



NJC

The index of ideality of correlation: models of flash points of ternary mixtures

Journal:	<i>New Journal of Chemistry</i>
Manuscript ID	NJ-ART-01-2020-000121.R1
Article Type:	Paper
Date Submitted by the Author:	14-Feb-2020
Complete List of Authors:	Toropova, Alla; Istituto di Ricerche Farmacologiche Mario Negri, Toropov, Andrey; Istituto di Ricerche Farmacologiche Mario Negri, Leszczynska, Danuta; Jackson State University, bCivil and Environmental Engineering Department Leszczynski, Jerzy; Jackson State University, Department of Chemistry

SCHOLARONE™
Manuscripts

The index of ideality of correlation: models of flash points of ternary mixtures

Alla P. Toropova^{*,1}, Andrey A. Toropov¹, Danuta Leszczynska², Jerzy Leszczynski³

¹*Laboratory of Environmental Chemistry and Toxicology, Department of Environmental Health Science, Istituto di Ricerche Farmacologiche Mario Negri IRCCS,*

Via La Masa 19, 20156 Milano, 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, Physics and Atmospheric Sciences*

Jackson State University, 1400 J. R. Lynch Street, P.O. Box 17910, Jackson,

MS 39217, USA

Abstract

Reliable information related to flash point of ternary mixtures assists in rational classification of different ternary mixtures of liquids. Hence, dependable computational models for predictions of the above endpoint can be useful. Simplified molecular input-line entry system (SMILES) is the representation of the molecular structure. Quasi-SMILES is the expansion of traditional SMILES by means of additional symbols that reflect "eclectic", which are able to influence physicochemical behaviour of substances. The application of the quasi-SMILES to build up model for flammability of ternary liquid mixtures has indicated that the approach provides very good model for the flash points of ternary mixtures of organic substances. The Index of Ideality of Correlation (*IIC*) is a criterion of predictive potential of QSPR/QSAR models. The attempts of applying of the *IIC* to improve models for flammability of ternary liquid mixtures confirm applicability of this criterion to improve predictive potential of the above models.

Keywords: Flash Point; Ternary Mixture; QSPR; *In silico* modelling; Monte Carlo method

^{*}) Corresponding author

Alla P. Toropova

Laboratory of Environmental Chemistry and Toxicology,

Istituto di Ricerche Farmacologiche Mario Negri IRCCS

Via La Masa 19, 20156 Milano, Italy

Tel: +39 02 3901 4595

Fax: +39 02 3901 4735

Email: alla.toropova@marionegri.it

INTRODUCTION

Flammable substances, such as organic solvents, are used in laboratories and industrial processes. The flash point (FP) is one of the most important parameters used to characterize the ignition hazards of these liquids.¹⁻³ Obviously, the evaluation of a risk assessment via mathematical equations for mixtures is more complex task in comparison with the risk assessment for pure substances. Development of computational models able to predict flash points of ternary mixtures is an attractive alternative to experimental definition of this endpoint.⁴

The CORAL software represents an efficient tool to build up quantitative structure – property / activity relationships (QSPRs/QSARs) for various endpoints.⁵ The input for the software is simplified molecular input-line entry system (SMILES).⁶ SMILES can be expanded by additional symbols, which represent various conditions by means of special quasi-SMILES.⁷⁻⁹ Quasi-SMILES can use as a tool to represent binary mixtures.^{10,11} Following the previous studies in the current work this approach is applied for ternary mixtures.

The Index of Ideality of correlation (*IIC*) has been assessed as a tool to improve predictive potential of QSPR/QSAR for various endpoints.¹²⁻¹⁷ The basic advantage of the *IIC* is sensitivity (i) to the correlation; and (ii) to the dispersion.

The aim of the present study is assessment of the CORAL models as a tool to build up predictive model for FP of ternary mixtures of organic liquids based on quasi-SMILES. In addition, the comparison of models calculated with and without the *IIC* is another objective of the study.

METHOD

Data

The experimental dataset contains 808 flash points FP (°C) of ternary mixtures of organic liquids (hydrocarbons, alcohols, ketones, esters, pyridines, and acids) with different molar fractions. These data are extracted from the literature.⁴ All experimental flash points, examined here, were measured using standardized protocols ASTM D56, ASTM D93/ISO 2719 or ASTM D3828/ISO 3679.⁴ Quasi-SMILES were used to represent these ternary mixtures as objects for the modelling. In contrast to traditional SMILES⁶ the quasi-SMILES contains special symbols to take into account conditions (circumstances), besides the molecular structure.¹⁸

Building up Quasi-SMILES

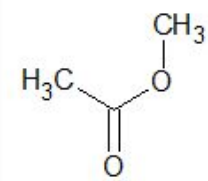
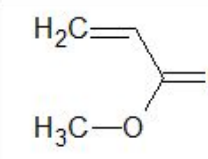
Quasi-SMILES are defined as strings that involve the following elements:

- (i) ID (numbering);
- (ii) SMILES-1 that represents first component of binary mixture;
- (iii) Symbols, which represent molar fraction of the first component;

- (iv) SMILES-2 that represents second component;
 (v) Symbols which represent molar fraction of second component;
 (vi) SMILES-3 that represents third component;
 (vii) Symbols which represent molar fraction of third component; and
 (viii) FP ($^{\circ}\text{C}$).

The first symbol of each line represents kind of set: training (+), invisible training (-), calibration (#), and validation (*). **Table 1** contains scale used to encode the molar fractions.

Figure 1 contains the general scheme of building up quasi-SMILES.

Structure-1	$\text{H}_3\text{C}-\text{OH}$	SMILES-1	CO
Structure-2		SMILES-2	CC (=O) OC
Structure-3		SMILES-3	C=CC (=O) OC

Molar fractions:

	x1	x2	x3		
...	↓	↓	↓	...	
+10.	CO%25	.CC (=O) OC%11	.C=CC (=O) OC%45	-3.000	Flash points
-11.	CO%21	.CC (=O) OC%11	.C=CC (=O) OC%50	-3.000	
#12.	CO%15	.CC (=O) OC%11	.C=CC (=O) OC%55	-2.500	
*13.	CO%11	.CC (=O) OC%11	.C=CC (=O) OC%60	-1.900	
#14.	CO%55	.CC (=O) OC%15	.C=CC (=O) OC%11	0.000	
-15.	CO%50	.CC (=O) OC%15	.C=CC (=O) OC%15	-2.500	
...	↑	↑	↑	...	
	SMILES-1	SMILES-2	SMILES-3		

Figure 1. The general scheme of building up quasi-SMILES

Table 1. Discretization of different molar fractions into the codes for quasi-SMILES

	R A N G E		Code for quasi-SMILES
	Min(x)	Max(x)	
1	0.000	0.020	%11
2	0.020	0.040	%12
3	0.040	0.060	%13
4	0.060	0.080	%14
5	0.080	0.100	%15
6	0.100	0.120	%16
7	0.120	0.140	%17
8	0.140	0.160	%18
9	0.160	0.180	%19
10	0.180	0.200	%19
11	0.200	0.220	%19
12	0.220	0.240	%19
13	0.240	0.260	%19
14	0.260	0.280	%19
15	0.280	0.300	%25
16	0.300	0.320	%26
17	0.320	0.340	%27
18	0.340	0.360	%28
19	0.360	0.380	%29
20	0.380	0.400	%30
21	0.400	0.420	%31
22	0.420	0.440	%32
23	0.440	0.460	%33
24	0.460	0.480	%34
25	0.480	0.500	%35
26	0.500	0.520	%36
27	0.520	0.540	%37
28	0.540	0.560	%38
29	0.560	0.580	%39
30	0.580	0.600	%40
31	0.600	0.620	%41
32	0.620	0.640	%42
33	0.640	0.660	%43
34	0.660	0.680	%44
35	0.680	0.700	%45
36	0.700	0.720	%46
37	0.720	0.740	%47
38	0.740	0.760	%48
39	0.760	0.780	%49
40	0.780	0.800	%50
41	0.800	0.820	%51
42	0.820	0.840	%52
43	0.840	0.860	%53

44	0.860	0.880	%54
45	0.880	0.900	%55
46	0.900	0.920	%56
47	0.920	0.940	%57
48	0.940	0.960	%58
49	0.960	0.980	%59
50	0.980	1.000	%60

Distribution of quasi-SMILES into the training and validation sets

According to the principle “QSPR is a random event”¹⁹ if a random split of the data into the training and validation sets is used to build up QSPR/QSAR then the careful checking up of the predictive potential of an approach should be based on consideration of a group of random splits. Therefore, the above-mentioned quasi-SMILES were distributed randomly, three times, into the active training set ($\approx 25\%$), the passive training set ($\approx 25\%$), the calibration set ($\approx 25\%$), and the validation set ($\approx 25\%$). **Table 2** indicates that three random splits examined in this work are not identical.

Table 2. The measure (%) of identity of splits into the active training set, the passive training set, the calibration, and the validation sets examined here

split	Set	Split 1	Split 2	Split 3
1	Active training	100*	35.1	42.6
	Passive training	100	35.0	40.0
	Calibration	100	32.0	30.6
	Validation	100	37.9	38.0
2	Active training		100	34.8
	Passive training		100	37.2
	Calibration		100	38.8
	Validation		100	40.3
3	Active training			100
	Passive training			100
	Calibration			100
	Validation			100

$$Identity(\%) = \frac{N_{i,j}}{0.5 * (N_i + N_j)} * 100$$

$N_{i,j}$ is the number of quasi-SMILES which are distributed into the same set for both i-th split and j-th split (i.e. set can be training, invisible training, calibration, or validation);

N_i is the number of quasi-SMILES which are distributed into the set for i-th split;

N_j is the number of quasi-SMILES which are distributed into the set for j-th split.

Each set has special task. The active training set is the “builder” of model, correlation weights extracted from quasi-SMILES distributed in this set are optimizing via the Monte Carlo

technique. The passive training set is the “inspector” of the model. Quasi-SMILES from this set are used to check up “whether the model is suitable for quasi-SMILES which are out of the active training set”. The calibration set is aimed to detect start of the overfitting: situation where excellent statistical quality for the training set is accompanied by poor statistical quality for the calibration set. The task of the validation set is final estimation of the model. Thus, each set has important task, consequently, equivalent distributions probably are quite reason strategy.^{10,12,15-18}

Optimal descriptor

The optimal descriptor used to build up model for flash points (⁰C) is defined as:

$$DCW(T^*, Nepoch^*) = CW(HARD) + \sum_{k=1}^{NA} CW(S_k) + \sum_{k=1}^{NA-1} CW(SS_k) + \sum_{k=1}^{NA-2} CW(SSS_k) \quad (1)$$

The S_k is so-called “quasi-SMILES-atom” i.e. one symbol (e.g. ‘C’, ‘N’, ‘O’, etc.) or group of symbols which cannot be examined separately (e.g. ‘Cl’, ‘Si’, %11, %35, etc.); the SS_k is a combination of two quasi-SMILES-atoms; the SSS_k is a combination of three quasi-SMILES-atoms the HARD is special SMILES attribute as described in the literature.²⁰ This is special code that characterized corresponding molecular system according to presence (absence) chemical elements, i.e. nitrogen, oxygen, sulphur, phosphorus, fluorine, chlorine, bromine, iodine, as well as presence of double, triple, and 3D (stereochemical) bonds.

The $CW(S_k)$, $CW(SS_k)$, $CW(SSS_k)$, and $CW(HARD)$ are so-called correlation weights of the above-mentioned attributes of SMILES.

The numerical data on the $CW(S_k)$, $CW(SS_k)$, $CW(SSS_k)$, and $CW(HARD)$ are calculated with the Monte Carlo method, i.e. via the optimization procedure which gives maximal value of a target function (TF). The NA is the number of attributes in SMILES.

The T is threshold i.e. integer to distribute all quasi-SMILES attributes into two categories (i) non-rare, if frequency of the attribute in the training set is larger than T ; and (ii) rare, otherwise. Correlation weights of all rare attributes are equal to zero (i.e. these are not involved in building up model). The $Nepoch$ is the number of epoch of the optimization. One epoch is sequence of modifications of all quasi-SMILES attributes involved in building up model. The $T=T^*$ and $Nepoch=Nepoch^*$ are values of the parameters which are preferable for statistical quality of the model for the calibration set.

Two kinds of the QSPR-models are studied: (i) models calculated with the Monte Carlo optimization based on target functions TF_1 and (ii) models calculated with the Monte Carlo optimization based on target functions TF_2 :^{12, 15, 21-24}

$$TF_1 = r_{TRN} + r_{iTRN} - |r_{TRN} - r_{iTRN}| * 0.1 \quad (2)$$

$$TF_2 = TF_1 + IIC_{CLB} * 0.1 \quad (3)$$

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:

$$IIC_{CLB} = r_{CLB} \frac{\min(-MAE_{CLB}, +MAE_{CLB})}{\max(-MAE_{CLB}, +MAE_{CLB})} \quad (4)$$

where

$$-MAE_{CLB} = \frac{1}{-N} \sum_{k=1}^{-N} |\Delta_k|, \quad \Delta_k < 0; -N \text{ is the number of } \Delta_k < 0 \quad (5)$$

$$+MAE_{CLB} = \frac{1}{+N} \sum_{k=1}^{+N} |\Delta_k|, \quad \Delta_k \geq 0; +N \text{ is the number of } \Delta_k \geq 0 \quad (6)$$

$$\Delta_k = observed_k - calculated_k \quad (7)$$

The “observed” and “calculated” are corresponding values of the endpoint.

Having the numerical data on the above-mentioned correlation weights the predictive model is calculated using quasi-SMILES of the training set:

$$FP(^0C) = C_0 + C_1 * DCW(T^*, N^*) \quad (8)$$

The predictive potential of the model should be checked up with the validation set. ^{12, 15, 21-24}

Domain of applicability

Domain of applicability of the model is defined according to distribution of quasi-SMILES attributes in the training and calibration sets as two-step:²⁰

Step 1: the definition of statistical defect (d_k) for each quasi-SMILES attribute involved (non- blocked) in building up a model:

$$d_k = \frac{|P(A_k) - P'(A_k)|}{N(A_k) + N(A_k)} \quad (9)$$

where $P(A_k)$ and $P'(A_k)$ are probability of attribute of quasi-SMILES A_k in the training and calibration sets, respectively; $N(A_k)$ and $N'(A_k)$ are frequencies of A_k in the training and calibration sets, respectively.

Step 2: the calculation for all substances of the statistical quasi-SMILES-defect (D_j):

$$D_j = \sum_{k=1}^{N_{act}} d_k \quad (10)$$

where N_{act} is the number of non-blocked SMILES attributes in the quasi-SMILES.

A substance falls in the domain of applicability if

$$D_j < 2 * \bar{D} \quad (11)$$

where \bar{D} is average of the statistical quasi-SMILES-defect for the training set.

RESULTS AND DISCUSSION

The essence of approach applied here is optimization of the above correlation weights of all fragments of the quasi-SMILES. One epoch of the optimization is variations of all correlation weights involved in the modelling process. Two versions of the Monte Carlo optimization are studied here. These optimization were carried out for 35 epochs with two different target functions TF_1 (Eq. 2) and TF_2 (Eq. 3).

There is significant difference between the results of the Monte Carlo optimization with the target function TF_1 and the optimization with the target function TF_2 . The comparison of “typical histories” of these Monte Carlo optimizations expressed via evolutions of correlation of the optimal descriptor and the endpoint for the active training set, the passive training set, the calibration set, and the validation set from epoch-1 till epoch-35 confirms that TF_2 function generates better model. Figure 1, 2, and 3 contain the graphical representations of these histories for split-1, split-2, and split-3, respectively.

Factually, the Index of Ideality of Correlation (IIC) (i.e. application of TF_2) results in decrease of correlation coefficients for the training and invisible training sets that is accompanied by the increase of correlation coefficient for calibration set.^{12, 15, 21-24} The increase of the correlation coefficient for the validation set as the rule is accompanied by improvement of statistical situation for the external validation set (**Table 2, Figure 2-4**).

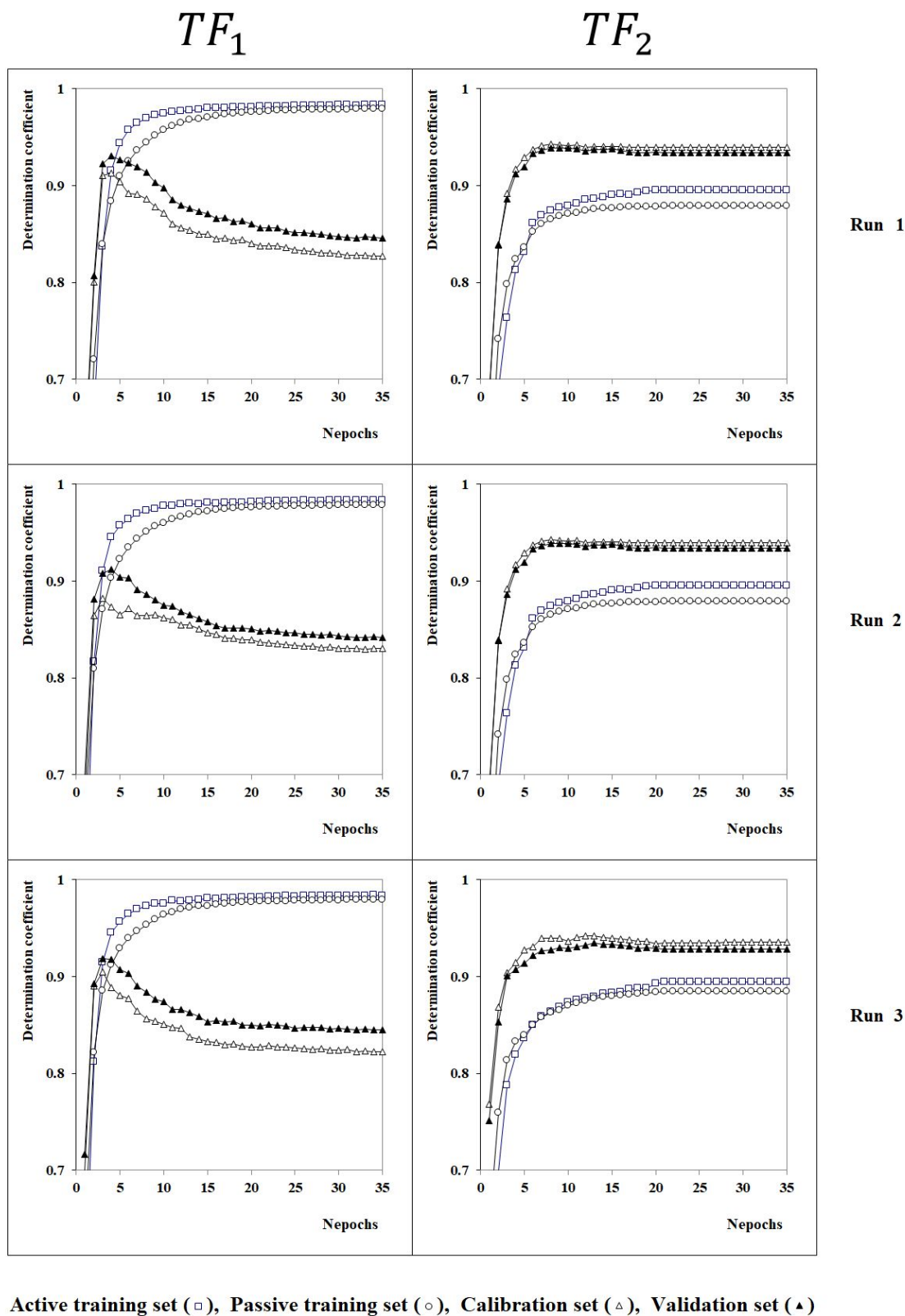


Figure 2. Comparison of evolution of correlations for the active training set, the passive training set, the calibration set, and the validation set in process of the Monte Carlo optimization with different target functions (TF_1 and TF_2) for the case of split 1

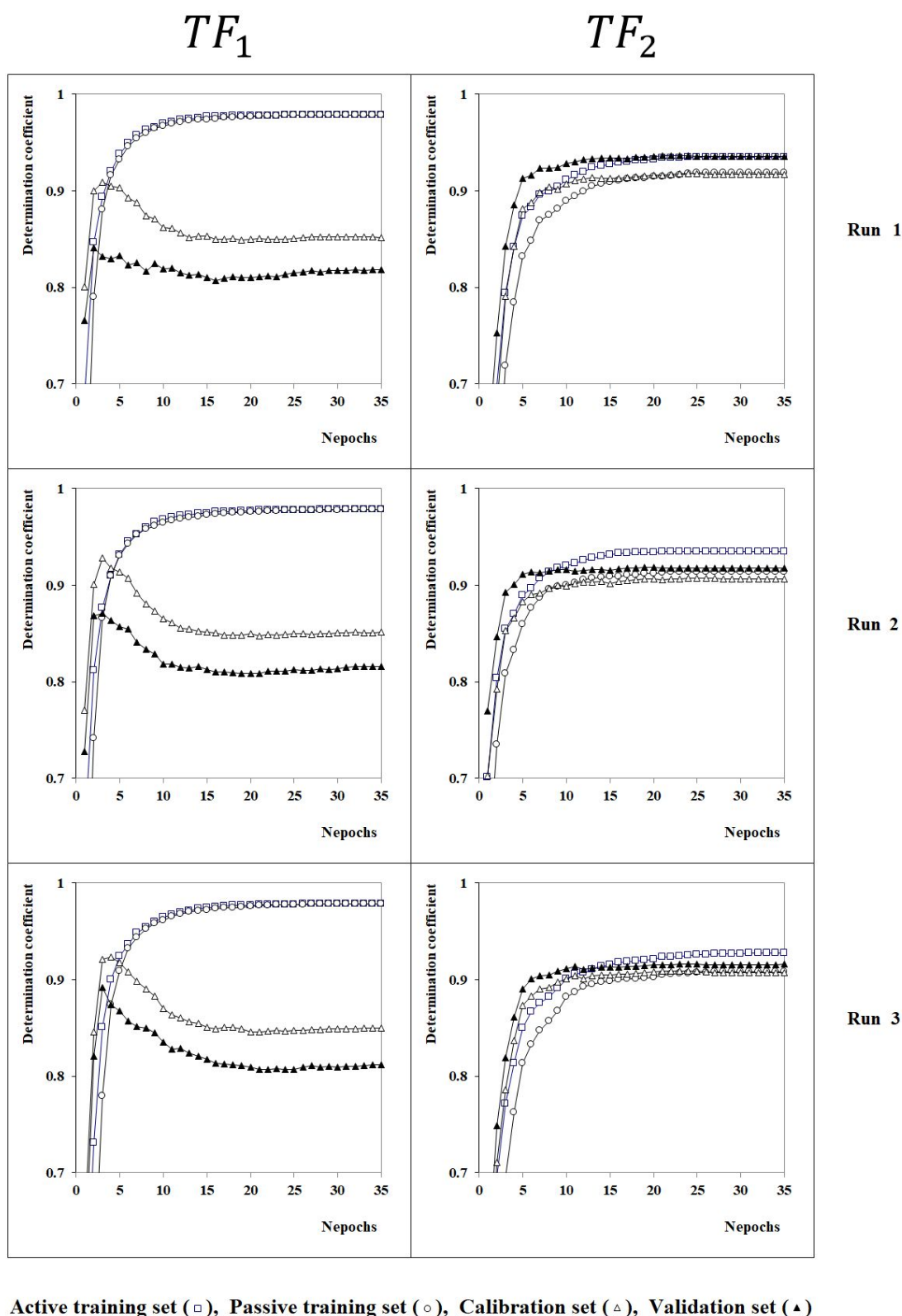


Figure 3. Comparison of evolution of correlations for the active training set, the passive training set, the calibration set, and the validation set in process of the Monte Carlo optimization with different target functions (TF_1 and TF_2) for the case of split 2

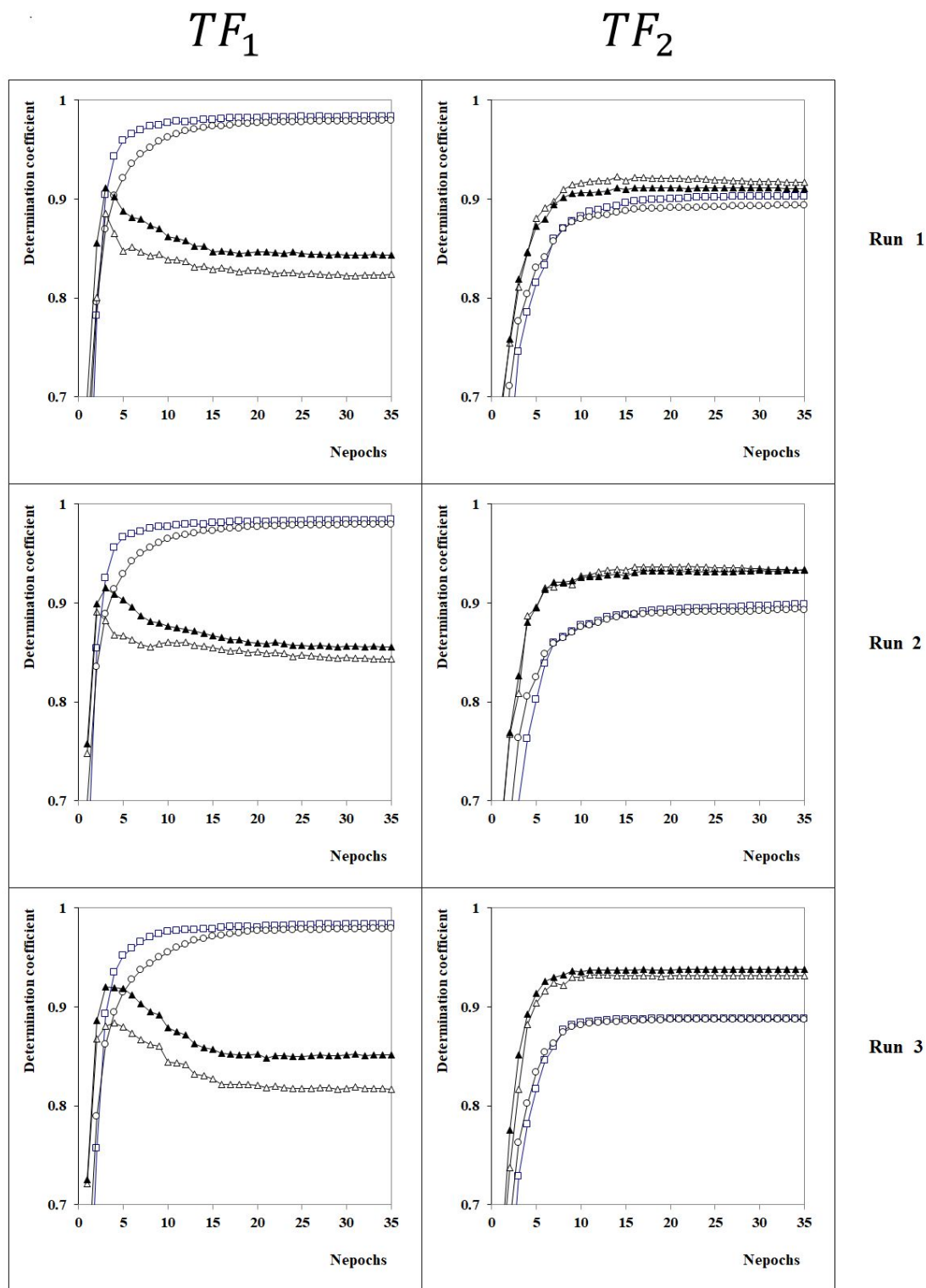


Figure 4. Comparison of evolution of correlations for the active training set, the passive training set, the calibration set, and the validation set in process of the Monte Carlo optimization with different target functions (TF_1 and TF_2) for the case of split 3

The CORAL models for FP obtained with TF_1 are the following:

$$FP(^{\circ}C) = -96.74(\pm 0.22) + 6.321(\pm 0.015) * DCW(1,3) \quad (12)$$

$$FP(^{\circ}C) = -70.84(\pm 0.20) + 4.544(\pm 0.012) * DCW(1,2) \quad (13)$$

$$FP(^{\circ}C) = -70.84(\pm 0.23) + 3.451(\pm 0.010) * DCW(1,2) \quad (14)$$

The CORAL models for FP obtained with TF_2 are the following:

$$FP(^{\circ}C) = -67.63(\pm 0.15) + 5.322(\pm 0.011) * DCW(1,15) \quad (17)$$

$$FP(^{\circ}C) = -75.59(\pm 0.18) + 5.040(\pm 0.011) * DCW(1,15) \quad (18)$$

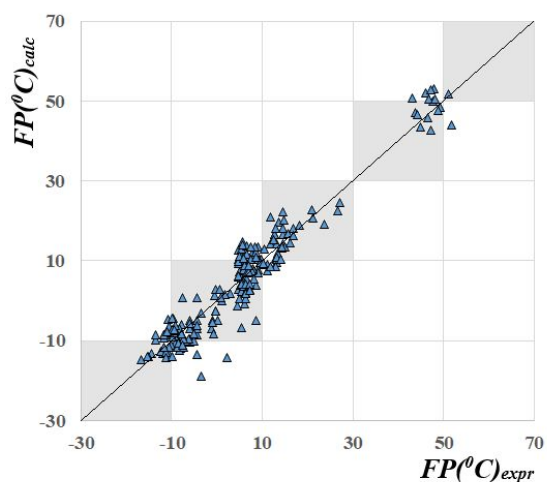
$$FP(^{\circ}C) = -82.88(\pm 0.22) + 6.179(\pm 0.016) * DCW(1,15) \quad (19)$$

Table 3 contains the statistical characteristics of these models. Data represented in Table 3 indicates that the Monte Carlo optimization with the TF_2 generates better model in comparison with models obtained using the TF_1 , if the statistical quality of the model for the validation set is applied as the basis for comparison of these models. Factually, taking into account *IIC* modify the Monte Carlo optimization. Moreover, a paradoxical situation takes place. There is an improvement of the statistical quality for the calibration and validation sets, but it is accompanied by the detriment of the active training set and passive training set. Does this situation is advantageous from practical point of view? We believe that rather yes, than no. Computational experiments recently described in the literature¹²⁻¹⁷ confirm this hypothesis. **Figure 5** contains graphical representations of models calculated with Eqs. 17-19, for splits 1-3.

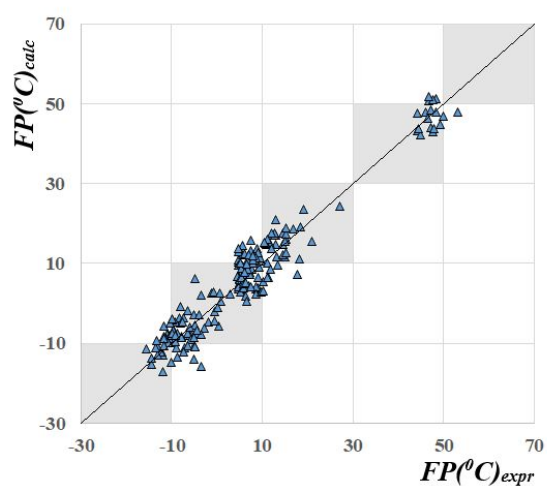
Table 3. The statistical characteristics of the CORAL models for flash point of ternary liquid mixtures of organic substances based on TF_1 and TF_2

Split	TF	Set	n*	R ²	CCC	IIC	Q ²	Q ² _{F1}	Q ² _{F2}	Q ² _{F3}	RMSE
1	1	Active training	200	0.8846	0.9388	0.7852	0.8815				6.34
		Passive training	206	0.8581	0.9239	0.8028	0.8543				7.15
		Calibration	195	0.8964	0.9450	0.9420	0.8937	0.8842	0.8828	0.9290	4.91
		Validation	207	0.9085	0.9516	0.8310	0.9063				4.86
	2	Active training	200	0.8779	0.9350	0.7981	0.8750				6.52
		Passive training	206	0.8755	0.9356	0.9091	0.8722				6.58
		Calibration	195	0.9262	0.9609	0.9619	0.9246	0.9181	0.9171	0.9498	4.13
		Validation	207	0.9345	0.9653	0.9234	0.9331				4.12
2	1	Active training	204	0.8015	0.8898	0.8779	0.7972				8.59
		Passive training	205	0.7313	0.8519	0.8215	0.7248				9.13
		Calibration	205	0.8096	0.8976	0.8905	0.8048	0.8004	0.7964	0.8881	6.14
		Validation	194	0.8888	0.9414	0.9264					5.55
	2	Active training	204	0.8795	0.9359	0.8671	0.8765				6.69
		Passive training	205	0.8614	0.9274	0.7248	0.8576				6.52
		Calibration	205	0.9203	0.9580	0.9593	0.9186	0.9214	0.9198	0.9559	3.86
		Validation	194	0.9392	0.9684	0.9447	0.9379				4.03
3	1	Active training	204	0.7630	0.8656	0.6896	0.7579				9.23
		Passive training	209	0.7738	0.8754	0.7947	0.7691				8.16
		Calibration	197	0.9165	0.9572	0.9217	0.9149	0.9128	0.9128	0.9313	4.70
		Validation	198	0.9105	0.9527	0.8223					4.63
	2	Active training	204	0.8595	0.9245	0.7768	0.8557				7.10
		Passive training	209	0.8633	0.9246	0.6541	0.8597				6.48
		Calibration	197	0.9386	0.9678	0.9688	0.9372	0.9325	0.9324	0.9468	4.13
		Validation	198	0.9372	0.9658	0.8784					4.02

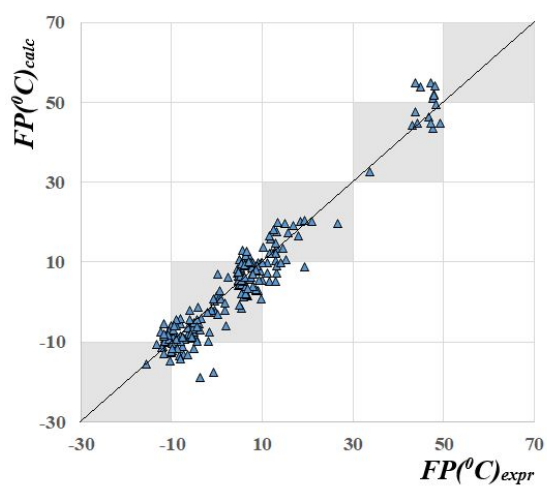
*) the n is number of compounds in a set; R^2 is correlation coefficient; CCC is concordance correlation coefficient; IIC is index of ideality of correlation; Q^2 is cross validated correlation coefficient; Q^2_{F1} , Q^2_{F2} , and Q^2_{F3} criteria of predictive potential suggested in the literature [34]; RMSE is root mean squared error.



Split 1 , Eq. 17



Split 2 , Eq. 18



Split 3 , Eq. 19

Figure 5. Graphical representation of models calculated with Eqs. 17-19 for splits 1-3 (validation sets)

1
2
3 In contrast to the Monte Carlo optimization with TF_1 , the optimization with TF_2 is sensitive
4 to both correlation and dispersion of points in diagram “experiment-calculation”. It seems, that
5 this is the reason why TF_2 gives better results. Nevertheless, different splits result in different
6 statistical quality (predictive potential) of models (**Table 3**).
7
8
9

10 It is to be noted, that any approach in several attempts to build up a model with different
11 “correct” or “random” splits of data into the training set and validation set can furnish “unpleasant”
12 results, e.g. the Kubinyi paradox.^{25,26} Perhaps, this is the main reason, why many published
13 QSPR/QSAR models, as a rule, are based on one “correct” (or “rational”) split, but do not provide
14 information on the results of applying of several random splits.
15
16
17
18

19 **Table 4** indicates that the statistical quality of the CORAL models and the statistical quality
20 of models, which were suggested in the literature^{4, 35-38} are comparable. However, in the model
21 available in the literature^{4,35-38} descriptors derived from quantum mechanics together with
22 additional physicochemical parameters and 3D descriptors were involved in building up models.
23 The CORAL models applied in our study are derived with data on molecular structure of
24 components and molar fractions (i.e. without additional physicochemical and 3D data). This
25 difference should be considered as the advantage of the CORAL approach in comparison with
26 above approaches.^{4,35-38} The CORAL models examined here, were checked up with several
27 distribution of available data into the training and validation sets. This indicates that the checking
28 up of the CORAL approach is more meticulous than the previously published works since only
29 one split into training and validation sets has been considered in those publications.
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Table 4. Comparison of QSPR models for flash point of ternary mixtures

Component1	Component2	Component3	Set	n*	R ²	MAE	Comments on Building up models
In literature							
methanol	methylacetate	methylacrylate	Total	68	-	4.7	Physicochemical parameters of atoms and molecules; topological and 3D descriptors; descriptors of quantum mechanics are used to build up model [4,35,36]
methanol	ethanol	acetone		66	-	14.7	
methanol	toluene	i-octane		147	-	7.6	
methanol	n-decane	acetone		119	-	7.3	
acetic acid	n-hexanol	cyclohexanone		59	-	19.1	
i-propanol	ethanol	n-octane		101	-	10.9	
2-butanol	ethanol	n-octane		109	-	4.8	
cyclohexanol	ethanol	n-octane		94	-	7.7	
			Training	1104	0.9350	-	Pure-binary-ternary mixtures; quantum mechanics descriptors [37]
			Test	276	0.9335	-	
			Total	1380	0.9405	-	
			Training	137	0.902	-	
Test	-	-	-				
In this work							
Organic compounds from list**			Total	808	0.8966	4.17	SMILES of components together with corresponding molar fractions are used to build up model
				808	0.8941	4.25	
				808	0.8903	4.15	

*) The n is the number of ternary mixtures; R² is the determination coefficient; MAE is mean absolute error.

**) The list of compounds is the following: methanol; methyl acetate; methyl ethyl ketone; ethanol; ethylene glycol butyl ether; acetone; toluene; i-propanol; i-octane; n-pentanol; n-heptane; n-octane; n-decane; acetic acid; n-hexanol; cyclohexanone; i-propanol; 2-butanol; cyclohexanol.

There is one more advantage of the CORAL models. These models are providers of the mechanistic interpretation in form of promoters of increase or decrease of an endpoint.^{12, 20-24} The promoters of increase are attributes of quasi-SMILES, which have only positive values in several probes of the Monte Carlo optimization. Vice versa, attributes of quasi-SMILES, which have only negative correlation weights in several probes of the optimization should be considered as promoters of decrease for an endpoint. It is necessary, however, to take into account the prevalence of corresponding attributes in the training set and validation set: rare attributes hardly should be considered as source of reliable heuristic hypotheses.

Based on our results the promoters of increase for flash point are (i) branching, i.e. presence brackets in SMILES; (ii) presence of cycles (digits in SMILES); (iii) presence of carbon and oxygen atoms; and (iv) presence of double bonds. *Supplementary materials* contains the list of promoters for flash points increase.

Finally, it should be noted that the CORAL software^{5, 15, 16, 21, 22, 30} as well as the technique of quasi-SMILES^{7-9, 18, 31-33} can be applied for building up predictive models of other endpoints. Thus, the described approach can be useful for technologists, because: (i) data required to calculate

1
2
3 the flash point include molecular structures (represented by SMILES) and molar fractions for two
4 compounds, without other data (e.g. 3D geometry, quantum mechanics descriptors, different
5 physicochemical parameters, etc.); and (ii) suggested models are freely available on the Internet.
6
7

8 **Supplementary materials** contain three splits of quasi-SMILES into the active training set,
9 passive training set, calibration set, and validation set for ternary liquid mixtures used here. In
10 addition, mechanistic interpretation in form of the list of molecular features, which are promoters
11 for increase of FP, and the domain of applicability, are presented in **Supplementary materials**.
12
13
14

15 CONCLUSIONS

16
17 The described approach based on quasi-SMILES provides possibility of fast modifications
18 of corresponding databases (corrections of wrong data as well as applying new data valuable from
19 practical and theoretical points of view).
20
21

22 The approach gives possibility to define mechanistic interpretation of a model via lists of
23 molecular features, which are promoters of increase (or decrease) for flash points. The quasi-
24 SMILES technique generates quite good models for flash points of ternary liquid mixtures of
25 organic substances.
26
27
28

29 The Index of Ideality of Correlation is important and useful component of the target
30 function in the Monte Carlo optimization, since this approach allows improving the predictive
31 potential of models of flash point, for the external invisible validation set.
32
33

34 AUTHOR CONTRIBUTIONS

35 Authors have done equivalent contributions to this work.
36

37 CONFLICT OF INTEREST

38 The authors confirm that this article content has no conflict of interest.
39

40 ACKNOWLEDGEMENT

41 AAT and APT thank the project LIFE-CONCERT contract (LIFE17 GIE/IT/000461) for
42 the support. JL and DL would like to thank NSF-CREST (Award No. HRD 154774) program for
43 the support.
44
45
46
47
48

49 REFERENCES

- 50
51 1. Fu, J. Flash points measurements and prediction of biofuels and biofuel blends with
52 aromatic fluids. *Fuel* **2019**, *241*, 892-900. <https://doi.org/10.1016/j.fuel.2018.12.105>
53
54 2. Gaudin, T.; Rotureau, P.; Fayet, G. Combining mixing rules with QSPR models for pure
55 chemicals to predict the flash points of binary organic liquid mixtures. *Fire Saf. J.* **2015**, *74*, 61-
56
57
58
59
60
70. <https://doi.org/10.1016/j.firesaf.2015.04.006>

- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
3. Hristova, M. Measurement and prediction of binary mixture flash point. *Cent. Eur. J. Chem.* **2013**, *11(1)*, 57-62. <https://doi.org/10.2478/s11532-012-0131-1>
4. Fayet, G.; Rotureau, P. New QSPR Models to Predict the Flammability of Binary Liquid Mixtures. *Mol. Inf.* **2019**, *38*, 180012. <https://doi.org/10.1002/minf.201800122>
5. Worachartcheewan, A.; Mandi, P.; Prachayasittikul, V.; Toropova, A.P.; Toropov, A.A.; Nantasenamat, C. Large-scale QSAR study of aromatase inhibitors using SMILES-based descriptors. *Chemom. Intell. Lab. Syst.* **2014**, *138*, 120-126. <https://doi.org/10.1016/j.chemolab.2014.07.017>
6. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31-36. DOI: 10.1021/ci00057a005
7. Toropov, A.A.; Toropova, A.P. Quasi-SMILES and nano-QFAR: United model for mutagenicity of fullerene and MWCNT under different conditions. *Chemosphere* **2015**, *139*, 18-22. <https://doi.org/10.1016/j.chemosphere.2015.05.042>
8. Toropova, A.P.; Toropov, A.A.; Rallo, R.; Leszczynska, D.; Leszczynski, J. Optimal descriptor as a translator of eclectic data into prediction of cytotoxicity for metal oxide nanoparticles under different conditions. *Ecotoxicol. Environ. Saf.* **2015**, *112*, 39-45. <https://doi.org/10.1016/j.ecoenv.2014.10.003>
9. Toropova, A.P.; Toropov, A.A. Mutagenicity: QSAR -quasi-QSAR -nano-QSAR. *Mini-Rev. Med. Chem.* **2015**, *15*, 608-621. <https://doi:10.2174/1389557515666150219121652>
10. Toropova, A.P.; Toropov, A.A.; Benfenati, E.; Gini, G.; Leszczynska, D.; Leszczynski, J. CORAL: Models of toxicity of binary mixtures. *Chemom. Intell. Lab. Syst.* **2012**, *119*, 39-43. <https://doi.org/10.1016/j.chemolab.2012.10.001>
11. Toropova, A.P.; Toropov, A.A.; Carnesecchi, E.; Benfenati, E.; Dorne, J.L. The index of ideality of correlation: models for flammability of binary liquid mixtures. *Chem. Pap.* Published online: 19 August **2019**. <https://doi.org/10.1007/s11696-019-00903-w>
12. Toropov, A.A.; Toropova, A.P. The index of ideality of correlation: A criterion of predictive potential of QSPR/QSAR models? *Mutat. Res. Genet. Toxicol. Environ. Mutagen.* **2017**, *819*, 31-37. <https://doi.org/10.1016/J.MRGENTOX.2017.05.008>
13. Golubović, M.; Lazarević, M.; Zlatanović, D.; Krtinić, D.; Stoičkov, V.; Mladenović, B.; Milić, D.J.; Sokolović, D.; Veselinović, A.M. The anesthetic action of some polyhalogenated ethers—Monte Carlo method based QSAR study. *Comput. Biol. Chem.* **2018**, *75*, 32-38. DOI: 10.1016/j.compbiolchem.2018.04.009

14. Stoičkov, V.; Stojanović, D.; Tasić, I.; Šarić, S.; Radenković, D.; Babović, P.; Sokolović, D.; Veselinović, A.M. QSAR study of 2,4-dihydro-3H-1,2,4-triazol-3-ones derivatives as angiotensin II AT1 receptor antagonists based on the Monte Carlo method. *Struct. Chem.* **2018**, *29*(2), 44-49. DOI: 10.1007/s11224-017-1041-9
15. Kumar, P.; Kumar, A.; Sindhu, J. Design and development of novel focal adhesion kinase (FAK) inhibitors using Monte Carlo method with index of ideality of correlation to validate QSAR. *SAR QSAR Environ. Res.* **2019**, *30*(2), 63-80. DOI: 10.1080/1062936X.2018.1564067
16. Kumar, P.; Kumar, A.; Sindhu, J.; Lal, S. QSAR Models for Nitrogen Containing Monophosphonate and Bisphosphonate Derivatives as Human Farnesyl Pyrophosphate Synthase Inhibitors Based on Monte Carlo Method. *Drug Res.* **2019**, *69*(3), 159-167. DOI: 10.1055/a-0652-5290
17. Jain, S.; Amin, S.A.; Adhikari, N.; Jha, T.; Gayen, S. Good and bad molecular fingerprints for human rhinovirus 3C protease inhibition: identification, validation, and application in designing of new inhibitors through Monte Carlo-based QSAR study. *J. Biomol. Struct. Dynamics*. Published online: 31 Jan **2019**. DOI: 10.1080/07391102.2019.1566093
18. Toropov, A.A.; Toropova, A.P. Quasi-QSAR for mutagenic potential of multi-walled carbon-nanotubes. *Chemosphere* 2015, *124*(1), 40-46. DOI: 10.1016/j.chemosphere.2014.10.067
19. Toropov, A.A.; Toropova, A.P. QSAR as a random event: criteria of predictive potential for a chance model. *Struct. Chem.* **2019**, *30*(5), 1677-1683. <https://doi.org/10.1007/s11224-019-01361-6>
20. Toropova, A.P.; Toropov, A.A.; Marzo, M.; Escher, S.E.; Dorne, J.L.; Georgiadis, N.; Benfenati, E. The application of new HARD-descriptor available from the CORAL software to building up NOAEL models. *Food Chem. Toxicol.* **2018**, *112*, 544-550. DOI: 10.1016/j.fct.2017.03.060
21. Toropova, A.P.; Toropov, A.A. The index of ideality of correlation: A criterion of predictability of QSAR models for skin permeability? *Sci. Total. Environ.* **2017**, *586*, 466-472. <https://doi.org/10.1016/J.SCITOTENV.2017.01.198>
22. Toropova, A.P.; Toropov, A.A. Use of the index of ideality of correlation to improve models of eco-toxicity. *Environ. Sci. Pollut. Res.* **2018**, *25*(31), 31771-31775. DOI: 10.1007/s11356-018-3291-5
23. Toropov, A.A.; Carbó-Dorca, R.; Toropova, A.P. Index of Ideality of Correlation: new possibilities to validate QSAR: a case study. *Struct. Chem.* **2018**, *29*(1), 33-38. <https://doi.org/10.1007/s11224-017-0997-9>

24. Toropov, A.A.; Raška I.; Toropova AP.; Raškova M.; Veselinović AM.; Veselinović JB. The study of the index of ideality of correlation as a new criterion of predictive potential of QSPR/QSAR-models. *Sci. Total. Environ.* **2019**, *659*, 1387-1394.

<https://doi.org/10.1016/j.scitotenv.2018.12.439>

25. Kubinyi, H.; Hamprecht, F.A.; Mietzner, T. Three-dimensional quantitative similarity-activity relationships (3D QSiAR) from SEAL similarity matrices. *J. Med. Chem.* **1998**, *41(14)*, 2553-2564. DOI: 10.1021/jm970732a

26. Hartung, T.; Hoffmann, S. Food for thought... on in silico methods in toxicology. *Altex* **2009**, *26(3)*, 155-166. DOI: 10.14573/altex.2009.3.155

27. Saldana, D.A.; Starck, L.; Mougin, P.; Rousseau, B.; Creton, B. Prediction of flash points for fuel mixtures using machine learning and a novel equation. *Energy. Fuels* **2013**, *27(7)*, 3811-3820. DOI: 10.1021/ef4005362

28. Gaudin, T.; Rotureau, P.; Fayet, G. Mixture Descriptors toward the Development of Quantitative Structure-Property Relationship Models for the Flash Points of Organic Mixtures. *Ind. Eng. Chem. Res.* **2015**, *54(25)*, 6596-6604. DOI: 10.1021/acs.iecr.5b01457

29. Jiao, L.; Zhang, X.; Qin, Y.; Wang, X.; Li, H. QSPR study on the flash point of organic binary mixtures by using electrotopological state index. *Chemom. Intell. Lab. Syst.* **2015**, *156*, 211-216. <https://doi.org/10.1016/j.chemolab.2016.05.023>

30. Ahmadi, S.; Akbari, A. Prediction of the adsorption coefficients of some aromatic compounds on multi-wall carbon nanotubes by the Monte Carlo method. *SAR QSAR Environ. Res.* **2018**, *29(11)*, 895-909. DOI: 10.1080/1062936X.2018.1526821

31. Toropova, A.P.; Toropov, A.A. QSPR and nano-QSPR: What is the difference? *J. Mol. Struct.* **2019**, *1182*, 141-149. [https://DOI: 10.1016/j.molstruc.2019.01.040](https://doi.org/10.1016/j.molstruc.2019.01.040)

32. Trinh TX, Choi J-S, Jeon H, Byun H-G, Yoon T-H, Kim J. Quasi-SMILES-Based Nano-Quantitative Structure-Activity Relationship Model to Predict the Cytotoxicity of Multiwalled Carbon Nanotubes to Human Lung Cells. *Chem. Res. Toxicol.* **2018**, *31(3)*, 183-190. DOI: 10.1021/acs.chemrestox.7b00303

33. Choi, J.-S.; Trinh, T.X.; Yoon, T.-H.; Kim, J.; Byun, H.-G. Quasi-QSAR for predicting the cell viability of human lung and skin cells exposed to different metal oxide nanomaterials. *Chemosphere* **2019**, *217*, 243-249. <https://doi.org/10.1016/j.chemosphere.2018.11.014>

34. Consonni, V.; Ballabio, D.; Todeschini, R. Comments on the definition of the Q^2 parameter for QSAR validation. *J. Chem. Inf. Model.* **2009**, *49(7)*, 1669-1678. DOI: 10.1021/ci900115y

1
2
3 35. Liaw, H.-J.; Tang, C.-L.; Lai, J.-S. A model for predicting the flash point of ternary
4 flammable solutions of liquid. *Combust. Flame* **2004**, *138*(4), 308-319. DOI:
5 10.1016/j.combustflame.2004.06.002
6
7
8
9

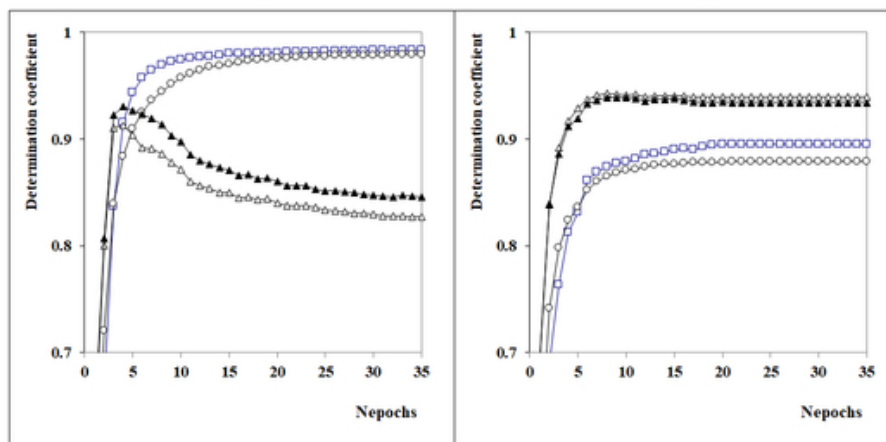
10
11 36. Zarringhalam Moghaddam, A.; Rafiei, A.; Khalili, T. Assessing prediction models on
12 calculating the flash point of organic acid, ketone and alcohol mixtures. *Fluid Ph. Equilibria* **2012**,
13 *316*, 117-121. DOI: 10.1016/j.fluid.2011.12.014
14
15
16

17 37. Wang, Y.; Yan, F.; Jia, Q.; Wang, Q. Distributive structure-properties relationship for
18 flash point of multiple components mixture. *Fluid Ph. Equilibria* 2018, *474*, 1-5. DOI:
19 10.1016/j.fluid.2018.07.005
20
21

22 38. Cheng, J.; Pan, Y.; Song, X.; Jiang, J.; Li, G.; Ding, L.; Chang, H. A new method for
23 the prediction of flash points for ternary miscible mixtures. *Process Saf. Environ.* **2015**, *95*, 102-
24 113. DOI: 10.1016/j.psep.2015.02.019
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Optimal Descriptor $[TF_2] = F(\text{component-1, molar fraction-1,}$
 $\text{component-2, molar fraction-2,}$
 $\text{component-3, molar fraction-3})$

$$\text{Flash point of ternary mixture} = C_0 + C_1 * \text{Optimal descriptor } [TF_2]$$

 TF_1 TF_2 

Active training set (\square), Passive training set (\circ), Calibration set (\triangle), Validation set (\blacktriangle)

49x38mm (300 x 300 DPI)