of defined endpoints related to defined nanomaterials (Toropov and Leszczynski, 2006; Sayes and Ivanov, 2010; Liu et al., 2013; Kleandrova et al., 2014a,b; Singh and Gupta, 2014; Melagraki and Afantitis, 2014; Luan et al., 2014; Speck-Planche et al., 2015). Both mentioned types of the researches are necessary and useful. In the case of works of second type, the results should be comfortable from point of view of "potential users". This means that the results should be simple, clear, and reproducible. The absence of reliable and systematic experimental data on endpoints related to nanomaterials was and is the limitation for this research field. This circumstance leads to the paradoxical situation: the total number of works of the first type is larger than the number of works of second type.

In addition, though importance of nanomaterials for basic research, industry, and practical applications has been growing over the years their physicochemical and biochemical data has not yet been properly evaluated and collected into large databases. This causes critical complications

and challenges for building up predictive models for nanomaterials' endpoints. Traditional quantitative structure – activity relationships (QSARs) related to endpoints of "standard" substances are aimed to predict endpoint as a mathematical function of the molecular structure. Dissimilarly, the quantitative feature - activity relationships (QFARs) are based on eclectic information (Toropov et al., 2015, 2016; Toropov and Toropova, 2015). The eclectic information includes description of all available conditions and circumstances (physicochemical, biochemical, medicinal ones).

Simplified molecular input-line entry system (SMILES) are lines of symbols, which are representing the molecular structure (Toropov et al., 2015; Toropov and Toropova, 2015a,b; Toropova et al., 2015). So-called quasi-SMILES being analogies of SMILES are representation of the available eclectic information by similar lines of symbols. The CORAL software has been developed and utilized to build up QSAR models for endpoints of standard substances as a mathematical function of the molecular structure represented by SMILES. Recently the above-mentioned quasi-SMILES have been adapted for applications to nanomaterials. They can be utilized to build up models for endpoints of nanomaterials as a mathematical function of the eclectic information (Toropov et al., 2015; Toropov and Toropova, 2015a,b; Toropova et al., 2015).

The ISA-TAB-NANO has been suggested as a possible way to extract data sets to build up "nano-QSAR" (Oksel et al., 2015). However, the extraction according to principle Investigation/Study/Assay being a fundamental idea remains far from practice, whereas quasi-SMILES give possibility to build up "nano-QFAR" based on eclectic available data sets (Toropova et al., 2014; Toropov and Toropova, 2014; Toropov and Toropova, 2015a,b; Toropova et al., 2015; Toropova et al., 2016; Toropov et al., 2016). In addition the possibility of integration of small data sets into united system has been demonstrated (Toropov and Toropova, 2015a,b). In fact, the quasi-SMILES is a flexible tool to build up predictive models for results of experimental works.

Building up QFAR model for bioavailability of metal oxide nanoparticles to *E. coli* using the CORAL software is the aim of this work.

2. Method

2.1. Data

The bioavailability of metal ions influences nanotoxicity of photocatalysts (Li et al., 2012; Hwang et al., 2012). Therefore, the data on bioavailability indicates the level of toxicity. The experimental data on the bioavailability (%), adopted for this study, is taken from the literature (Dasari et al., 2013). Table 1 contains details of translation of the experimental conditions (features) into quasi-SMILES (Toropov et al., 2015).

2.2. Optimal descriptors

Optimal descriptors are calculated with quasi-SMILES as the following:

$$DCW(T^*, N^*) = \sum CW(SA_k) \quad (1)$$

where SA_k is attribute of qiasi-SMILES; the $CW(SA_k)$ represents correlation weight of SA_k. The numerical data on the $CW(SA_k)$ are calculated with the Monte Carlo method. Threshold (T) and the number of epochs (N) are parameters of the Monte Carlo optimization. The T^* and N^* are values of the above-mentioned parameters which provide preferable statistical quality for the calibration set (Toropov et al., 2015; Toropov and Toropova, 2015a,b; Toropova et al., 2015). Having the numerical data on the $CW(SA_k)$ one can calculate $DCW(T^*,N^*)$ for all quasi-SMILES. The next step involves application of the quasi-SMILES of the training set to build up bioavailability model:

$$\text{Bioavailability (\%)} = C_0 + C_1 * DCW(T^*,N^*) \quad (2)$$

After development of the model one more step is required. The model calculated with Eq.2 should be checked up with validation set (i.e. with quasi-SMILES which are not involved in building up the model).

3. Results and Discussion

The statistical characteristics of a model depend on the splitting of experimental data into three sets: training, calibration, and validation. Here three different splits were examined. The described approach based on quasi-SMILES (Table 1) gives the following models:

$$\text{Bioavailability (\%)} = -13864.0\ (\pm 1517.1) + 4617.3\ (\pm 504.9) * DCW(1,4) \quad (3)$$

$$\text{Bioavailability (\%)} = -9640.6\ (\pm 948.4) + 3218.1\ (\pm 316.3) * DCW(1,3) \quad (4)$$

$$\text{Bioavailability (\%)} = -15153.5\ (\pm 1700.9) + 5046.3\ (\pm 566.0) * DCW(1,5) \quad (5)$$

Each model is characterized by different statistical characteristics. Table 2 displays obtained statistical characteristics of these models for three splits. The details including experimental and calculated values of the bioavailability are given in the Table 3. In addition, Table 3 contains three splits of the experimental data into the training, calibration and validation sets. It should be noted, that the prevalence of attributes in the training and calibration sets is important indicator of quality of a selected split. Apparently, the frequency of features of quasi-SMILES in the training, and calibration sets should be as large as possible (Toropova et al., 2014; Toropov and Toropova, 2014;

Toropov and Toropova, 2015a,b). Of course, this is correct, also, for the validation set. Table 4 contains the numerical data on the correlation weights of attributes of quasi-SMILES calculated by the Monte Carlo technique.

One can see from the data presented in the Table 3 that the number of quasi-SMILES available for the QFAR analysis is twenty-four, i.e. it is limited. Consequently, the prevalence of features in the training and calibration sets is considerably different for examined splits. This leads to considerable difference of the predictive potential of the models. Unfortunately, this is disadvantage of models for small data sets (Toropova et al., 2014). Nevertheless, in the case of increase of available data (the total number of available quasi-SMILES), the statistical characteristics of the CORAL models becomes more stable (Toropova et al., 2015).

Hence, the suggested model has predictive potential confirmed for three random splits. Thus, on the one hand, the predictive potential of the approach based on the quasi-SMILES is confirmed; on the other hand, the model can be extended and generalized only based on feedback mechanism with the results of experiments (i.e. with increase of the number of available quasi-SMILES).

In fact, the results of experiments related to various endpoints of nanomaterials should involve ideas derived from theoretical and computational models of the endpoints and, vice versa the developers of computational models should assure that they include in their models all available eclectic details of the experimental work. Thus, the application of quasi-SMILES is one of the possible ways to organize dialog between the experimentalists and the developers of predictive models.

The possibility to build up integrated models for congeneric datasets is attractive advantage of models based on quasi-SMILES (Toropov and Toropova, 2015a,b). For example, modification of Table 1 if additional data become available is an non complex extension of “Cryptography” list.

The development of models based on quasi-SMILES obey the OECD principles for validation QSAR models (OECD, 2007).

Finally, the quasi-SMILES can be used as a tool for the practical realization of the ISA-TAB-NANO conception (Oksel et al., 2015), i.e. the standardization of available data into the format "Investigation - Study - Assay".

Conclusions

The predictive model for bioavailability of four metal oxides nanoparticles is built up using the QFAR. The CORAL software based on the Monte Carlo method was applied to develop three models for different random splits of available eclectic data represented by described quasi-SMILES (Table 1) into the training, calibration, and validation sets. One can see that representation

of experimental conditions by quasi-SMILES provides statistically robust predictive models of the investigated endpoint (Table 2). The methodological attraction of paradigm "Endpoint is a mathematical fucntion of eclectic data (conditions)" is confirmed.

Acknowledgements

Authors thank the EC project PeptiCAPS (Project reference: 686141); the project EU-ToxRisk (Project reference:681002); EFSA contract (NP//EFSA/AFSCO/2016/1) and the National Science Foundation (NSF/CREST HRD-1547754), and EPSCoR (Award #: 362492-190200-01/NSFEPS-090378) for financial support.

References

Dasari, T.P., Pathakoti, K., Hwang, H.M., 2013. Determination of the mechanism of photoinduced toxicity of selected metal oxide nanoparticles (ZnO, CuO, Co_3O_4 and TiO_2) to *E. coli* bacteria. J. Environ. Sci. 25(5), 882‑888.

Hwang, H.M., Ray, P.C., Yu, H., He, X., 2012. Toxicology of designer/engineered metallic nanoparticles. In Book: Sustainable Preparation of Metal Nanoparticles: Methods and Applications (Luque, R., Varma, R., eds). Royal Society of Chemistry, Cambridge, United Kingdom, pp.190‑212.

Gottschalk, F., Sun, T.Y., Nowack, B., 2013. Environmental concentrations of engineered nanomaterials: Review of modeling and analytical studies. Environ. Pollut. 181, 287-300.

Kleandrova, V.V., Luan, F., González-Díaz, H., Ruso, J.M., Speck-Planche, A., Cordeiro, M.N.D.S., 2014a. 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.

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

Li, Y., Zhang, W., Niu, J.F., Chen, Y.S., 2012. Mechanism of photogenerated reactive oxygen species and correlation with the antibacterial properties of engineered metal-oxide nanoparticles. ACS Nano 6(6), 5164‑ 5173.

Liu, X., Tang, K., Harper, S., Harper, B., Steevens, J.A., Xu, R., 2013. Predictive modeling of nanomaterial exposure effects in biological systems. Int. J. Nanomedicine 8(Suppl 1), 31–43.

Luan, F., Kleandrova, V. V., Gonzalez-Diaz, 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, 10623-10630.

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

OECD document, 2007. Guidance document on the guidance document on the validation of (quantitative) structure-activity relationships [(Q)SAR] models.

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

Posner, J.D., 2009. Engineered nanomaterials: Where they go,nobody knows. Nano Today 4, 114—115.

Puzyn, T., Leszczynska, D., Leszczynski, J., 2009. Toward the development of "nano-QSARs": Advances and challenges. Small 5: 2494–2509.

Sayes, C., Ivanov, I., 2010. Comparative Study of Predictive Computational Models for Nanoparticle Induced Cytotoxicity. Risk Analysis 30, 1723-1734.

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

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 (Lond) 10, 193-204.

Toropov, A.A., Leszczynski, J., 2006. A new approach to the characterization of nanomaterials: Predicting Young's modulus by correlation weighting of nanomaterials codes. Chem. Phys. Lett. 433, 125–129.

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., 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, 1837-1844.

Toropov, A.A., Toropova, A.P., 2015a. Quasi-QSAR for mutagenic potential of multi-walled carbon-nanotubes. Chemosphere 124, 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., Achary, P.G.R., Toropova, A.P., 2016. Quasi-SMILES and nano-QFPR: The predictive model for zeta potentials of metal oxide nanoparticles. Chem. Phys. Lett. 660, 107-110.

Toropova, A.P., Toropov, A.A., Benfenati, E., Puzyn, T., Leszczynska, D., 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 TiO_2 nanoparticles. Ecotoxicol. Environ. Saf. 108: 203-209.

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

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.

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.

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

Winkler, D.A., 2016. Recent advances, and unresolved issues, in the application of computational modelling to the prediction of the biological effects of nanomaterials. Toxicol. Appl. Pharmacol. 299, 96-100.

Vanli, O.A., Chen, L., Tsai, C., Zhang, C., Wang, B., 2014. An uncertainty quantification method for nanomaterial prediction models. Int. J. Adv. Manuf. Technol. 70, 33.

Ying, J., Zhang, T., Tang, M., 2015. Metal Oxide Nanomaterial QNAR Models: Available Structural Descriptors and Understanding of Toxicity Mechanisms. Nanomaterials 5, 1620-1637.

Table 1

The scheme of translation of the experimental conditions and bioavailability (%) into quasi-SMILES and bioavailability

Experimental conditions and bioavailability				**"Cryptography"**	**Quasi-SMILES vs bioavailability**				
Zn^{2+}	Light	LC_{50}	79.64		1	%11*	%20	%50	79.64
Cu^{2+}	Light	LC_{50}	75.77		2	%12	%20	%50	75.77
Co^{2+}	Light	LC_{50}	1.07		3	%13	%20	%50	1.07
Ti^{2+}	Light	LC_{50}	9.21		4	%14	%20	%50	9.21
Zn^{2+}	Light	LC_{25}	16.21		5	%11	%20	%25	16.21
Cu^{2+}	Light	LC_{25}	66.03		6	%12	%20	%25	66.03
Co^{2+}	Light	LC_{25}	1.48		7	%13	%20	%25	1.48
Ti^{2+}	Light	LC_{25}	13.95	Zn^{2+} = %11	8	%14	%20	%25	13.95
Zn^{2+}	Light	LC_{10}	21.70	Cu^{2+} = %12	9	%11	%20	%10	21.70
Cu^{2+}	Light	LC_{10}	42.39	Co^{2+} = %13	10	%12	%20	%10	42.39
Co^{2+}	Light	LC_{10}	2.55	Ti^{2+} = %14	11	%13	%20	%10	2.55
Ti^{2+}	Light	LC_{10}	31.53		12	%14	%20	%10	31.53
Zn^{2+}	Dark	LC_{50}	15.63	Light = %20	13	%11	%30	%50	15.63
Cu^{2+}	Dark	LC_{50}	9.12	Dark = %30	14	%12	%30	%50	9.12
Co^{2+}	Dark	LC_{50}	0.66		15	%13	%30	%50	0.66
Ti^{2+}	Dark	LC_{50}	2.39	LC_{50} = %50	16	%14	%30	%50	2.39
Zn^{2+}	Dark	LC_{25}	10.10	LC_{25} = %25	17	%11	%30	%25	10.10
Cu^{2+}	Dark	LC_{25}	15.01	LC_{10} = %10	18	%12	%30	%25	15.01
Co^{2+}	Dark	LC_{25}	0.71		19	%13	%30	%25	0.71
Ti^{2+}	Dark	LC_{25}	0.56		20	%14	%30	%25	0.56
Zn^{2+}	Dark	LC_{10}	9.49		21	%11	%30	%10	9.49
Cu^{2+}	Dark	LC_{10}	18.03		22	%12	%30	%10	18.03
Co^{2+}	Dark	LC_{10}	0.43		23	%13	%30	%10	0.43
Ti^{2+}	Dark	LC_{10}	0.52		24	%14	%30	%10	0.52

*) In contrast to previous works where quasi-SMILES were defined using different symbols and digits in this study quasi-SMILES are constructed using denomination of presence cycles for

281 molecules which contain ten and more cycles (Weininger, 1988; Weininger et al., 1989; Weininger,
282 1990). This gives possibility (i) to avoid wrong interpretation of symbols by the CORAL software
283 (e.g. interpretation of two conditions represented as ‘C’ and ‘L’ as one condition ‘Cl’); and (ii) use
284 of ninety identifiers for various conditions (i.e. %10, %11, …%99).
285
286

Table 2

The statistical characteristics of developed models for three splits of data into the training, calibration and validation sets

Split	**Set**	**n**	**r^2**	**RMSE**
1	Training	13	0.5740	16.5
	Calibration	5	0.7553	15.5
	validation	6	0.7587	13.9
2	Training	14	0.6287	12.6
	Calibration	5	0.5546	25.9
	validation	5	0.7481	17.9
3	Training	14	0.5384	16.3
	Calibration	5	0.8843	17.7
	validation	5	0.8967	11.9

Table 3

Experimental and calculated values of bioavailability

ID	Split1	Split2	Split3	Quasi-SMILES	Experiment	Eq. 3	Eq. 4	Eq. 5
1	T*	T	T	%11%20%50	79.64	44.9096	33.1803	40.8236
2	V	V	C	%12%20%50	75.77	57.5662	44.0076	47.9860
3	T	T	T	%13%20%50	1.07	20.4431	11.3971	12.7975
4	V	V	T	%14%20%50	9.21	28.2878	26.5923	20.5947
5	T	T	T	%11%20%25	16.21	36.0448	21.1883	38.5189
6	T	T	T	%12%20%25	66.03	48.7014	32.0156	45.6813
7	T	T	C	%13%20%25	1.48	11.5783	-0.5949	10.4928
8	C	C	V	%14%20%25	13.95	19.4230	14.6003	18.2900
9	T	T	T	%11%20%10	21.70	28.1326	32.4596	32.7313
10	C	V	V	%12%20%10	42.39	40.7893	43.2869	39.8937
11	V	T	T	%13%20%10	2.55	3.6662	10.6763	4.7052
12	C	C	C	%14%20%10	31.53	11.5108	25.8715	12.5024
13	T	T	T	%11%30%50	15.63	19.4186	13.5088	17.0777
14	T	T	T	%12%30%50	9.12	32.0753	24.3361	24.2401
15	C	C	V	%13%30%50	0.66	-5.0478	-8.2744	-10.9484
16	T	T	V	%14%30%50	2.39	2.7968	6.9208	-3.1512
17	V	V	T	%11%30%25	10.10	10.5538	1.5169	14.7730
18	V	V	T	%12%30%25	15.01	23.2105	12.3441	21.9354
19	T	T	T	%13%30%25	0.71	-13.9126	-20.2664	-13.2531
20	T	T	C	%14%30%25	0.56	-6.0679	-5.0712	-5.4559
21	T	T	C	%11%30%10	9.49	2.6417	12.7881	8.9854
22	T	T	T	%12%30%10	18.03	15.2983	23.6154	16.1478
23	C	C	V	%13%30%10	0.43	-21.8248	-8.9951	-19.0407
24	V	V	T	%14%30%10	0.52	-13.9801	6.2001	-11.2435

*) T=training set; C=calibration set; and V=validation set

Table 4

The numerical data on the correlation weights of attributes of quasi-SMILES calculated with the Monte Carlo technique

SA_k	CW(SA_k)	Prevalence of SA_k in training set	Prevalence of SA_k in calibration set	DEFECTof SA_k^*
Split 1				
%10	0.99982	3	3	0.0615
%11	1.00544	5	0	1.0000
%12	1.00818	3	1	0.0077
%13	1.00014	3	2	0.0338
%14	1.00184	2	2	0.0615
%20	1.00347	6	3	0.0154
%25	1.00154	5	1	0.0308
%30	0.99795	7	2	0.0154
%50	1.00346	5	1	0.0308
Split 2				
%10	1.00042	6	2	0.0036
%11	1.00003	3	2	0.0371
%12	1.00340	5	0	1.0000
%13	0.99326	2	2	0.0643
%14	0.99798	4	1	0.0171
%20	1.00537	6	3	0.0190
%25	0.99691	3	2	0.0371
%30	0.99926	8	2	0.0171
%50	1.00064	5	1	0.0262
Split 3				
%10	1.00337	4	2	0.0190
%11	1.00347	5	1	0.0262
%12	1.00489	4	1	0.0171
%13	0.99792	3	1	0.0036
%14	0.99946	2	2	0.0643
%20	1.00254	7	3	0.0100

%25	1.00451	5	2	0.0061
%30	0.99783	7	2	0.0111
%50	1.00497	5	1	0.0262

311

312 *) The defect of attribute of quasi-SMILES is defined as difference between probability of SAk in
313 training set and probability of SA_k in the calibration set:

314 $$DEFECT_{SA} = P_{train}(SA_k) - P_{calib}(SA_k)$$

315 If $P_{calib}(SA_k) = 0$ then $DEFECT_{SA} = 1$