Environmental Science and Pollution Research (2018) 25:31771–31775
https://doi.org/10.1007/s11356-018-3291-5

SHORT RESEARCH AND DISCUSSION ARTICLE

Use of the index of ideality of correlation to improve models of eco-toxicity

Alla P. Toropova[1] · Andrey A. Toropov[1]

Received: 26 April 2018 / Accepted: 18 September 2018 / Published online: 25 September 2018

Abstract

Persistent organic pollutants are compounds used for various everyday purposes, such as personal care products, food, pesticides, and pharmaceuticals. Decomposition of considerable part of the above pollutants is a long-time process. Under such circumstances, estimation of toxicity for large arrays of organic substances corresponding to the above category of pollutants is a necessary component of theoretical chemistry. The CORAL software is a tool to establish quantitative structure—activity relationships (QSARs). The index of ideality of correlation (*IIC*) was suggested as a criterion of predictive potential of QSAR. The statistical quality of models for eco-toxicity of organic pollutants, which are built up, with use of the *IIC* is better than statistical quality of models, which are built up without use of data on the *IIC*.

Keywords Eco-toxicity · QSAR · Index of ideality of correlation · Monte Carlo method · CORAL software

Introduction

Eco-toxicity of nonreactive organic pollutants (personal care products, food, pesticides, and pharmaceuticals) is important data for development and improvement of chemical technology (Concu et al. 2017; Castillo-Garit et al. 2016; Kleandrova et al. 2014a, b). Exposure of chemical contaminants to the aquatic environment (Baun et al. 2000; Sánchez-Bayo 2006; Parvez et al. 2008) to air (Raevsky et al. 2011) poses serious threats to the preservation of environmental quality and to human health and is recognized as a global problem (Kleandrova et al. 2014a, b; Castillo-Garit et al. 2008; Papa et al. 2005; de Morais e Silva et al. 2018). In addition, ionic liquids are important class of the organic pollutants caused by their use of everyday life (Peric et al. 2015; Ma et al. 2015). Other source of eco-toxicologic pollutants is associated with the massive use of petroleum-derived organic solvents (Perales et al. 2017). Finally, nanomaterials become additional source of eco-toxic effects (Nowack and Mitrano 2018). Thus, the development of databases together with predictive models related to eco-toxicity data for nonreactive pollutants becomes an important task of biochemistry and medicinal chemistry.

The aim of this study is estimation of the CORAL software (Toropova and Toropov 2014) as a possible tool to build up predictive models for eco-toxicity. The index of ideality of correlation (*IIC*) (Toropova and Toropov 2017; Toropov and Toropova 2017; Toropov et al. 2018; Toropov and Toropova 2018) is examined as a criterion of predictive potential of the CORAL model of eco-toxicity.

Method

Data

The experimental values measured for EC50 (effective molar concentration) (mol/L) are represented by negative decimal logarithm pEC50. The data taken in the literature (de Morais e Silva et al. 2018). These numerical data ($n = 111$) were randomly distributed into the training ($n = 28$), invisible training ($n = 27$), calibration ($n = 29$), and external validation ($n = 27$) sets. Table 1 confirms that the percentage of the identical distribution is not large.

Responsible editor: Philippe Garrigues

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11356-018-3291-5) contains supplementary material, which is available to authorized users.

✉ Alla P. Toropova
alla.toropova@marionegri.it

[1] Laboratory of Environmental Chemistry and Toxicology, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Via La Masa 19, 20156 Milan, Italy

Table 1 Percentage of identical distribution of compounds into the training, invisible training, calibration, and validation sets

	Set	Split 1	Split 2	Split 3
Split 1	Training	100	28.6	25.0
	Invisible training	100	18.5	29.6
	Calibration	100	20.7	24.1
	Validation	100	37.0	22.2
Split 2	Training		100	28.6
	Invisible training		100	40.7
	Calibration		100	27.6
	Validation		100	44.4
Split 3	Training			100
	Invisible training			100
	Calibration			100
	Validation			100

$$\text{Identify}\,(\%) = \frac{N_{i,j}}{0.5^{*}\left(N_i+N_j\right)} \times 100$$

where $N_{i,j}$ is the number of substances which are distributed into the same set for both i-th split and j-th split (set = training, invisible training, calibration, and validation); N_i is the number of substances which are distributed into the set for i-th split; N_j is the number of substances which are distributed into the set for j-th split

Optimal descriptor

The optimal descriptor (Toropova and Toropov 2014) used here is calculated as the following:

$$DCW\left(T^{*},N^{*}\right) = \sum_{k=1}^{NA} CW(S_k) + \sum_{k=1}^{NA-1} CW(SS_k) \tag{1}$$

The S_k is the "SMILES-atom," i.e., one symbol or two symbols (e.g.. "C," "N," and "O") which cannot be examined separately (e.g., "Cl" and "Si"); the SS_k is a combination of two SMILES-atoms. The $CW(S_k)$ and $CW(SS_k)$ are so-called correlation weights of the above-mentioned attributes of SMILES. The numerical data on the $CW(S_k)$ and $CW(SSS_k)$ are calculated with the Monte Carlo method, i.e., the optimization procedure which gives maximal value of a target function (*TF*).

QSAR models, calculated with the Monte Carlo optimization of target functions TF_1 and TF_2:

$$TF_1 = r_{TRN} + r_{iTRN} - |r_{TRN} - r_{iTRN}| * 0.1 \tag{2}$$

$$TF_2 = TF_3 + IIC_{CLB} * 0.1 \tag{3}$$

The r_{TRN} and r_{iTRN} are correlation coefficient between observed and predicted endpoint for the training and invisible training sets, respectively.

The IIC_{CLB} is calculated with data on the calibration (*CLB*) set as the following:

$$IIC_{CLB} = r_{CLB}\frac{\min\left({}^{-}MAE_{CLB}, {}^{+}MAE_{CLB}\right)}{max\left({}^{-}MAE_{CLB}, {}^{+}MAE_{CLB}\right)} \tag{4}$$

$${}^{-}M\,AE_{CLB} = \frac{1}{{}^{-}N}\sum_{k=1}^{{}^{-}N} |\,\Delta_k|, \Delta_k < 0; {}^{-}N \tag{5}$$

is the number of $\Delta_k < 0$

$${}^{+}M\,AE_{CLB} = \frac{1}{{}^{+}N}\sum_{k=1}^{{}^{+}N} |\,\Delta_k|, \Delta_k \geqslant 0; {}^{+}N \text{ is the number of } \Delta_k \geqslant 0 \tag{6}$$

$$\Delta_k = \text{observed}_k - \text{calculated}_k \tag{7}$$

The observed and calculated are corresponding values of pEC50.

Having the numerical data on the $CW(S_k)$ and $CW(SS_k)$, the predictive model is calculated by the least squares method with compounds from the training set:

$$pEC_{50} = C_0 + C_1 * DCW\left(T^{*},N^{*}\right) \tag{8}$$

Results and discussion

Three models for pEC50 are built up using three random splits with two versions of target function TF_1 calculated with Eq. 2 and TF_2 calculated with Eq. 3.

In the case of TF_1 these models are the following:

$$\text{pEC50} = 1.732(\pm 0.027) + 0.3695(\pm 0.0047) * \text{DCW}(1,2) \tag{9}$$

$$\text{pEC50} = 1.842(\pm 0.042) + 0.3694(\pm 0.0063) * \text{DCW}(1,6) \tag{10}$$

$$\text{pEC50} = 1.784(\pm 0.023) + 0.4488(\pm 0.0046) * \text{DCW}(1,2) \tag{11}$$

In the case of TF_2, these models are the following:

$$\text{pEC50} = 1.582(\pm 0.048) + 0.3745(\pm 0.0069) * \text{DCW}(1,15) \tag{12}$$

$$\text{pEC50} = 1.366(\pm 0.054) + 0.2766(\pm 0.0052) * \text{DCW}(1,15) \tag{13}$$

$$\text{pEC50} = 2.009(\pm 0.036) + 0.4891(\pm 0.0091) * \text{DCW}(1,15) \tag{14}$$

Table 2 contains the statistical characteristics of the models calculated with Eqs. 3–5. Comparison of these

Table 2 The statistical characteristics of models for eco-toxicity

Split	*TF*	Set	*n*	r^2	RMSE	*CCC*[a]	$<R_m^2>$[b]	*IIC*
1	TF_1	Training	28	0.8921	0.291			
		Invisible training	27	0.8699	0.378			
		Calibration	29	0.7248	0.446	0.8343	0.5840	0.4738
		Validation	27	0.9062	0.267			
	TF_2	Training	28	0.7877	0.409			
		Invisible training	27	0.8157	0.420			
		Calibration	29	0.8162	0.345	*0.8937*	*0.7068*	*0.9028*
		Validation	27	0.9515	0.223			
2	TF_1	Training	28	0.8431	0.326			
		Invisible training	27	0.8166	0.424			
		Calibration	29	0.8878	0.295	0.9417	0.8376	0.6284
		Validation	27	0.8556	0.322			
	TF_2	Training	28	0.8633	0.304			
		Invisible training	27	0.7251	0.476			
		Calibration	29	0.8718	0.315	0.9330	0.8152	*0.9325*
		Validation	27	0.9224	0.228			
3	TF_1	Training	28	0.9062	0.262			
		Invisible training	27	0.9060	0.297			
		Calibration	29	0.6890	0.454	0.8080	0.5310	0.6061
		Validation	27	0.8454	0.368			
	TF_2	Training	28	0.8346	0.348			
		Invisible training	27	0.8433	0.407			
		Calibration	29	0.8312	0.283	*0.9078*	*0.7584*	*0.9113*
		Validation	27	0.9335	0.225			

[a] The *CCC* is concordance correlation coefficient (I-Kuei Lin 1989); [b] $<R_m^2>$ is Rm2 metric (Roy et al. 2009; Ojha et al. 2011)

Model suggested in the literature (de Morais e Silva et al. 2018) has the following statistical quality n=86, r^2 =0.8221, RMSE=0.353 (training set) and n=25, r^2 =0.8981, RMSE=0.299 (validation set)

models with model from the literature (de Morais e Silva et al. 2018) shows that the CORAL-models are better for the external validation set.

Figure 1 contains comparison of co-evolutions of correlations between observed and calculated pEC50 for training, invisible training, and calibration sets. The absence of overtraining is the main difference between the optimization with TF_2 and optimization with TF_1. Factually, this is an advantage of the optimization with TF_2.

Concordance correlation coefficient (*CCC*) (I-Kuei Lin 1989) and average $<R_m^2>$ (Roy et al. 2009; Ojha et al. 2011) are widely used criteria of predictive potential of a QSAR model. In other words, if there are model-1 and model-2 and *CCC*-1 is larger than *CCC*-2, then the model-1 should has better predictive potential for external compounds. Analogically, if there are model-1 and model-2 and R_m^2-1 is larger than R_m^2-2, then the model-1 should has better predictive potential for external compounds. The same principle is related to *IIC*: larger value of *IIC* should be observed for model with better predictive potential. The *CCC* and $<R_m^2>$ give correct recommendation for pair of models built up with TF_1 and TF_2 for split #1 and #3, but for split #2 these criteria give wrong recommendation (Table 2). The *IIC* gives correct recommendations for all splits #1, #2, and #3. Thus, *CCC* (I-Kuei Lin 1989), $<R_m^2>$ (Roy et al. 2009; Ojha et al. 2011) and *IIC* (Toropova and Toropov 2017; Toropov and Toropova 2017; Toropov et al. 2018; Toropov and Toropova 2018) are different criteria of predictive potential.

Supplementary materials contain confirmation of the compliances of the CORAL approach to OECD principles: Table S1 contains definition of the domain of applicability; Table S2 contains mechanistic interpretation of the CORAL model in terms of SMILES-attributes, which are promoters of increase or decrease for pEC50. Table S3 contains observed and calculated pEC50 together with distribution into the training, invisible training, calibration, and validation sets.

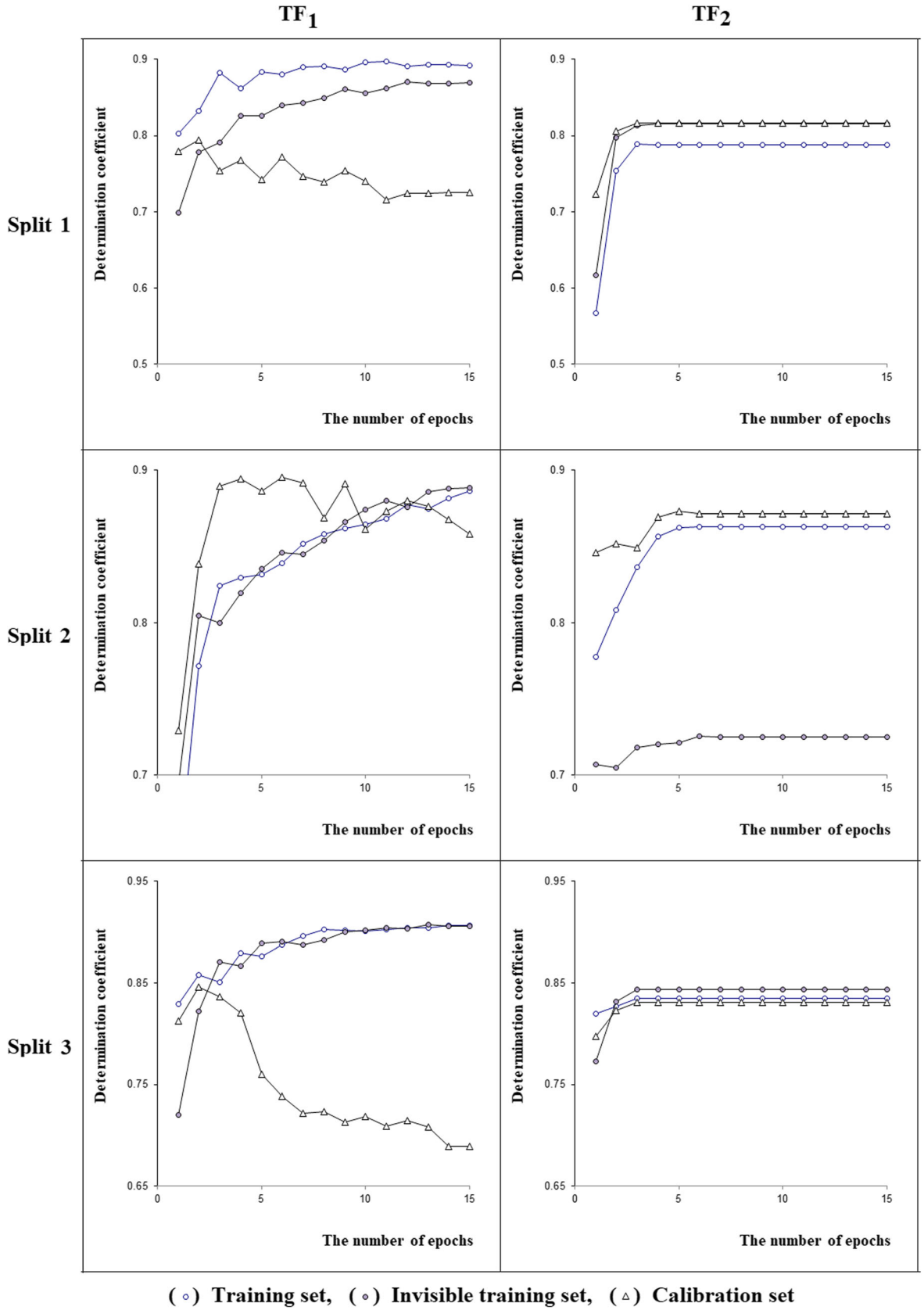

Fig. 1 Co-evolution of correlations between $pEC50_{observed}$ and $pEC50_{calculated}$ for training (white circle), invisible training (dark circle), and calibration (white triangle) sets with applying target function TF_1 (Eq. 2) and TF_2 (Eq. 3)

Conclusions

The CORAL software factually is a tool to build up predictive models for eco-toxicity of compounds examined here. The target function TF_2 gives models with better predictive potential in comparison with models based on the Monte Carlo optimization with TF_1. In other words, the *IIC* is checked up with three random splits. Hence, the *IIC* can be a useful criterion of the predictive potential of QSAR models of eco-toxicity.

Author contributions Authors have done equivalent contributions to this work.

Funding information This research was supported by the LIFE-CONCERT project (LIFE17 GIE/IT/000461).

References

Baun A, Jensen SD, Bjerg PL, Christensen TH, Nyholm N (2000) Toxicity of organic chemical pollution in groundwater downgradient of a Landfill (Grindsted, Denmark). Environ Sci Technol 34(9): 1647–1652. https://doi.org/10.1021/es9902524

Castillo-Garit JA, Marrero-Ponce Y, Escobar J, Torrens F, Rotondo R (2008) A novel approach to predict aquatic toxicity from molecular structure. Chemosphere 73(3):415–427. https://doi.org/10.1016/j.chemosphere.2008.05.024

Castillo-Garit JA, Abad C, Casañola-Martin GM, Barigye SJ, Torrens F, Torreblanca A (2016) Prediction of aquatic toxicity of benzene derivatives to tetrahymena pyriformis according to OECD principles. Curr Pharm Des 22(33):5085–5094. https://doi.org/10.2174/1381612822666160804095107

Concu R, Kleandrova VV, Speck-Planche A, Cordeiro MNDS (2017) Probing the toxicity of nanoparticles: a unified in silico machine learning model based on perturbation theory. Nanotoxicology 11(7):891–906. https://doi.org/10.1080/17435390.2017.1379567

de Morais e Silva L, Alves MF, Scotti L, Lopes WS, Scotti MT (2018) Predictive ecotoxicity of MoA 1 of organic chemicals using in silico approaches. Ecotoxicol Environ Saf 153:151–159. https://doi.org/10.1016/j.ecoenv.2018.01.054

I-Kuei Lin L (1989) A concordance correlation coefficient to evaluate reproducibility. Biometrics 45(1):255–268. https://doi.org/10.2307/2532051

Kleandrova VV, Luan F, González-Díaz H, Ruso JM, Melo A, Speck-Planche A, Cordeiro MNDS (2014a) Computational ecotoxicology: simultaneous prediction of ecotoxic effects of nanoparticles under different experimental conditions. Environ Int 73:288–294. https://doi.org/10.1016/j.envint.2014.08.009

Kleandrova VV, Luan F, González-Díaz H, Ruso JM, Speck-Planche A, Cordeiro MNDS (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. https://doi.org/10.1021/es503861x

Ma S, Lv M, Deng F, Zhang X, Zhai H, Lv W (2015) Predicting the ecotoxicity of ionic liquids towards Vibrio fischeri using genetic function approximation and least squares support vector machine. J Hazard Mater 283:591–598. https://doi.org/10.1016/j.jhazmat.2014.10.011

Nowack B, Mitrano DM (2018) Procedures for the production and use of synthetically aged and product released nanomaterials for further environmental and ecotoxicity testing. NanoImpact 10:70–80. https://doi.org/10.1016/j.impact.2017.12.001

Ojha PK, Mitra I, Das RN, Roy K (2011) Further exploring R_m^2 metrics for validation of QSPR models. Chemom Intell Lab Syst 107(1): 194–205. https://doi.org/10.1016/j.chemolab.2011.03.011

Papa E, Battaini F, Gramatica P (2005) Ranking of aquatic toxicity of esters modelled by QSAR. Chemosphere 58(5):559–570. https://doi.org/10.1016/j.chemosphere.2004.08.003

Parvez S, Venkataraman C, Mukherji S (2008) Toxicity assessment of organic pollutants: reliability of bioluminescence inhibition assay and univariate QSAR models using freshly prepared Vibrio fischeri. Toxicol in Vitro 22(7):1806–1813. https://doi.org/10.1016/j.tiv.2008.07.011

Perales E, García JI, Pires E, Aldea L, Lomba L, Giner B (2017) Ecotoxicity and QSAR studies of glycerol ethers in Daphnia magna. Chemosphere 183:277–285. https://doi.org/10.1016/j.chemosphere.2017.05.107

Peric B, Sierra J, Martí E, Cruañas R, Garau MA (2015) Quantitative structure-activity relationship (QSAR) prediction of (eco)toxicity of short aliphatic protic ionic liquids. Ecotoxicol Environ Saf 115: 257–262. https://doi.org/10.1016/j.ecoenv.2015.02.027

Raevsky OA, Modina EA, Raevskaya OE (2011) QSAR models of the inhalation toxicity of organic compounds. Pharm Chem J 45(3): 165–169. https://doi.org/10.1007/s11094-011-0585-z

Roy PP, Paul S, Mitra I, Roy K (2009) On two novel parameters for validation of predictive QSAR models. Molecules 14(5):1660–1701. https://doi.org/10.3390/molecules14051660

Sánchez-Bayo F (2006) Comparative acute toxicity of organic pollutants and reference values for crustaceans. I. Branchiopoda, Copepoda and Ostracoda. Environ Pollut 139(3):385–420. https://doi.org/10.1016/j.envpol.2005.06.016

Toropov AA, Toropova AP (2017) The index of ideality of correlation: a criterion of predictive potential of QSPR/QSAR models? Mutat Res Genet Toxicol Environ Mutagen 819:31–37. https://doi.org/10.1016/j.mrgentox.2017.05.008

Toropov AA, Toropova AP (2018) Application of the Monte Carlo method for building up models for octanol-water partition coefficient of platinum complexes. Chem Phys Lett 701:137–146. https://doi.org/10.1016/j.cplett.2018.04.012

Toropov AA, Carbó-Dorca R, Toropova AP (2018) Index of ideality of correlation: new possibilities to validate QSAR: a case study. Struct Chem 29(1):33–38. https://doi.org/10.1007/s11224-017-0997-9

Toropova AP, Toropov AA (2014) CORAL software: prediction of carcinogenicity of drugs by means of the Monte Carlo method. Eur J Pharm Sci 52(1):21–25. https://doi.org/10.1016/j.ejps.2013.10.005

Toropova AP, Toropov AA (2017) The index of ideality of correlation: a criterion of predictability of QSAR models for skin permeability? Sci Total Environ 586:466–472. https://doi.org/10.1016/j.scitotenv.2017.01.198